Electrical Engineer's Reference Book

VISIT…

LANZAROTE
Caliente.com

Important notice
Many practical techniques described in this book involve potentially dangerous applications of electricity and engineering equipment. The authors, editors and publishers cannot take responsibility for any personal, professional or financial risk involved in carrying out these techniques, or any resulting injury, accident or loss. The techniques described in this book should only be implemented by professional and fully qualified electrical engineers using their own professional judgement and due regard to health and safety issues.

Electrical Engineer's Reference Book

Sixteenth edition

M. A. Laughton CEng., FIEE

D. J. Warne CEng., FIEE

Newnes

OXFORD AMSTERDAM BOSTON NEW YORK
LONDON PARIS SAN DIEGO SAN FRANCISCO
SINGAPORE SYDNEY TOKYO

Newnes
An imprint of Elsevier Science
Linacre House, Jordan Hill, Oxford OX2 8DP
200 Wheeler Road, Burlington, MA 01803
A division of Reed Educational and Professional Publishing Ltd

A member of the Reed Elsevier plc group

First published in 1945 by George Newnes Ltd
Fifteenth edition 1993
Sixteenth edition 2003

Copyright © Elsevier Science, 2003. All rights reserved

No part of this publication may be
reproduced in any material form (including photocopying or storing in any medium by electronic means and whether or not transiently or incidentally to some other use of this publication) without the written permission of the copyright holder except in accordance with the provisions of the Copyright, Designs and Patents Act 1988 or under the terms of a licence issued by the Copyright Licensing Agency Ltd, 90 Tottenham Court Road, London, England W1T 4LP. Applications for the copyright holder's written permission to reproduce any part of this publication should be addressed to the publishers

British Library Cataloguing in Publication Data
A catalogue record for this book is available from the
British Library

ISBN 0 7506 46373

For information on all Newnes publications
visit our website at www.newnespress.com

Typeset in India by Integra Software Services Pvt. Ltd,
Pondicherry 605 005, India. www.integra-india.com
Printed and bound in Great Britain

Contents

Preface

Section A – General Principles

1 Units, Mathematics and Physical Quantities

International unit system . Mathematics . Physical quantities . Physical properties . Electricity

2 Electrotechnology

Nomenclature . Thermal effects . Electrochemical effects . Magnetic field effects . Electric field effects . Electromagnetic field effects . Electrical discharges

3 Network Analysis

Introduction . Basic network analysis . Power-system network analysis

Section B – Materials & Processes

4 Fundamental Properties of Materials

Introduction . Mechanical properties . Thermal properties . Electrically conducting materials . Magnetic materials . Dielectric materials . Optical materials . The plasma state

5 Conductors and Superconductors

Conducting materials . Superconductors

6 Semiconductors, Thick and Thin-Film Microcircuits

Silicon, silicon dioxide, thick- and thin-film technology . Thick- and thin-film microcircuits

7 Insulation

Insulating materials . Properties and testing . Gaseous dielectrics . Liquid dielectrics . Semi-fluid and fusible materials . Varnishes, enamels, paints and lacquers . Solid dielectrics . Composite solid/liquid dielectrics . Irradiation effects . Fundamentals of dielectric theory . Polymeric insulation for high voltage outdoor applications

8 Magnetic Materials

Ferromagnetics . Electrical steels including silicon steels . Soft irons and relay steels . Ferrites . Nickel–iron alloys . Iron–cobalt alloys . Permanent magnet materials

9 Electroheat and Materials Processing

Introduction . Direct resistance heating . Indirect resistance heating . Electric ovens and furnaces . Induction heating . Metal melting . Dielectric heating . Ultraviolet processes . Plasma torches . Semiconductor plasma processing . Lasers

10 Welding and Soldering

Arc welding . Resistance welding . Fuses . Contacts . Special alloys . Solders . Rare and precious metals . Temperature-sensitive bimetals . Nuclear-reactor materials . Amorphous materials

Section C – Control

11 Electrical Measurement

Introduction . Terminology . The role of measurement traceability in product quality . National and international measurement standards . Direct-acting analogue measuring instruments . Integrating (energy) metering . Electronic instrumentation . Oscilloscopes . Potentiometers and bridges . Measuring and protection transformers . Magnetic measurements . Transducers . Data recording

12 Industrial Instrumentation

Introduction . Temperature . Flow . Pressure . Level transducers . Position transducers . Velocity and acceleration . Strain gauges, loadcells and weighing . Fieldbus systems . Installation notes

13 Control Systems

Introduction . Laplace transforms and the transfer function . Block diagrams . Feedback . Generally desirable and acceptable behaviour . Stability . Classification of system and static accuracy. Transient behaviour . Root-locus method . Frequency-response methods . State-space description . Sampled-data systems . Some necessary mathematical preliminaries . Sampler and zero-order hold . Block diagrams . Closed-loop systems . Stability . Example . Dead-beat response . Simulation . Multivariable control . Dealing with non linear elements .

Disturbances . Ratio control . Transit delays . Stability . Industrial controllers . Digital control algorithms . Auto-tuners . Practical tuning methods

14 Digital Control Systems

Introduction . Logic families . Combinational logic . Storage . Timers and monostables . Arithmetic circuits . Counters and shift registers . Sequencing and event driven logic . Analog interfacing . Practical considerations . Data sheet notations

15 Microprocessors

Introduction . Structured design of programmable logic systems . Microprogrammable systems . Programmable systems . Processor instruction sets . Program structures . Reduced instruction set computers (RISC) . Software design . Embedded systems

16 Programmable Controllers

Introduction . The programmable controller . Programming methods . Numerics . Distributed systems and fieldbus . Graphics . Software engineering . Safety

Section D – Power Electronics and Drives

17 Power Semiconductor Devices

Junction diodes . Bipolar power transistors and Darlingtons . Thyristors . Schottky barrier diodes . MOSFET . The insulated gate bipolar transistor (IGBT)

18 Electronic Power Conversion

Electronic power conversion principles . Switch-mode power supplies . D.c/a.c. conversion . A.c./d.c. conversion . A.c./a.c. conversion . Resonant techniques . Modular systems . Further reading

19 Electrical Machine Drives

Introduction . Fundamental control requirements for electrical machines . Drive power circuits . Drive control . Applications and drive selection . Electromagnetic compatibility

20 Motors and Actuators

Energy conversion . Electromagnetic devices . Industrial rotary and linear motors

Section E – Environment

21 Lighting

Light and vision . Quantities and units . Photometric concepts . Lighting design technology . Lamps . Lighting design . Design techniques . Lighting applications

22 Environmental Control

Introduction . Environmental comfort . Energy requirements . Heating and warm-air systems . Control . Energy conservation . Interfaces and associated data

23 Electromagnetic Compatibility

Introduction . Common terms . The EMC model . EMC requirements . Product design . Device selection . Printed circuit boards . Interfaces . Power supplies and power-line filters . Signal line filters . Enclosure design . Interface cable connections . Golden rules for effective design for EMC . System design . Buildings . Conformity assessment . EMC testing and measurements . Management plans

24 Health and Safety

The scope of electrical safety considerations . The nature of electrical injuries . Failure of electrical equipment

25 Hazardous Area Technology

A brief UK history . General certification requirements . Gas group and temperature class . Explosion protection concepts . ATEX certification . Global view . Useful websites

Section F – Power Generation

26 Prime Movers

Steam generating plant . Steam turbine plant . Gas turbine plant . Hydroelectric plant . Diesel-engine plant

27 Alternative Energy Sources

Introduction . Solar . Marine energy . Hydro . Wind . Geothermal energy. Biofuels . Direct conversion . Fuel cells . Heat pumps

28 Alternating Current Generators

Introduction . Airgap flux and open-circuit e.m.f. . Alternating current windings . Coils and insulation . Temperature rise . Output equation . Armature reaction . Reactances and time constants . Steady-state operation . Synchronising . Operating charts . On-load excitation . Sudden three phase short circuit . Excitation systems . Turbogenerators . Generator–transformer connection . Hydrogenerators . Salient-pole generators other than hydrogenerators . Synchronous compensators . Induction generators . Standards

29 Batteries

Introduction . Cells and batteries . Primary cells . Secondary cells and batteries . Battery applications . Anodising . Electrodeposition . Hydrogen and oxygen electrolysis

Section G – Transmission and Distribution

30 Overhead Lines

General . Conductors and earth wires . Conductor fittings . Electrical characteristics . Insulators . Supports . Lightning . Loadings

31 Cables

Introduction . Cable components . General wiring cables and flexible cords . Supply distribution cables . Transmission cables . Current-carrying capacity . Jointing and accessories . Cable fault location

32 HVDC

Introduction . Applications of HVDC . Principles of HVDC converters . Transmission arrangements . Converter station design . Insulation co-ordination of HVDC converter stations . HVDC thyristor valves . Design of harmonic filters for HVDC converters . Reactive power considerations . Control of HVDC . A.c. system damping controls . Interaction between a.c. and d.c. systems . Multiterminal HVDC systems . Future trends

33 Power Transformers

Introduction . Magnetic circuit . Windings and insulation . Connections . Three-winding transformers . Quadrature booster transformers . On-load tap changing . Cooling . Fittings . Parallel operation . Auto-transformers . Special types . Testing . Maintenance . Surge protection . Purchasing specifications

34 Switchgear

Circuit-switching devices . Materials . Primary-circuit-protection devices . LV switchgear . HV secondary distribution switchgear . HV primary distribution switchgear . HV transmission switchgear . Generator switchgear . Switching conditions . Switchgear testing . Diagnostic monitoring . Electromagnetic compatibility . Future developments

35 Protection

Overcurrent and earth leakage protection . Application of protective systems . Testing and commissioning . Overvoltage protection

36 Electromagnetic Transients

Introduction . Basic concepts of transient analysis . Protection of system and equipment against transient overvoltage . Power system simulators . Waveforms associated with the electromagnetic transient phenomena

37 Optical Fibres in Power Systems

Introduction . Optical fibre fundamentals . Optical fibre cables . British and International Standards . Optical fibre telemetry on overhead power lines . Power equipment monitoring with optical fibre sensors

38 Installation

Layout . Regulations and specifications . High-voltage supplies . Fault currents . Substations . Wiring systems . Lighting and small power . Floor trunking . Stand-by and emergency supplies . Special buildings . Low-voltage switchgear and protection . Transformers . Power-factor correction . Earthing . Inspection and testing

Section H – Power Systems

39 Power System Planning

The changing electricity supply industry (ESI) . Nature of an electrical power system . Types of generating plant and characteristics . Security and reliability of a power system . Revenue collection . Environmental sustainable planning

40 Power System Operation and Control

Introduction . Objectives and requirements . System description . Data acquisition and telemetering . Decentralised control: excitation systems and control characteristics of synchronous machines . Decentralised control: electronic turbine controllers . Decentralised control: substation automation . Decentralised control: pulse controllers for voltage control with tap-changing transformers. Centralised control . System operation . System control in liberalised electricity markets . Distribution automation and demand side management . Reliability considerations for system control

41 Reactive Power Plant and FACTS Controllers

Introduction . Basic concepts . Variations of voltage with load . The management of vars . The development of FACTS controllers . Shunt compensation . Series compensation . Controllers with shunt and series components . Special aspects of var compensation . Future prospects

42 Electricity Economics and Trading

Introduction . Summary of electricity pricing principles . Electricity markets . Market models . Reactive market

43 Power Quality

Introduction . Definition of power quality terms . Sources of problems . Effects of power quality problems . Measuring power quality . Amelioration of power quality problems . Power quality codes and standards

Section I – Sectors of Electricity Use

44 Road Transport

Electrical equipment of road transport vehicles . Light rail transit . Battery vehicles . Road traffic control and information systems

45 Railways

Railway electrification . Diesel-electric traction . Systems, EMC and standards . Railway signalling and control

46 Ships

Introduction . Regulations . Conditions of service . D.c. installations . A.c. installations . Earthing . Machines and transformers . Switchgear . Cables . Emergency power . Steering gear . Refrigerated cargo spaces . Lighting . Heating . Watertight doors . Ventilating fans . Radio interference and electromagnetic compatibility . Deck auxiliaries . Remote and automatic control systems . Tankers . Steam plant . Generators . Diesel engines . Electric propulsion

47 Aircraft

Introduction . Engine technology . Wing technology . Integrated active controls . Flight-control systems . Systems technology . Hydraulic systems . Air-frame mounted accessory drives . Electrohydraulic flight controls . Electromechanical flight controls . Aircraft electric power . Summary of power systems . Environmental control system . Digital power/digital load management

48 Mining Applications

General . Power supplies . Winders . Underground transport . Coal-face layout . Power loaders . Heading machines . Flameproof and intrinsically safe equipment . Gate-end boxes . Flameproof motors . Cables, couplers, plugs and sockets . Drilling machines . Underground lighting . Monitoring and control

49 Standards and Certification

Introduction . Organisations preparing electrical standards . The structure and application of standards . Testing, certification and approval to standard recommendations . Sources of standards information

Index

Preface

The *Electrical Engineer's Reference Book* was first published in 1945: its original aims, to reflect the state of the art in electrical science and technology, have been kept in view throughout the succeeding decades during which subsequent editions have appeared at regular intervals.

Publication of a new edition gives the opportunity to respond to many of the changes occurring in the practice of electrical engineering, reflecting not only the current commercial and environmental concerns of society, but also industrial practice and experience plus academic insights into fundamentals. For this 16th edition, thirty-nine chapters are either new, have been extensively rewritten, or augmented and updated with new material. As in earlier editions this wide range of material is brought within the scope of a single volume. To maintain the overall length within the possible bounds some of the older material has been deleted to make way for new text.

The organisation of the book has been recast in the following format with the aim of facilitating quick access to information.

General Principles (Chapters 1–3) covers basic scientific background material relevant to electrical engineering. It includes chapters on units, mathematics and physical quantities, electrotechnology and network analysis.

Materials & Processes (Chapters 4–10) describes the fundamentals and range of materials encountered in electrical engineering in terms of their electromechanical, thermoelectric and electromagnetic properties. Included are chapters on the fundamental properties of materials, conductors and superconductors, semiconductors, insulation, magnetic materials, electroheat and materials processing and welding and soldering.

Control (Chapters 11–16) is a largely new section with six chapters on electrical measurement and instruments, industrial instrumentation for process control, classical control systems theory, fundamentals of digital control, microprocessors and programmable controllers.

Power Electronics and Drives (Chapters 17–20) reflect the significance of upto 50% of all electrical power passing through semiconductor conversion. The subjects included of greatest importance to industry, particularly those related to the area of electrical variable speed drives, comprise power semiconductor devices, electronic power conversion, electrical machine drives, motors and actuators.

Environment (Chapters 21–25) is a new section of particular relevance to current concerns in this area including lighting, environmental control, electromagnetic compatibility, health and safety, and hazardous area technology.

Power Generation (Chapters 26–29) sees some rationalisation of contributions to previous editions in the largely mechanical engineering area of prime movers, but with an expanded treatment of the increasingly important topic of alternative energy sources, along with further chapters on alternating current generators and batteries.

Transmission and Distribution (Chapters 30–38) is concerned with the methods and equipment involved in the delivery of electric power from the generator to the consumer. It deals with overhead lines, cables, HVDC transmission, power transformers, switchgear, protection, and optical fibres in power systems and aspects of installation with an additional chapter on the nature of electromagnetic transients.

Power Systems (Chapters 39–43) gathers together those topics concerned with present day power system planning and power system operation and control, together with aspects of related reactive power plant and FACTS controllers. Chapters are included on electricity economics and trading in the liberalised electricity supply industry now existing in many countries, plus an analysis of the power supply quality necessary for modern industrialised nations.

Sectors of Electricity Use (Chapters 44–49) is a concluding section comprising chapters on the special requirements of agriculture and horticulture, roads, railways, ships, aircraft, and mining with a final chapter providing a preliminary guide to Standards and Certification.

Although every effort has been made to cover the scope of electrical engineering, the nature of the subject and the manner in which it is evolving makes it inevitable that improvements and additions are possible and desirable. In order to ensure that the reference information provided remains accurate and relevant, communications from professional engineers are invited and all are given careful consideration in the revision and preparation of new editions of the book.

The expert contributions made by all the authors involved and their patience through the editorial process is gratefully acknowledged.

M. A. Laughton
D. F. Warne
2002

Electrical Engineer's Reference Book—online edition

As this book goes to press an online electronic version is also in preparation. The online edition will feature

- the complete text of the book
- access to the latest revisions (a rolling chapter-by-chapter revision will take place between print editions)
- additional material not included in the print version

To find out more, please visit the *Electrical Engineer's Reference Book* web page:
http://www.bh.com/newness?isbn = 0750646373

or send an e-mail to newnes@elsevier.com

Section A

General Principles

1

Units, Mathematics and Physical Quantities

M G Say PhD, MSc, CEng, ACGI, DIC, FIEE, FRSE
Formerly of Heriot-Watt University

M A Laughton BASc, PhD, DSc(Eng), FREng, CEng, FIEE
Formerly of Queen Mary & Westfield College, University of London
(Section 1.2.10)

Contents

1.1 International unit system 1/3
  1.1.1 Base units 1/3
  1.1.2 Supplementary units 1/3
  1.1.3 Notes 1/3
  1.1.4 Derived units 1/3
  1.1.5 Auxiliary units 1/4
  1.1.6 Conversion factors 1/4
  1.1.7 CGS electrostatic and electromagnetic units 1/4

1.2 Mathematics 1/4
  1.2.1 Trigonometric relations 1/6
  1.2.2 Exponential and hyperbolic relations 1/7
  1.2.3 Bessel functions 1/9
  1.2.4 Series 1/9
  1.2.5 Fourier series 1/9
  1.2.6 Derivatives and integrals 1/10
  1.2.7 Laplace transforms 1/10
  1.2.8 Binary numeration 1/10
  1.2.9 Power ratio 1/13
  1.2.10 Matrices and vectors 1/13

1.3 Physical quantities 1/17
  1.3.1 Energy 1/17
  1.3.2 Structure of matter 1/19

1.4 Physical properties 1/26

1.5 Electricity 1/26
  1.5.1 Charges at rest 1/26
  1.5.2 Charges in motion 1/26
  1.5.3 Charges in acceleration 1/28

This reference section provides (a) a statement of the International System (SI) of Units, with conversion factors; (b) basic mathematical functions, series and tables; and (c) some physical properties of materials.

1.1 International unit system

The International System of Units (SI) is a metric system giving a fully coherent set of units for science, technology and engineering, involving no conversion factors. The starting point is the selection and definition of a minimum set of independent 'base' units. From these, 'derived' units are obtained by forming products or quotients in various combinations, again without numerical factors. For convenience, certain combinations are given shortened names. A single SI unit of energy (joule = kilogram metre-squared per second-squared) is, for example, applied to energy of any kind, whether it be kinetic, potential, electrical, thermal, chemical . . . , thus unifying usage throughout science and technology.

The SI system has seven *base* units, and two *supplementary* units of angle. Combinations of these are *derived* for all other units. Each physical quantity has a quantity symbol (e.g. m for mass, P for power) that represents it in physical equations, and a unit symbol (e.g. kg for kilogram, W for watt) to indicate its SI unit of measure.

1.1.1 Base units

Definitions of the seven base units have been laid down in the following terms. The quantity symbol is given in italic, the unit symbol (with its standard abbreviation) in roman type. As measurements become more precise, changes are occasionally made in the definitions.

Length: l, metre (m) The metre was defined in 1983 as the length of the path travelled by light in a vacuum during a time interval of 1/299 792 458 of a second.

Mass: m, kilogram (kg) The mass of the international prototype (a block of platinum preserved at the International Bureau of Weights and Measures, Sèvres).

Time: t, second (s) The duration of 9 192 631 770 periods of the radiation corresponding to the transition between the two hyperfine levels of the ground state of the caesium-133 atom.

Electric current: i, ampere (A) The current which, maintained in two straight parallel conductors of infinite length, of negligible circular cross-section and 1 m apart in vacuum, produces a force equal to 2×10^{-7} newton per metre of length.

Thermodynamic temperature: T, kelvin (K) The fraction 1/273.16 of the thermodynamic (absolute) temperature of the triple point of water.

Luminous intensity: I, candela (cd) The luminous intensity in the perpendicular direction of a surface of 1/600 000 m^2 of a black body at the temperature of freezing platinum under a pressure of 101 325 newton per square metre.

Amount of substance: Q, mole (mol) The amount of substance of a system which contains as many elementary entities as there are atoms in 0.012 kg of carbon-12. The elementary entity must be specified and may be an atom, a molecule, an ion, an electron . . . , or a specified group of such entities.

1.1.2 Supplementary units

Plane angle: α, β, . . ., radian (rad) The plane angle between two radii of a circle which cut off on the circumference of the circle an arc of length equal to the radius.

Solid angle: Ω, steradian (sr) The solid angle which, having its vertex at the centre of a sphere, cuts off an area of the surface of the sphere equal to a square having sides equal to the radius.

1.1.3 Notes

Temperature At zero K, bodies possess no thermal energy. Specified points (273.16 and 373.16 K) define the Celsius (centigrade) scale (0 and 100°C). In terms of *intervals*, 1°C = 1 K. In terms of *levels*, a scale Celsius temperature θ corresponds to (θ + 273.16) K.

Force The SI unit is the newton (N). A force of 1 N endows a mass of 1 kg with an acceleration of 1 m/s^2.

Weight The weight of a mass depends on gravitational effect. The standard weight of a mass of 1 kg at the surface of the earth is 9.807 N.

1.1.4 Derived units

All physical quantities have units derived from the base and supplementary SI units, and some of them have been given names for convenience in use. A tabulation of those of interest in electrical technology is appended to the list in *Table 1.1*.

Table 1.1 SI base, supplementary and derived units

Quantity	*Unit name*	*Derivation*	*Unit symbol*
Length	metre		m
Mass	kilogram		kg
Time	second		s
Electric current	ampere		A
Thermodynamic temperature	kelvin		K
Luminous intensity	candela		cd
Amount of substance	mole		mol
Plane angle	radian		rad
Solid angle	steradian		sr
Force	newton	kg m/s^2	N
Pressure, stress	pascal	N/m^2	Pa
Energy	joule	N m, W s	J
Power	watt	J/s	W
Electric charge, flux	coulomb	A s	C
Magnetic flux	weber	V s	Wb
Electric potential	volt	J/C	V
Magnetic flux density	tesla	Wb/m^2	T
Resistance	ohm	V/A	Ω
Inductance	henry	Wb/A, V s/A	H
Capacitance	farad	C/V, A s/V	F
Conductance	siemens	A/V	S
Frequency	hertz	s^{-1}	Hz
Luminous flux	lumen	cd sr	lm
Illuminance	lux	lm/m^2	lx
Radiation activity	becquerel	s^{-1}	Bq
Absorbed dose	gray	J/kg	Gy
Mass density	kilogram per cubic metre		kg/m^3
Dynamic viscosity	pascal-second		Pa s
Concentration	mole per cubic metre		mol/m^3
Linear velocity	metre per second		m/s
Linear acceleration	metre per second-squared		m/s^2
Angular velocity	radian per second		rad/s

cont'd

Table 1.1 (*continued*)

Quantity	*Unit name*	*Derivation*	*Unit symbol*
Angular acceleration	radian per second-squared		rad/s^2
Torque	newton metre		N m
Electric field strength	volt per metre		V/m
Magnetic field strength	ampere per metre		A/m
Current density	ampere per square metre		A/m^2
Resistivity	ohm metre		Ω m
Conductivity	siemens per metre		S/m
Permeability	henry per metre		H/m
Permittivity	farad per metre		F/m
Thermal capacity	joule per kelvin		J/K
Specific heat capacity	joule per kilogram kelvin		J/(kg K)
Thermal conductivity	watt per metre kelvin		W/(m K)
Luminance	candela per square metre		cd/m^2

Decimal multiples and submultiples of SI units are indicated by prefix letters as listed in *Table 1.2*. Thus, kA is the unit symbol for kiloampere, and μF that for microfarad. There is a preference in technology for steps of 10^3. Prefixes for the kilogram are expressed in terms of the gram: thus, 1000 kg = 1 Mg, not 1 kkg.

1.1.5 Auxiliary units

Some quantities are still used in special fields (such as vacuum physics, irradiation, etc.) having non-SI units. Some of these are given in *Table 1.3* with their SI equivalents.

1.1.6 Conversion factors

Imperial and other non-SI units still in use are listed in *Table 1.4*, expressed in the most convenient multiples or submultiples of the basic SI unit [ ] under classified headings.

1.1.7 CGS electrostatic and electromagnetic units

Although obsolescent, electrostatic and electromagnetic units (e.s.u., e.m.u.) appear in older works of reference. Neither system is ‘rationalised’, nor are the two mutually compatible. In e.s.u. the electric space constant is $\varepsilon_0 = 1$, in e.m.u. the magnetic space constant is $\mu_0 = 1$; but the SI units take account of the fact that $1/\sqrt{(\varepsilon_0 \mu_0)}$ is the velocity of electromagnetic wave propagation in free space. *Table 1.5* lists SI units with the equivalent number *n* of e.s.u. and e.m.u. Where these lack names, they are expressed as SI unit names with the prefix ‘st’ (‘electrostatic’) for e.s.u. and ‘ab’ (‘absolute’) for e.m.u. Thus, 1 V corresponds to $10^{-2}/3$ stV and to 10^8 abV, so that 1 stV = 300 V and 1 abV = 10^{-8} V.

1.2 Mathematics

Mathematical symbolism is set out in *Table 1.6*. This subsection gives trigonometric and hyperbolic relations, series (including Fourier series for a number of common wave forms), binary enumeration and a list of common derivatives and integrals.

Table 1.2 Decimal prefixes

10^{18} exa E	10^9 giga G	10^2 hecto h	10^{-3} milli m	10^{-12} pico p
10^{15} peta P	10^6 mega M	10^1 deca da	10^{-6} micro $\mu\varsigma$	10^{-15} femto f
10^{12} tera T	10^3 kilo k	10^{-1} deci d	10^{-9} nano n	10^{-18} atto a
		10^{-2} centi c		

Table 1.3 Auxiliary units

Quantity	*Symbol*	*SI*		*Quantity*	*Symbol*	*SI*	
Angle				Mass			
degree	(°)	π/180	rad	tonne	t	1000	kg
minute	(′)	—	—				
second	(″)	—	—	Nucleonics, Radiation			
				becquerel	Bq	1.0	s^{-1}
Area				gray	Gy	1.0	J/kg
are	a	100	m^2	curie	Ci	3.7×10^{10}	Bq
hectare	ha	0.01	km^2	rad	rd	0.01	Gy
barn	barn	10^{-28}	m^2	roentgen	R	2.6×10^{-4}	C/kg
Energy				Pressure			
erg	erg	0.1	μJ	bar	b	100	kPa
calorie	cal	4.186	J	torr	Torr	133.3	Pa
electron-volt	eV	0.160	aJ	Time			
gauss-oersted	Ga Oe	7.96	μJ/m^3	minute	min	60	s
Force				hour	h	3600	s
dyne	dyn	10	μN	day	d	86 400	s
Length				Volume			
Ångstrom	Å	0.1	μm	litre	l or L	1.0	dm^3

Table 1.4 Conversion factors

Length [m]	
1 mil	25.40 µm
1 in	25.40 mm
1 ft	0.3048 m
1 yd	0.9144 m
1 fathom	1.829 m
1 mile	1.6093 km
1 nautical mile	1.852 km
Area [m^2]	
1 circular mil	506.7 μm^2
1 in^2	645.2 mm^2
1 ft^2	0.0929 m^2
1 yd^2	0.8361 m^2
1 acre	4047 m^2
1 $mile^2$	2.590 km^2
Volume [m^3]	
1 in^3	16.39 cm^3
1 ft^3	0.0283 m^3
1 yd^3	0.7646 m^3
1 UKgal	4.546 dm^3
Velocity [m/s, rad/s] Acceleration [m/s^2, rad/s^2]	
1 ft/min	5.080 mm/s
1 in/s	25.40 mm/s
1 ft/s	0.3048 m/s
1 mile/h	0.4470 m/s
1 knot	0.5144 m/s
1 deg/s	17.45 mrad/s
1 rev/min	0.1047 rad/s
1 rev/s	6.283 rad/s
1 ft/s^2	0.3048 m/s^2
1 mile/h per s	0.4470 m/s^2
Mass [kg]	
1 oz	28.35 g
1 lb	0.454 kg
1 slug	14.59 kg
1 cwt	50.80 kg
1 UKton	1016 kg
Energy [J], Power [W]	
1 ft lbf	1.356 J
1 m kgf	9.807 J
1 Btu	1055 J
1 therm	105.5 kJ
1 hp h	2.685 MJ
1 kW h	3.60 MJ
1 Btu/h	0.293 W
1 ft lbf/s	1.356 W
1 m kgf/s	9.807 W
1 hp	745.9 W
Thermal quantities [W, J, kg, K]	
1 W/in^2	1.550 kW/m^2
1 Btu/(ft^2 h)	3.155 W/m^2
1 Btu/(ft^3 h)	10.35 W/m^3
1 Btu/(ft h °F)	1.731 W/(m K)
1 ft lbf/lb	2.989 J/kg
1 Btu/lb	2326 J/kg
1 Btu/ft^3	37.26 KJ/m^3
1 ft lbf/(lb °F)	5.380 J/(kg K)
1 Btu/(lb °F)	4.187 kJ/(kg K)
1 Btu/(ft^3 °F)	67.07 kJ/m^3 K
Density [kg/m, kg/m^3]	
1 lb/in	17.86 kg/m
1 lb/ft	1.488 kg/m
1 lb/yd	0.496 kg/m
1 lb/in^3	27.68 Mg/m^3
1 lb/ft^3	16.02 kg/m^3
1 ton/yd^3	1329 kg/m^3
Flow rate [kg/s, m^3/s]	
1 lb/h	0.1260 g/s
1 ton/h	0.2822 kg/s
1 lb/s	0.4536 kg/s
1 ft^3/h	7.866 cm^3/s
1 ft^3/s	0.0283 m^3/s
1 gal/h	1.263 cm^3/s
1 gal/min	75.77 cm^3/s
1 gal/s	4.546 dm^3/s
Force [N], Pressure [Pa]	
1 dyn	10.0 µN
1 kgf	9.807 N
1 ozf	0.278 N
1 lbf	4.445 N
1 tonf	9.964 kN
1 dyn/cm^2	0.10 Pa
1 lbf/ft^2	47.88 Pa
1 lbf/in^2	6.895 kPa
1 $tonf/ft^2$	107.2 kPa
1 $tonf/in^2$	15.44 MPa
1 kgf/m^2	9.807 Pa
1 kgf/cm^2	98.07 kPa
1 mmHg	133.3 Pa
1 inHg	3.386 kPa
1 inH_2O	149.1 Pa
1 ftH_2O	2.989 kPa
Torque [N m]	
1 ozf in	7.062 nN m
1 lbf in	0.113 N m
1 lbf ft	1.356 N m
1 tonf ft	3.307 kN m
1 kgf m	9.806 N m
Inertia [kg m^2] Momentum [kg m/s, kg m^2/s]	
1 oz in^2	0.018 g m^2
1 lb in^2	0.293 g m^2
1 lb ft^2	0.0421 kg m^2
1 slug ft^2	1.355 kg m^2
1 ton ft^2	94.30 kg m^2
1 lb ft/s	0.138 kg m/s
1 lb ft^2/s	0.042 kg m^2/s
Viscosity [Pa s, m^2/s]	
1 poise	9.807 Pa s
1 kgf s/m^2	9.807 Pa s
1 lbf s/ft^2	47.88 Pa s
1 lbf h/ft^2	172.4 kPa s
1 stokes	1.0 cm^2/s
1 in^2/s	6.452 cm^2/s
1 ft^2/s	929.0 cm^2/s
Illumination [cd, lm]	
1 lm/ft^2	10.76 lm/m^2
1 cd/ft^2	10.76 cd/m^2
1 cd/in^2	1550 cd/m^2

Table 1.5 Relation between SI, e.s. and e.m. units

Quantity	*SI unit*	*Equivalent number n of*			
		e.s.u.		*e.m.u.*	
Length	m	10^2	cm	10^2	cm
Mass	kg	10^3	g	10^3	g
Time	s	1	s	1	s
Force	N	10^5	dyn	10^5	dyn
Torque	N m	10^7	dyn cm	10^7	dyn cm
Energy	J	10^7	erg	10^7	erg
Power	W	10^7	erg/s	10^7	erg/s
Charge, electric flux	C	3×10^9	stC	10^{-1}	abC
density	C/m^2	3×10^5	stC/cm^2	10^{-5}	abC/cm^2
Potential, e.m.f.	V	$10^{-2}/3$	stV	10^8	abV
Electric field strength	V/m	$10^{-4}/3$	stV/cm	10^6	abV/cm
Current	A	3×10^9	stA	10^{-1}	abA
density	A/m^2	3×10^5	stA/cm^2	10^{-5}	abA/cm^2
Magnetic flux	Wb	$10^{-2}/3$	stWb	10^8	Mx
density	T	$10^{-6}/3$	stWb/cm^2	10^4	Gs
Mag. fd. strength	A/m	$12\pi \times 10^7$	stA/cm	$4\pi \times 10^{-3}$	Oe
M.M.F.	A	$12\pi \times 10^9$	stA	$4\pi \times 10^{-1}$	Gb
Resistivity	Ω m	$10^{-9}/9$	stΩ cm	10^{11}	abΩ cm
Conductivity	S/m	9×10^9	stS/cm	10^{-11}	abS/cm
Permeability (abs)	H/m	$10^{-13}/36\pi$ς	—	$10^7/4\pi$ς	—
Permittivity (abs)	F/m	$36\pi \times 10^9$	—	$4\pi \times 10^{-11}$	—
Resistance	Ω	$10^{-11}/9$	stΩ	10^9	abΩ
Conductance	S	9×10^{11}	stS	10^{-9}	abS
Inductance	H	$10^{-12}/9$	stH	10^9	cm
Capacitance	F	9×10^{11}	cm	9×10^{11}	abF
Reluctance	A/Wb	$36\pi \times 10^{11}$	—	$4\pi \times 10^{-8}$	Gb/Mx
Permeance	Wb/A	$10^{11}/36\pi$ς	—	$10^9/4\pi$ς	Mx/Gb

Gb = gilbert; Gs = gauss; Mx = maxwell; Oe = oersted.

1.2.1 Trigonometric relations

The trigonometric functions (sine, cosine, tangent, cosecant, secant, cotangent) of an angle θ are based on the circle, given by $x^2 + y^2 = h^2$. Let two radii of the circle enclose an angle θςand form the sector area $S_c = (\pi h^2)(\theta/2\pi)$ shown shaded in *Figure 1.1* (left): then θςcan be defined as $2S_c/h^2$. The *right-angled* triangle with sides h (hypotenuse), a (adjacent side) and p (opposite side) give ratios defining the trigonometric functions

$$\sin\theta = p/h \qquad \operatorname{cosec}\theta = 1/\sin\theta = h/p$$

$$\cos\theta = a/h \qquad \sec\theta = 1/\cos\theta = h/a$$

$$\tan\theta = p/a \qquad \operatorname{cotan}\theta = 1/\tan\theta = a/p$$

In *any* triangle (*Figure 1.1*, right) with angles, A, B and C at the corners opposite, respectively, to sides a, b and c, then $A + B + C = \pi$ rad (180°) and the following relations hold:

$$a = b\cos C + c\cos B$$

$$b = c\cos A + a\cos C$$

$$c = a\cos B + b\cos A$$

$$a/\sin A = b/\sin B = c/\sin C$$

$$a^2 = b^2 + c^2 + 2bc\cos A$$

$$(a+b)/(a-b) = (\sin A + \sin B)/(\sin A - \sin B)\Leftarrow$$

Other useful relationships are:

$$\sin(x \pm y) = \sin x \cdot \cos y \pm \cos x \cdot \sin y$$

$$\cos(x \pm y) = \cos x \cdot \cos y \mp \sin x \cdot \sin y$$

$$\tan(x \pm y) = (\tan x \cdot \tan y)/(1 \mp \tan x \cdot \tan y)\Leftarrow$$

$$\sin^2 x = \tfrac{1}{2}(1 - \cos 2x)\cos^2 x = -\tfrac{1}{2}(1 + \cos 2x)$$

$$\sin^2 x + \cos^2 x = 1 \qquad \sin^3 x = -\tfrac{1}{4}(3\sin x - \sin 3x)$$

$$\cos^3 x = \tfrac{1}{4}(3\cos x + \cos 3x)$$

$$\sin x \pm \sin y = 2\left[\begin{matrix}\cos\\ \sin\end{matrix}\,\tfrac{1}{2}(x-y) \cdot \begin{matrix}\sin\\ \Leftarrow\\ \cos\end{matrix}\,\tfrac{1}{2}(x+y)\right]\Bigg]$$

$$\cos x \pm \cos y = -2\left[\begin{matrix}\cos\\ \sin\end{matrix}\,\tfrac{1}{2}(x-y) \cdot \begin{matrix}\sin\\ \Leftarrow\\ \cos\end{matrix}\,\tfrac{1}{2}(x+y)\right]\Bigg]$$

$$\tan x \pm \tan y = \sin(x \pm y)/\cos x \cdot \cos y$$

$$\sin^2 x - \sin^2 y = \sin(x+y) \cdot \sin(x-y)\Leftarrow$$

$$\cos^2 x - \cos^2 y = -\sin(x+y) \cdot \sin(x-y)\Leftarrow$$

$$\cos^2 x - \sin^2 y = \cos(x+y) \cdot \cos(x-y)\Leftarrow$$

$$\mathrm{d}(\sin x)/\mathrm{d}x = \cos x \qquad \int \sin x \cdot \mathrm{d}x = -\cos x + k$$

$$\mathrm{d}(\cos x)/\mathrm{d}x = -\sin x \qquad \int \cos x \cdot \mathrm{d}x = \sin x + k$$

$$\mathrm{d}(\tan x)/\mathrm{d}x = \sec^2 x \qquad \int \tan x \cdot \mathrm{d}x = -\ln|\cos x| + k$$

Values of sin θ, cos θ and tan θ for $0°\Leftarrow < \theta < 90°$ ⇐(or $0 < \theta$ ς < 1.571 rad) are given in *Table 1.7* as a check list, as they can generally be obtained directly from calculators.

Table 1.6 Mathematical symbolism

Term	*Symbol*
Base of natural logarithms	e (= 2.718 28 . . .)
Complex number	$\mathbf{C} = A + jB = C \exp(j\theta)$ $= C \angle \theta$
argument; modulus	$\arg \mathbf{C} = \theta$; $\operatorname{mod} \mathbf{C} = C$
conjugate	$\mathbf{C}^* = A - jB = C \exp(-j\theta)$ $= C \angle -\theta$
real part; imaginary part	$\operatorname{Re} \mathbf{C} = A$; $\operatorname{Im} \mathbf{C} = B$
Co-ordinates	
cartesian	x, y, z
cylindrical; spherical	r, ϕ, z; r, θ, ϕ
Function of x	
general	$f(x)$, $g(x)$, $F(x)$
Bessel	$J_n(x)$
circular	$\sin x$, $\cos x$, $\tan x \ldots$
inverse	$\arcsin x$, $\arccos x$, $\arctan x \ldots$
differential	dx
partial	∂x
exponential	$\exp(x)$
hyperbolic	$\sinh x$, $\cosh x$, $\tanh x \ldots$
inverse	$\operatorname{arsinh} x$, $\operatorname{arcosh} x$, $\operatorname{artanh} x \ldots$
increment	Δx, δx
limit	$\lim x$
logarithm	
base b	$\log_b x$
common; natural	$\lg x$; $\ln x$ (or $\log x$; $\log_e x$)
Matrix	A, B
complex conjugate	A^*, B^*
product	AB
square, determinant	$\det A$
inverse	A^{-1}
transpose	A^t
unit	I
Operator	
Heaviside	p ($\equiv d/dt$)
impulse function	$\delta(t)$
Laplace $L[f(t)] = F(s)$	s ($= \sigma + j\omega$)
nabla, del	∇
rotation $\pi/2$ rad;	j
$2\pi/3$ rad	h
step function	$H(t)$, $u(t)$
Vector	**A, a, B, b**
curl of A	curl **A**, $\nabla \times \mathbf{A}$
divergence of A	div **A**, $\nabla \cdot \mathbf{A}$
gradient of ϕ	grad ϕ, $\nabla \phi$
product: scalar; vector	$\mathbf{A} \cdot \mathbf{B}$; $\mathbf{A} \times \mathbf{B}$
units in cartesian axes	**i, j, k**

Table 1.7 Trigonometric functions of θ

θ		*sin* θ	*cos* θ	*tan* θ
deg	rad			
0	0.0	0.0	1.0	0.0
5	0.087	0.087	0.996	0.087
10	0.175	0.174	0.985	0.176
15	0.262	0.259	0.966	0.268
20	0.349	0.342	0.940	0.364
25	0.436	0.423	0.906	0.466
30	0.524	0.500	0.866	0.577
35	0.611	0.574	0.819	0.700
40	0.698	0.643	0.766	0.839
45	0.766	0.707	0.707	1.0
50	0.873	0.766	0.643	1.192
55	0.960	0.819	0.574	1.428
60	1.047	0.866	0.500	1.732
65	1.134	0.906	0.423	2.145
70	1.222	0.940	0.342	2.747
75	1.309	0.966	0.259	3.732
80	1.396	0.985	0.174	5.671
85	1.484	0.996	0.097	11.43
90	1.571	1.0	0.0	∞

1.2.2 Exponential and hyperbolic relations

Exponential functions For a positive datum ('real') number u, the exponential functions $\exp(u)$ and $\exp(-u)$ are given by the summation to infinity of the series

$$\exp(\pm u) = 1 \pm u + u^2/2! \pm u^3/3! + u^4/4! \pm \cdots$$

with $\exp(+u)$ increasing and $\exp(-u)$ decreasing at a rate proportional to u.

If $u = 1$, then

$$\exp(+1) = 1 + 1 + 1/2 + 1/6 + 1/24 + \cdots = e = 2.718 \cdots$$

$$\exp(-1) = 1 - 1 + 1/2 - 1/6 + 1/24 - \cdots = 1/e = 0.368 \cdots$$

In the electrical technology of transients, u is most commonly a negative function of time t given by $u = -(t/T)$. It then has the graphical form shown in *Figure 1.2* (left) as a time dependent variable. With an initial value k, i.e. $y = k \exp(-t/T)$, the rate of reduction with time is $dy/dt = -(k/T)\exp(-t/T)$. The initial rate at $t = 0$ is $-k/T$. If this rate were maintained, y would reach zero at $t = T$, defining the *time constant* T. Actually, after time T the value of y is $k \exp(-t/T) = k \exp(-1) = 0.368k$. Each successive interval T decreases y by the factor 0.368. At a time $t = 4.6T$ the value of y is $0.01k$, and at $t = 6.9T$ it is $0.001k$.

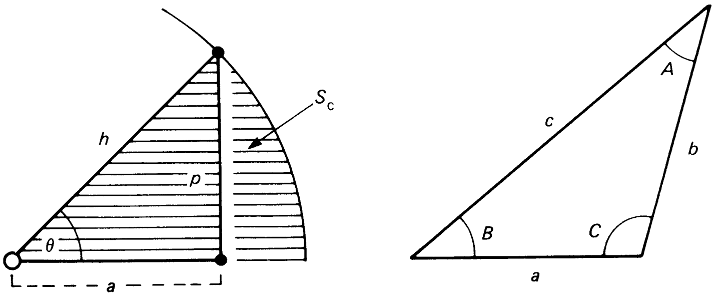


Figure 1.1 Trigonometric relations

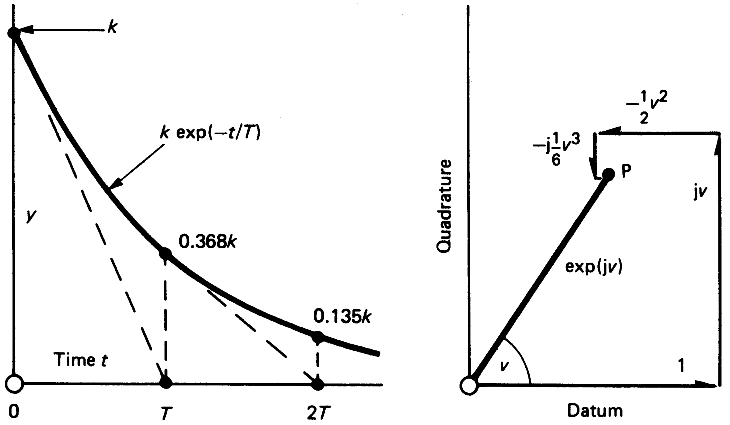


Figure 1.2 Exponential relations

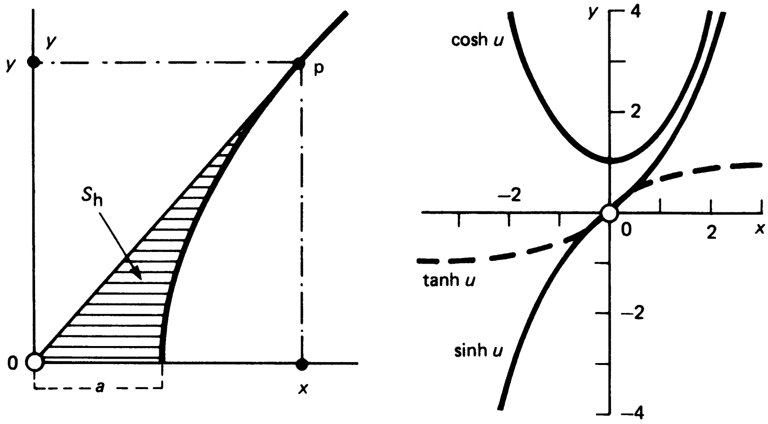


Figure 1.3 Hyperbolic relations

If u is a quadrature ('imaginary') number $\pm$jv, then

$$\exp(\pm \mathrm{j}v) = 1 \pm \mathrm{j}v - v^2/2! \mp \mathrm{j}v^3/3! + v^4/4! \pm \ldots$$

because $\mathrm{j}^2 = -1$, $\mathrm{j}^3 = -\mathrm{j}1$, $\mathrm{j}^4 = +1$, etc. *Figure 1.2* (right) shows the summation of the first five terms for exp(j1), i.e.

$$\exp(\mathrm{j}1) = 1 + \mathrm{j}1 - 1/2 - \mathrm{j}1/6 + 1/24$$

a complex or expression converging to a point P. The length OP is unity and the angle of OP to the datum axis is, in fact, 1 rad. In general, exp(jv) is equivalent to a shift by $\angle v$ rad. It follows that $\exp(\pm\mathrm{j}v) = \cos v \pm \mathrm{j}\sin v$, and that

$$\exp(\mathrm{j}v) + \exp(-\mathrm{j}v) = 2\cos v \quad \exp(\mathrm{j}v) - \exp(-\mathrm{j}v) = \mathrm{j}2\sin v$$

For a complex number $(u + \mathrm{j}v)$, then

$$\exp(u + \mathrm{j}v) = \exp(u) \cdot \exp(\mathrm{j}v) = \exp(u) \cdot \angle v$$

Hyperbolic functions A point P on a rectangular hyperbola $(x/a)^2 - \Leftarrow (y/a)^2 = 1$ defines the hyperbolic 'sector' area $S_\mathrm{h} = \frac{1}{2}a^2 \ln[(x/a - (y/a)]$ shown shaded in *Figure 1.3* (left). By analogy with $\theta \Leftarrow 2S_c/h^2$ for the trigonometrical angle θ, the hyperbolic entity (not an angle in the ordinary sense) is $u = 2S_\mathrm{h}/a^2$, where a is the major semi-axis. Then the hyperbolic functions of u for point P are:

$$\sinh u = y/a \qquad \mathrm{cosech}\, u = a/y$$
$$\cosh u = x/a \qquad \mathrm{sech}\, u = a/x$$
$$\tanh u = y/x \qquad \coth u = x/y$$

The principal relations yield the curves shown in the diagram (right) for values of u between 0 and 3. For higher values sinh u approaches $\pm$cosh u, and tanh u becomes asymptotic to ± 1. Inspection shows that $\cosh(-u) = \cosh u$, $\sinh(-u) = -\sinh u$ and $\cosh^2 u - \sinh^2 u = 1$.

The hyperbolic functions can also be expressed in the exponential form through the series

$$\cosh u = 1 + u^2/2! + u^4/4! + u^6/6! + \cdots \Leftarrow$$
$$\sinh u = u + u^3/3! + u^5/5! + u^7/7! + \cdots \Leftarrow$$

so that

$$\cosh u = \tfrac{1}{2}[\exp(u) + \exp(-u)] \Leftarrow \sinh u = \tfrac{1}{2}[\exp(u) - \exp(-u)]$$
$$\cosh u + \sinh u = \exp(u) \Leftarrow \qquad \cosh u - \sinh u = \exp(-u) \Leftarrow$$

Other relations are:

$$\sinh u + \sinh v = 2\sinh\tfrac{1}{2}(u + v) \cdot \cosh\tfrac{1}{2}(u - v)$$
$$\cosh u + \cosh v = 2\cosh\tfrac{1}{2}(u + v) \cdot \cosh\tfrac{1}{2}(u - v)$$
$$\cosh u - \cosh v = 2\sinh\tfrac{1}{2}(u + v) \cdot \sinh\tfrac{1}{2}(u - v)$$
$$\sinh(u \pm v) = \sinh u \cdot \cosh v \pm \cosh u \cdot \sinh v$$
$$\cosh(u \pm v) = \cosh u \cdot \cosh v \pm \sinh u \cdot \sinh v$$
$$\tanh(u \pm v) = (\tanh u \pm \tanh v)/(1 \pm \tanh u \cdot \tanh v) \Leftarrow$$

Table 1.8 Exponential and hyperbolic functions

u	*exp*(*u*)	*exp*(−*u*)	*sinh u*	*cosh u*	*tanh u*
0.0	1.0	1.0	0.0	1.0	0.0
0.1	1.1052	0.9048	0.1092	1.0050	0.0997
0.2	1.2214	0.8187	0.2013	1.0201	0.1974
0.3	1.3499	0.7408	0.3045	1.0453	0.2913
0.4	1.4918	0.6703	0.4108	1.0811	0.3799
0.5	1.6487	0.6065	0.5211	1.1276	0.4621
0.6	1.8221	0.5488	0.6367	1.1855	0.5370
0.7	2.0138	0.4966	0.7586	1.2552	0.6044
0.8	2.2255	0.4493	0.8881	1.3374	0.6640
0.9	2.4596	0.4066	1.0265	1.4331	0.7163
1.0	2.7183	0.3679	1.1752	1.5431	0.7616
1.2	3.320	0.3012	1.5095	1.8107	0.8337
1.4	4.055	0.2466	1.9043	2.1509	0.8854
1.6	4.953	0.2019	2.376	2.577	0.9217
1.8	6.050	0.1653	2.942	3.107	0.9468
2.0	7.389	0.1353	3.627	3.762	0.9640
2.303	10.00	0.100	4.950	5.049	0.9802
2.5	12.18	0.0821	6.050	6.132	0.9866
2.75	15.64	0.0639	7.789	7.853	0.9919
3.0	20.09	0.0498	10.02	10.07	0.9951
3.5	33.12	0.0302	16.54	16.57	0.9982
4.0	54.60	0.0183	27.29	27.31	0.9993
4.5	90.02	0.0111	45.00	45.01	0.9998
4.605	100.0	0.0100	49.77	49.80	0.9999
5.0	148.4	0.0067	74.20	74.21	0.9999
5.5	244.7	0.0041	122.3	cosh *u* = sinh *u* = $\frac{1}{2}$ exp(*u*)	tanh *u* = 1.0
6.0	403.4	0.0025	201.7		
6.908	1000	0.0010	500		

$$\sinh(u \pm jv) = (\sinh u \cdot \cos v) \pm j(\cosh u \cdot \sin v)$$

$$\cosh(u \pm jv) = (\cosh u \cdot \cos v) \pm j(\sinh u \cdot \sin v)$$

$$d(\sinh u)/du = \cosh u \qquad \int \sinh u \cdot du = \cosh u$$

$$d(\cosh u)/du = \sinh u \qquad \int \cosh u \cdot du = \sinh u$$

Exponential and hyperbolic functions of *u* between zero and 6.908 are listed in *Table 1.8*. Many calculators can give such values directly.

1.2.3 Bessel functions

Problems in a wide range of technology (e.g. in eddy currents, frequency modulation, etc.) can be set in the form of the Bessel equation

$$\frac{d^2y}{dx^2} + \frac{1}{x} \cdot \frac{dy}{dx} + \left[1 - \frac{n^2}{x^2}\right] y = 0$$

and its solutions are called Bessel functions of order *n*. For $n = 0$ the solution is

$$J_0(x) = 1 - (x^2/2^2) + (x^4/2^2 \cdot 4^2) - (x^6/2^2 \cdot 4^2 \cdot 6^2) + \cdots$$

and for $n = 1, 2, 3 \ldots$

$$J_n(x) = \frac{x^n}{2^n n!}\left[1 - \frac{x^2}{2(2n+2)} + \frac{x^4}{2 \cdot 4(2n+2)(2n+4)} - \cdots\right]$$

Table 1.9 gives values of $J_n(x)$ for various values of *n* and *x*.

1.2.4 Series

Factorials In several of the following the factorial (*n*!) of integral numbers appears. For *n* between 2 and 10 these are

2! = 2	1/2! = 0.5
3! = 6	1/3! = 0.1667
4! = 24	$1/4! = 0.417 \times 10^{-1}$
5! = 120	$1/5! = 0.833 \times 10^{-2}$
6! = 720	$1/6! = 0.139 \times 10^{-2}$
7! = 5 040	$1/7! = 0.198 \times 10^{-3}$
8! = 40 320	$1/8! = 0.248 \times 10^{-4}$
9! = 362 880	$1/9! = 0.276 \times 10^{-5}$
10! = 3 628 800	$1/10! = 0.276 \times 10^{-6}$

Progression

Arithmetic $a + (a + d) + (a + 2d) + \cdots + [a + (n-1)d]$
$= \frac{1}{2} n$ (sum of 1st and *n*th terms)

Geometric $a + ar + ar^2 + \cdots + ar^{n-1} = a(1-r^n)/(1-r)$

Trigonometric See Section 1.2.1.

Exponential and hyperbolic See Section 1.2.2.

Binomial

$$(1 \pm x)^n = 1 \pm nx + \frac{n(n-1)}{2!}x^2 \pm \frac{n(n-1)(n-2)}{3!}x^3 + \cdots + (-1)^r \frac{n!}{r!(n-r)!}x^r + \cdots$$

$$(a \pm x)^n = a^n[1 \pm (x/a)]^n$$

Binomial coefficients $n!/[r!\,(n-r)!]$ are tabulated:

Term $r =$	0	1	2	3	4	5	6	7	8	9	10
$n = 1$	1	1									
2	1	2	1								
3	1	3	3	1							
4	1	4	6	4	1						
5	1	5	10	10	5	1					
6	1	6	15	20	15	6	1				
7	1	7	21	35	35	21	7	1			
8	1	8	28	56	70	56	28	8	1		
9	1	9	36	84	126	126	84	36	9	1	
10	1	10	45	120	210	252	210	120	45	10	1

Power If there is a power series for a function $f(h)$, it is given by

$$f(h) = f(0) + hf^{(\mathrm{i})}(0) + (h^2/2!)f^{(\mathrm{ii})}(0) + (h^3/3!)f^{(\mathrm{iii})}(0) + \cdots + (h^r/r!)f^{(r)}(0) + \cdots \qquad (\mathit{Maclaurin})$$

$$f(x+h) = f(x) + hf^{(\mathrm{i})}(x) + (h^2/2!)f^{(\mathrm{ii})}(x) + \cdots + (h^r/r!)f^{(r)}(x) + \cdots \qquad (\mathit{Taylor})$$

Permutation, combination

$${}^nP_r = n(n-1)(n-2)(n-3)\ldots(n-r+1) = n!/(n-r)!$$

$${}^nC_r = (1/r!)[n(n-1)(n-2)(n-3)\ldots(n-r+1)] = n!/r!(n-r)!$$

Bessel See Section 1.2.3.
Fourier See Section 1.2.5.

1.2.5 Fourier series

A univalued periodic wave form $f(\theta)$ of period 2π is represented by a summation in general of sine and cosine waves of fundamental period 2π and of integral harmonic orders n $(= 2, 3, 4, \ldots)$ as

$$f(\theta) = c_0 + a_1 \cos\theta + a_2 \cos 2\theta + \cdots + a_n \cos n\theta + \cdots + b_1 \sin\theta + b_2 \sin 2\theta + \cdots + b_n \sin n\theta + \cdots$$

The mean value of $f(\theta)$ over a full period 2π is

$$c_0 = \frac{1}{2\pi}\int_0^{2\pi} f(\theta)\cdot \mathrm{d}\theta$$

and the harmonic-component amplitudes a and b are

$$a_n = \frac{1}{\pi}\int_0^{2\pi} f(\theta)\cdot \cos n\theta\,\mathrm{d}\theta, \quad b_n = \frac{1}{\pi}\int_0^{2\pi} f(\theta)\cdot \sin n\theta\,\mathrm{d}\theta$$

Table 1.10 gives for a number of typical wave forms the harmonic series in square brackets, preceded by the mean value c_0 where it is not zero.

1.2.6 Derivatives and integrals

Some basic forms are listed in *Table 1.11*. Entries in a given column are the integrals of those in the column to its left and the derivatives of those to its right. Constants of integration are omitted.

1.2.7 Laplace transforms

Laplace transformation is a method of deriving the response of a system to any stimulus. The system has a basic equation of behaviour, and the stimulus is a pulse, step, sine wave or other variable with time. Such a response involves integration: the Laplace transform method removes integration difficulties, as tables are available for the direct solution of a great variety of problems. The process is analogous to evaluation (for example) of $y = 2.1^{3.6}$ by *transformation* into a logarithmic form $\log y = 3.6 \times \log(2.1)$, and a subsequent *inverse transformation* back into arithmetic by use of a table of antilogarithms.

The Laplace transform (L.t.) of a time-varying function $f(t)$ is

$$\mathrm{L}[f(t)] = \mathrm{F(s)} = \int_0^{\infty} \exp(-st)\cdot f(t)\cdot \mathrm{d}t$$

and the inverse transformation of F(s) to give $f(t)$ is

$$\mathrm{L}^{-1}[\mathrm{F(s)}] = f(t) = \lim \frac{1}{2\pi}\int_{c-\mathrm{j}\omega}^{c+\mathrm{j}\omega} \exp(st)\cdot \mathrm{F(s)}\cdot \mathrm{ds}$$

The process, illustrated by the response of a current $i(t)$ in an electrical network of impedance z to a voltage $v(t)$ applied at $t = 0$, is to write down the transform equation

$$\mathrm{I(s)} = \mathrm{V(s)}/\mathrm{Z(s)}$$

where I(s) is the L.t. of the current $i(t)$, V(s) is the L.t. of the voltage $v(t)$, and Z(s) is the *operational* impedance. Z(s) is obtained from the network resistance R, inductance L and capacitance C by leaving R unchanged but replacing L by Ls and C by $1/C$s. The process is equivalent to writing the network impedance for a steady state frequency ω and then replacing $\mathrm{j}\omega$ by s. V(s) and Z(s) are polynomials in s: the quotient V(s)/Z(s) is reduced algebraically to a form recognisable in the transform table. The resulting current/time relation $i(t)$ is read out: it contains the complete solution. However, if at $t = 0$ the network has initial energy (i.e. if currents flow in inductors or charges are stored in capacitors), the equation becomes

$$\mathrm{I(s)} = [\mathrm{V(s)} + \mathrm{U(s)}]/\mathrm{Z(s)}$$

where U(s) contains such terms as LI_0 and $(1/\mathrm{s})V_0$ for the inductors or capacitors at $t = 0$.

A number of useful transform pairs is listed in *Table 1.12*.

1.2.8 Binary numeration

A number N in decimal notation can be represented by an ordered set of binary digits $a_n, a_{n-2}, \ldots, a_2, a_1, a_0$ such that

$$N = 2^n a_n + 2^{n-1} a_{n-1} + \cdots + 2a_1 + a_0$$

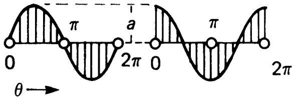
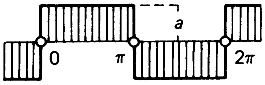
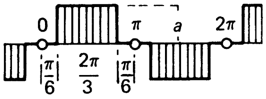
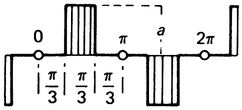
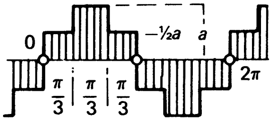
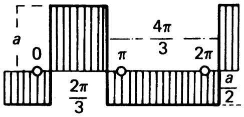
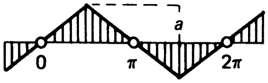
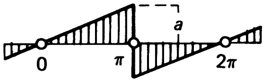
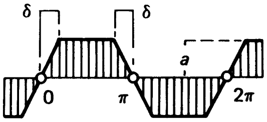
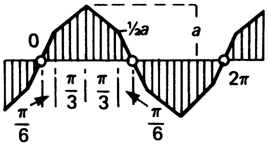
Decimal	1	2	3	4	5	6	7	8	9	10	100
Binary	1	10	11	100	101	110	111	1000	1001	1010	1100100

Table 1.9 Bessel functions $J_n(x)$

n	$J_n(1)$	$J_n(2)$	$J_n(3)$	$J_n(4)$	$J_n(5)$	$J_n(6)$	$J_n(7)$	$J_n(8)$	$J_n(9)$	$J_n(10)$	$J_n(11)$	$J_n(12)$	$J_n(13)$	$J_n(14)$	$J_n(15)$
0	0.7652	0.2239	−0.2601	−0.3971	−0.1776	0.1506	0.3001	0.1717	−0.0903	−0.2459	−0.1712	0.0477	0.2069	0.1711	−0.0142
1	0.4401	0.5767	0.3391	−0.0660	−0.3276	−0.2767	−0.0047	0.2346	0.2453	0.0435	−0.1768	−0.2234	−0.0703	0.1334	0.2051
2	0.1149	0.3528	0.4861	0.3641	0.0466	−0.2429	−0.3014	−0.1130	0.1448	0.2546	0.1390	−0.0849	−0.2177	−0.1520	0.0416
3	0.0196	0.1289	0.3091	0.4302	0.3648	0.1148	−0.1676	−0.2911	−0.1809	0.0584	0.2273	0.1951	0.0033	−0.1768	−0.1940
4	—	0.0340	0.1320	0.2811	0.3912	0.3567	0.1578	−0.1054	−0.2655	−0.2196	−0.0150	0.1825	0.2193	0.0762	−0.1192
5	—	—	0.0430	0.1321	0.2611	0.3621	0.3479	0.1858	−0.0550	−0.2341	−0.2383	−0.0735	0.1316	0.2204	0.1305
6	—	—	0.0114	0.0491	0.1310	0.2458	0.3392	0.3376	0.2043	−0.0145	−0.2016	−0.2437	−0.1180	0.0812	0.2061
7	—	—	—	0.0152	0.0534	0.1296	0.2336	0.3206	0.3275	0.2167	0.0184	−0.1703	−0.2406	−0.1508	0.0345
8	—	—	—	—	0.0184	0.0565	0.1280	0.2235	0.3051	0.3179	0.2250	0.0451	−0.1410	−0.2320	−0.1740
9	—	—	—	—	—	0.0212	0.0589	0.1263	0.2149	0.2919	0.3089	0.2304	0.0670	−0.1143	−0.2200
10	—	—	—	—	—	—	0.0235	0.0608	0.1247	0.2075	0.2804	0.3005	0.2338	0.0850	−0.0901
11	—	—	—	—	—	—	—	0.0256	0.0622	0.1231	0.2010	0.2704	0.2927	0.2357	0.0999
12	—	—	—	—	—	—	—	—	0.0274	0.0634	0.1216	0.1953	0.2615	0.2855	0.2367
13	—	—	—	—	—	—	—	—	0.0108	0.0290	0.0643	0.1201	0.1901	0.2536	0.2787
14	—	—	—	—	—	—	—	—	—	0.0119	0.0304	0.0650	0.1188	0.1855	0.2464
15	—	—	—	—	—	—	—	—	—	—	0.0130	0.0316	0.0656	0.1174	0.1813

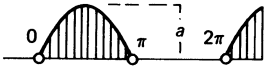
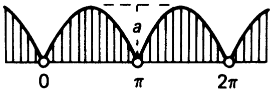
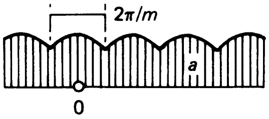
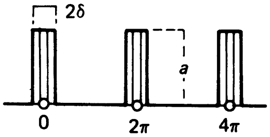
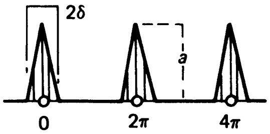
Values below 0.01 not tabulated.

Table 1.10 Fourier series

	Wave form	*Series*
1		Sine: $a \sin \theta$ Cosine: $a \sin \theta$
2		Square: $a\frac{4}{\pi}\left[\frac{\sin\theta}{1}+\frac{\sin 3\theta}{3}+\frac{\sin 5\theta}{5}+\frac{\sin 7\theta}{7}+\cdots\right]$
5		Rectangular block: $a\frac{2\sqrt{3}}{\pi}\left[\frac{\sin\theta}{1}-\frac{\sin 5\theta}{5}-\frac{\sin 7\theta}{7}+\frac{\sin 11\theta}{11}+\frac{\sin 13\theta}{13}-\frac{\sin 17\theta}{17}-\cdots\right]$
4		Rectangular block: $a\frac{4}{\pi}\left[\frac{\sin\theta}{2\cdot 1}-\frac{\sin 3\theta}{3}+\frac{\sin 5\theta}{2\cdot 5}+\frac{\sin 7\theta}{2\cdot 7}-\frac{\sin 9\theta}{9}+\frac{\sin 11\theta}{2\cdot 11}+\frac{\sin 13\theta}{2\cdot 13}-\frac{\sin 15\theta}{15}+\frac{\sin 17\theta}{2\cdot 17}+\cdots\right]$
4a		Stepped rectangle: $a\frac{3}{\pi}\left[\frac{\sin\theta}{1}+\frac{\sin 5\theta}{5}+\frac{\sin 7\theta}{7}+\frac{\sin 11\theta}{11}+\frac{\sin 13\theta}{13}+\frac{\sin 17\theta}{17}+\cdots\right]$
3		Asymmetric rectangle: $a\frac{3\sqrt{3}}{2\pi}\left[\frac{\sin\theta}{1}-\frac{\sin 5\theta}{5}-\frac{\sin 7\theta}{7}+\frac{\sin 11\theta}{11}+\frac{\sin 13\theta}{13}-\cdots -\frac{\cos 2\theta}{2}-\frac{\cos 4\theta}{4}+\frac{\cos 8\theta}{8}+\frac{\cos 10\theta}{10}-\cdots\right]$
		Triangle: $a\frac{8}{\pi^2}\left[\frac{\sin\theta}{1}-\frac{\sin 3\theta}{9}+\frac{\sin 5\theta}{25}-\frac{\sin 7\theta}{49}+\frac{\sin 9\theta}{81}-\frac{\sin 11\theta}{121}+\cdots\right]$
		Sawtooth: $a\frac{3}{\pi}\left[\frac{\sin\theta}{1}-\frac{\sin 2\theta}{2}+\frac{\sin 3\theta}{3}-\frac{\sin 4\theta}{4}+\frac{\sin 5\theta}{5}-\cdots\right]$
		Trapeze: $a\frac{4}{\pi\delta}\left[\frac{\sin\delta\sin\theta}{1}+\frac{\sin 3\delta\sin 3\theta}{9}+\frac{\sin 5\delta\sin 5\theta}{25}+\cdots\right]$ $a\frac{6\sqrt{3}}{\pi^2}\left[\frac{\sin\theta}{1}-\frac{\sin 5\theta}{25}+\frac{\sin 7\theta}{49}-\frac{\sin 11\theta}{121}+\cdots\right]$ for $\delta=\pi/3$
		Trapeze-triangle: $a\frac{9}{\pi^2}\left[\frac{\sin\theta}{1}+\frac{\sin 5\theta}{25}-\frac{\sin 7\theta}{49}+\frac{\sin 11\theta}{121}-\frac{\sin 13\theta}{169}+\cdots\right]$

cont'd

Table 1.10 (*continued*)

Wave form	*Series*
	Rectified sine (half-wave): $a\frac{1}{\pi}+a\frac{2}{\pi}\left[\frac{\pi\sin\theta}{4}-\frac{\cos 2\theta}{1\cdot 3}-\frac{\cos 4\theta}{3\cdot 5}-\frac{\cos 6\theta}{5\cdot 7}-\cdots\right]$
	Rectified sine (full-wave): $a\frac{2}{\pi}-a\frac{4}{\pi}\left[\frac{\cos 2\theta}{1\cdot 3}+\frac{\cos 4\theta}{3\cdot 5}+\frac{\cos 6\theta}{5\cdot 7}+\frac{\cos 8\theta}{7\cdot 9}+\cdots\right]$
	Rectified sine (*m*-phase): $a\frac{m}{\pi}\sin\frac{\pi}{m}+a\frac{2m}{\pi}\sin\frac{\pi}{m}\left[\frac{\cos m\theta}{m^2-1}-\frac{\cos 2m\theta}{4m^2-1}+\frac{\cos 3m\theta}{9m^2-1}-\cdots\right]$
	Rectangular pulse train: $a\frac{\delta}{\pi}+a\frac{2}{\pi}\left[\frac{\sin\delta\cos\theta}{1}+\frac{\sin 2\delta\cos 2\theta}{2}+\frac{\sin 3\delta\cos 3\theta}{3}+\cdots\right]$ $a\frac{\delta}{\pi}+a\frac{2\delta}{\pi}\left[\frac{\cos\theta}{1}+\frac{\cos 2\theta}{2}+\frac{\cos 3\theta}{3}+\cdots\right]$ for $\delta\ll\pi$
	Triangular pulse train: $a\frac{\delta}{2\pi}+a\frac{4}{\pi\delta}\left[\frac{\sin^2(\frac{1}{2}\delta)}{1}\cos\theta+\frac{\sin^2 2(\frac{1}{2}\delta)}{4}\cos 2\theta+\frac{\sin^2 3(\frac{1}{2}\delta)}{9}\cos 3\theta+\cdots\right]$ $a\frac{\delta}{2\pi}+a\frac{\delta}{\pi}[\cos\theta+\cos 2\theta+\cos 3\theta+\cdots]$ for $\delta\ll\pi$

where the *a*s have the values either 1 or 0. Thus, if $N=19$,

$$19=16+2+1=(2^4)1+(2^3)0+(2^2)0+(2^1)1+(2^0)1=10011$$

in binary notation. The rules of addition and multiplication are

$$0+0=0,\ 0+1=1,\ 1+1=10;\ 0\times 0=0,\ 0\times 1=0,\ 1\times 1=1$$

1.2.9 Power ratio

In communication networks the powers P_1 and P_2 at two specified points may differ widely as the result of amplification or attenuation. The power ratio P_1/P_2 is more convenient in logarithmic terms.

Neper [Np] This is the natural logarithm of a voltage or current ratio, given by

$$a=\ln(V_1/V_2)\quad\text{or}\quad a=\ln(I_1/I_2)\ \text{Np}$$

If the voltages are applied to, or the currents flow in, identical impedances, then the power ratio is

$$a=\ln(V_1/V_2)^2=2\ln(V_1/V_2)$$

and similarly for current.

Decibel [dB] The power gain is given by the common logarithm $\lg(P_1/P_2)$ in bel [B], or most commonly by $A=10\log(P_1/P_2)$ decibel [dB]. With again the proviso that the powers are developed in identical impedances, the power gain is

$$A=10\log(P_1/P_2)=10\log(V_1/V_2)^2=20\log(V_1/V_2)\ \text{dB}$$

Table 1.13 gives the power ratio corresponding to a gain *A* (in dB) and the related identical-impedance voltage (or current) ratios. Approximately, 3 dB corresponds to a power ratio of 2, and 6 dB to a power ratio of 4. The decibel equivalent of 1 Np is 8.69 dB.

1.2.10 Matrices and vectors

1.2.10.1 Definitions

If $a_{11}, a_{12}, a_{13}, a_{14}\ldots$ is a set of elements, then the rectangular array

$$\mathbf{A}=\begin{bmatrix} a_{11} & a_{12} & a_{13} & a_{14}\ldots a_{1n}\\ a_{21} & a_{22} & a_{23} & a_{24}\ldots a_{2n}\\ \\ a_{m1} & a_{m2} & a_{m3} & a_{m4}\ldots a_{mn}\end{bmatrix}$$

arranged in *m* rows and *n* columns is called an $(m\times n)$ *matrix*. If $m=n$ then **A** is *n-square*.

Table 1.11 Derivatives and integrals

$\mathrm{d}[f(x)]/\mathrm{d}x$	$f(x)$	$\int f(x)\cdot \mathrm{d}x$
1	x	$\frac{1}{2}x^2$
nx^{n-1}	$x^n \ (n \neq -1)$	$x^{n+1}/(n+1)$
$-1/x^2$	$1/x$	$\ln x$
$1/x$	$\ln x$	$x \ln x - x$
$\exp x$	$\exp x$	$\exp x$
$\cos x$	$\sin x$	$-\cos x$
$-\sin x$	$\cos x$	$\sin x$
$\sec^2 x$	$\tan x$	$\ln(\sec x)$
$-\mathrm{cosec}\, x \cdot \cot x$	$\mathrm{cosec}\, x$	$\ln(\tan \frac{1}{2}x)$
$\sec x \cdot \tan x$	$\sec x$	$\ln(\sec x + \tan x)$
$-\mathrm{cosec}^2 x$	$\cot x$	$\ln(\sin x)$
$1/\sqrt{(a^2-x^2)}$	$\arcsin(x/a)$	$x \arcsin(x/a) + \sqrt{(a^2-x^2)}$
$-1/\sqrt{(a^2-x^2)}$	$\arccos(x/a)$	$x \arccos(x/a) - \sqrt{(a^2-x^2)}$
$a/(a^2+x^2)$	$\arctan(x/a)$	$x \arctan(x/a) - \frac{1}{2}a \ln(a^2+x^2)$
$-a/x\sqrt{(x^2-a^2)}$	$\mathrm{arccosec}(x/a)$	$x\, \mathrm{arccosec}(x/a) + a \ln \mid x + \sqrt{(x^2-a^2)} \mid$
$a/x\sqrt{(x^2-a^2)}$	$\mathrm{arcsec}(x/a)$	$x\, \mathrm{arcsec}(x/a) - a \ln \mid x + \sqrt{(x^2-a^2)} \mid$
$-a/(a^2+x^2)$	$\mathrm{arccot}(x/a)$	$x\, \mathrm{arccot}(x/a) + \frac{1}{2}a \ln(a^2+x^2)$
$\cosh x$	$\sinh x$	$\cosh x$
$\sinh x$	$\cosh x$	$\sinh x$
$\mathrm{sech}^2 x$	$\tanh x$	$\ln(\cosh x)$
$-\mathrm{cosech}\, x \cdot \coth x$	$\mathrm{cosech}\, x$	$-\ln(\tanh \frac{1}{2}x)$
$-\mathrm{sech}\, x \cdot \tanh x$	$\mathrm{sech}\, x$	$2 \arctan(\exp x)$
$-\mathrm{cosech}^2 x$	$\coth x$	$\ln(\sinh x)$
$1/\sqrt{(x^2+1)}$	$\mathrm{arsinh}\, x$	$x\, \mathrm{arsinh}\, x - \sqrt{(1+x^2)}$
$1/\sqrt{(x^2-1)}$	$\mathrm{arcosh}\, x$	$x\, \mathrm{arcosh}\, x - \sqrt{(x^2-1)}$
$1/(1-x^2)$	$\mathrm{artanh}\, x$	$x\, \mathrm{artanh}\, x + \frac{1}{2}\ln(1-x^2)$
$-1/x\sqrt{(x^2+1)}$	$\mathrm{arcosech}\, x$	$x\, \mathrm{arcosech}\, x + \mathrm{arsinh}\, x$
$-1/x\sqrt{(1-x^2)}$	$\mathrm{arsech}\, x$	$x\, \mathrm{arsech}\, x + \arcsin x$
$1/(1-x^2)$	$\mathrm{arcoth}\, x$	$x\, \mathrm{arcoth}\, x + \frac{1}{2}\ln(x^2-1)$
$u\frac{\mathrm{d}v}{\mathrm{d}x} + v\frac{\mathrm{d}u}{\mathrm{d}x}$	$u(x) \cdot v(x)$	$uv - \int v \frac{\mathrm{d}u}{\mathrm{d}v}\, \mathrm{d}v$
$\frac{1}{v}\frac{\mathrm{d}u}{\mathrm{d}x} - \frac{u}{v^2}\frac{\mathrm{d}v}{\mathrm{d}x}$	$\frac{u(x)}{v(x)}$	—
$r \exp(xa) \times \sin(\omega x + \phi + \theta)$	$\exp(ax) \times \sin(\omega x + \phi)$	$(1/r)\exp(ax)\sin(\omega x + \phi - \theta)$ $r = \sqrt{(\omega^2 + a^2)}\ \theta = \arctan(\omega/a)$

An ordered set of elements $\mathbf{x} = [x_1, x_2, x_3 \ldots x_n]$ is called an *n-vector*.

An $(n \times 1)$ matrix is called a *column vector* and a $(1 \times n)$ matrix a *row vector*.

1.2.10.2 Basic operations

If $\mathbf{A} = (a_{rs})$, $\mathbf{B} = (b_{rs})$,

(i) *Sum* $\mathbf{C} = \mathbf{A} + \mathbf{B}$ is defined by $c_{rs} = a_{rs} + b_{rs}$, for $r = 1 \ldots m$; $s = 1 \ldots n$.

(ii) *Product* If $\mathbf{A}$ is an $(m \times q)$ matrix and $\mathbf{B}$ is a $(q \times n)$ matrix, then the product $\mathbf{C} = \mathbf{AB}$ is an $(m \times n)$ matrix defined by $(c_{rs}) = \Sigma_p a_{rp} b_{ps}$, $p = 1 \ldots q$; $r = 1 \ldots m$; $s = 1 \ldots n$. If $\mathbf{AB} = \mathbf{BA}$ then $\mathbf{A}$ and $\mathbf{B}$ are said to *commute*.

(iii) *Matrix-vector product* If $\mathbf{x} = [x_1 \ldots x_n]$, then $\mathbf{b} = \mathbf{Ax}$ is defined by $(b_r) = \Sigma_p\, a_{rp}\, x_p$, $p = 1 \ldots n$; $r = 1 \ldots m$.

(iv) *Multiplication of a matrix by a (scalar) element* If k is an element then $\mathbf{C} = k\mathbf{A} = \mathbf{A}k$ is defined by $(c_{rs}) = k(a_{rs})$.

(v) *Equality* If $\mathbf{A} = \mathbf{B}$, then $(a_{ij}) = (b_{ij})$, for $i = 1 \ldots n$; $j = 1 \ldots m$.

1.2.10.3 Rules of operation

(i) *Associativity* $\mathbf{A} + (\mathbf{B} + \mathbf{C}) = (\mathbf{A} + \mathbf{B}) + \mathbf{C}$, $\mathbf{A(BC)} = \mathbf{(AB)C} = \mathbf{ABC}$.

(ii) *Distributivity* $\mathbf{A(B + C)} = \mathbf{AB} + \mathbf{AC}$, $\mathbf{(B + C)A} = \mathbf{BA} + \mathbf{CA}$.

(iii) *Identity* If $\mathbf{U}$ is the $(n \times n)$ matrix (δij), $i, j = 1 \ldots n$, where $\delta ij = 1$ if $i = j$ and 0 otherwise, then $\mathbf{U}$ is the diagonal *unit matrix* and $\mathbf{A\,U} = \mathbf{A}$.

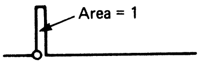
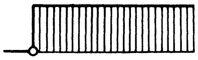
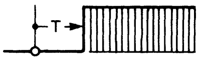
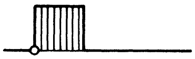
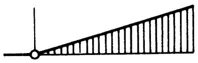
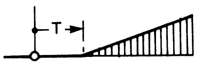
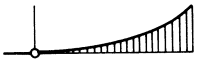
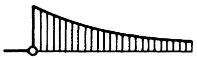
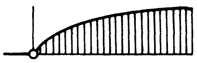
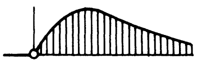
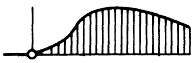
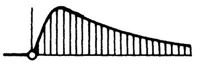
(iv) *Inverse* If the product $\mathbf{U} = \mathbf{AB}$ exists, then $\mathbf{B} = \mathbf{A}^{-1}$, the inverse matrix of $\mathbf{A}$. If both inverses $\mathbf{A}^{-1}$ and $\mathbf{B}^{-1}$ exist, then $(\mathbf{A\,B})^{-1} = \mathbf{B}^{-1}\,\mathbf{A}^{-1}$.

(v) *Transposition* The transpose of $\mathbf{A}$ is written as $\mathbf{A}^{\mathbf{T}}$ and is the matrix whose rows are the columns of $\mathbf{A}$. If the product $\mathbf{C} = \mathbf{AB}$ exists then $\mathbf{C}^{\mathbf{T}} = (\mathbf{AB})^{\mathbf{T}} = \mathbf{B}^{\mathbf{T}}\mathbf{A}^{\mathbf{T}}$.

(vi) *Conjugate* For $\mathbf{A} = (a_{rs})$, the congugate of $\mathbf{A}$ is denoted by $\mathbf{A}^* = (a_{rs}{}^*)$.

(vii) *Orthogonality* Matrix A is orthogonal if $\mathbf{AA}^{\mathbf{T}} = \mathbf{U}$.

Table 1.12 Laplace transforms

Definition	*f(t) from* $t=0+$	$F(s)=\mathcal{L}[f(t)]=\int_{0-}^{\infty} f(t)\cdot\exp(-st)\cdot dt$	
Sum	$af_1(t)+bf_2(t)$	$aF_1(s)+bF_2(s)$	
First derivative	$(d/dt)f(t)$	$sF(s)-f(0-)$	
nth derivative	$(d^n/dt^n)f(t)$	$s^nF(s)-s^{n-1}f(0-)-s^{n-2}f^{(1)}(0-)-\cdots-f^{(n-1)}(0-)$	
Definite integral	$\int_{0-}^{T} f(t)\cdot dt$	$\frac{1}{s}F(s)$	
Shift by T	$f(t-T)$	$\exp(-sT)\cdot F(s)$	
Periodic function (period T)	$f(t)$	$\frac{1}{1-\exp(-sT)}\int_{0-}^{T}\exp(-sT)\cdot f(t)\cdot dt$	
Initial value	$f(t),\ t\to 0+$	$sF(s),\ s\to\infty$	
Final value	$f(t),\ t\to\infty$	$sF(s),\ s\to 0$	
Description	*f(t)*	F(s)	*f(t) to base t*
1. Unit impulse	$\delta(t)$	1	
2. Unit step	$H(t)$	$\frac{1}{s}$	
3. Delayed step	$H(t-T)$	$\frac{\exp(-st)}{s}$	
4. Rectangular pulse (duration T)	$H(t)-H(t-T)$	$\frac{1-\exp(-sT)}{s}$	
5. Unit ramp	t	$\frac{1}{s^2}$	
6. Delayed ramp	$(t-T)H(t-T)$	$\frac{\exp(-sT)}{s^2}$	
7. nth-order ramp	t^n	$\frac{n!}{s^{n+1}}$	
8. Exponential decay	$\exp(-\alpha t)$	$\frac{1}{s+\alpha}$	
9. Exponential rise	$1-\exp(-\alpha t)$	$\frac{\alpha}{s(s+\alpha)}$	
10. Exponential × t	$t\exp(-\alpha t)$	$\frac{1}{(s+\alpha)^2}$	
11. Exponential × t^n	$t^n\exp(-\alpha t)$	$\frac{n!}{(s+\alpha)^{n+1}}$	
12. Difference of exponentials	$\exp(-\alpha t)-\exp(-\beta t)$	$\frac{\beta-\alpha}{(s+\alpha)(s+\beta)}$	

cont'd

Table 1.12 (*continued*)

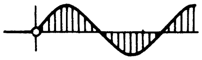
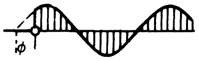
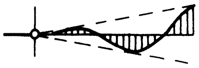
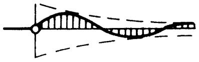
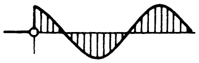
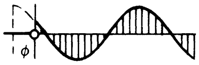
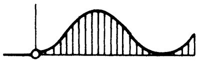
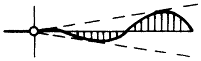
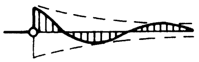
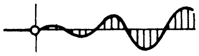
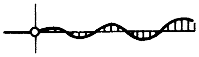
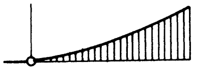
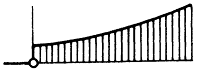
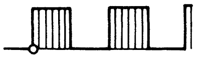
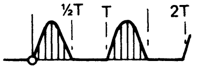
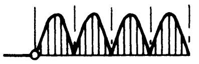
Definition	*f(t) from t = 0+*	$F(s) = L[f(t)] = \int_{0-}^{\infty} f(t) \cdot \exp(-st) \cdot dt$	
13. Sinusoidal	$\sin \omega t$	$\frac{\omega\varsigma}{s^2+\omega^2}$	
14. Phase-advanced sine	$\sin(\omega t+\phi)$	$\frac{\omega\cos\phi\varsigma + s\sin\phi\varsigma}{s^2+\omega^2}$	
15. Sine × *t*	$t \sin \omega t$	$\frac{2\omega s}{(s^2+\omega^2)^2}$	
16. Exponentially decaying sine	$\exp(-\alpha t)\sin \omega t$	$\frac{\omega\varsigma}{(s+\alpha)^2+\omega^2}$	
17. Cosinusoidal	$\cos \omega t$	$\frac{s}{s^2+\omega^2}$	
18. Phase-advanced cosine	$\cos(\omega t+\phi)$	$\frac{s\cos\phi\varsigma - \omega\sin\phi\varsigma}{s^2+\omega^2}$	
19. Offset cosine	$1-\cos \omega t$	$\frac{\omega^2}{s(s^2+\omega^2)}$	
20. Cosine × *t*	$t \cos \omega t$	$\frac{s^2-\omega^2}{(s^2+\omega^2)^2}$	
21. Exponentially decaying cosine	$\exp(-\alpha t)\cos \omega t$	$\frac{(s+\alpha)}{(s+\alpha)^2+\omega^2}$	
22. Trigonometrical function G(*t*)	$\sin \omega t - \omega t \cos \omega t$	$\frac{2\omega^3}{(s^2+\omega^2)^2}$	
23. Exponentially decaying trigonometrical function	$\exp(-\alpha t)\cdot G(t)$	$\frac{2\omega^3}{[(s+\alpha)^2+\omega^2]^2}$	
24. Hyperbolic sine	$\sinh \omega t$	$\frac{\omega\varsigma}{s^2-\omega^2}$	
25. Hyperbolic cosine	$\cosh \omega t$	$\frac{s}{s^2-\omega^2}$	
26. Rectangular wave (period *T*)	*f*(*t*)	$\frac{1+\tanh(sT/4)}{2s}$	
27. Half-wave rectified sine ($T = 2\pi/\omega$)	*f*(*t*)	$\frac{\omega\exp(sT/2)\mathrm{cosech}(sT/2)}{2(s^2+\omega^2)}$	
28. Full-wave rectified sine ($T = 2\pi/\omega$)	*f*(*t*)	$\frac{\omega\coth(sT/2)}{s^2+\omega^2}$	

Table 1.13 Decibel gain: power and voltage ratios

A	P_1/P_2	V_1/V_2	A	P_1/P_2	V_1/V_2
0	1.000	1.000	9	7.943	2.818
0.1	1.023	1.012	10	10.00	3.162
0.2	1.047	1.023	12	15.85	3.981
0.3	1.072	1.032	14	25.12	5.012
0.4	1.096	1.047	16	39.81	6.310
0.6	1.148	1.072	18	63.10	7.943
0.8	1.202	1.096	20	100.0	10.00
1.0	1.259	1.122	25	316.2	17.78
1.2	1.318	1.148	30	1000	31.62
1.5	1.413	1.189	35	3162	56.23
2.0	1.585	1.259	40	1.0×10^4	100.0
2.5	1.778	1.333	45	3.2×10^4	177.8
3.0	1.995	1.413	50	1.0×10^5	316.2
3.5	2.239	1.496	55	3.2×10^5	562.3
4.0	2.512	1.585	60	1.0×10^6	1 000
4.5	2.818	1.679	65	3.2×10^6	1 778
5.0	3.162	1.778	70	1.0×10^7	3 160
6.0	3.981	1.995	80	1.0×10^8	10 000
7.0	5.012	2.239	90	1.0×10^9	31 620
8.0	6.310	2.512	100	1.0×10^{10}	100 000

1.2.10.4 Determinant and trace

(i) The *determinant* of a square matrix **A** denoted by $|\mathbf{A}|$, also det(**A**), is defined by the recursive formula $|\mathbf{A}| = a_{11}\ \mathrm{M}_{11} - a_{12}\ \mathrm{M}_{12} + a_{13}\ \mathrm{M}_{13} - \cdots (\Leftarrow 1)^n a_{1n}\ \mathrm{M}_{1n}$ where M_{11} is the determinant of the matrix with row 1 and column 1 missing, M_{12} is the determinant of the matrix with row 1 and column 2 missing etc.

(ii) The *Trace* of **A** is denoted by $\mathrm{tr}(\mathbf{A}) = \Sigma_i\, a_{ii}$, $i = 1, 2 \ldots n$.

(iii) *Singularity* The square matrix **A** is singular if $\det(\mathbf{A}) = 0$.

(iv) The *Characteristic Polynomial* $\mathrm{P}(\lambda) = \det(\mathbf{A} - \lambda \mathbf{U})$.

1.2.10.5 Eigensystems

(i) *Eigenvalues* The eigenvalues of a matrix $\lambda(\mathbf{A})$ are the n complex roots $\lambda_1(\mathbf{A}), \lambda_2(\mathbf{A}) \ldots \lambda_n(\mathbf{A})$ of the characteristic polynomial $\det(\mathbf{A} - \lambda\mathbf{U}) = 0$. Normally in most engineering systems there are no equal roots so the eigenvalues are distinct.

(ii) *Eigenvectors* For any distinct eigenvalue $\lambda_i(\mathbf{A})$, there is an associated non-zero *right eigenvector* $\mathbf{X}_i$ satisfying the homogeneous equations $(\mathbf{A} - \lambda_i\mathbf{U})\ \mathbf{X}_i = \mathbf{0}$, $i = 1, 2 \ldots n$. The matrix $(\mathbf{A} - \lambda_i\mathbf{U})$ is singular, however, because the $\det(\mathbf{A} - \lambda_i\mathbf{U}) = 0$; hence $\mathbf{X}_i$ is not unique. In each set of equations $(\mathbf{A} - \lambda_i\mathbf{U})\ \mathbf{X}_i = \mathbf{0}$ one equation is redundant and only the relative values of the elements of $\mathbf{X}_i$ can be determined. Thus the eigenvectors can be scaled arbitrarily, one element being assigned a value and the other elements determined accordingly from the remaining non-homogeneous equations.

The equations can be written also as $\mathbf{AX}_i = \lambda_i\mathbf{X}_i$, or combining all eigenvalues and right eigenvectors, $\mathbf{AX} = \Lambda\mathbf{X}$, where Λ is a diagonal matrix of the eigenvalues and **X** is a square matrix containing all the right eigenvectors in corresponding order.

Since the eigenvalues of **A** and $\mathbf{A}^{\mathbf{T}}$ are identical, for every eigenvalue λ_i associated with an eigenvector $\mathbf{X}_i$ of **A** there is also an eigenvector $\mathbf{P}_i$ of $\mathbf{A}^{\mathbf{T}}$ such that $\mathbf{A}^{\mathbf{T}}\mathbf{P}_i = \lambda_i\mathbf{P}_i$. Alternatively the eigenvector $\mathbf{P}_i$ can be considered to be the *left eigenvector* of **A** by transposing the equation to give $\mathbf{P}_i^{\mathbf{T}}\mathbf{A} = \lambda_i\mathbf{P}_i^{\mathbf{T}}$, or combining into one matrix equation, $\mathbf{P}^{\mathbf{T}}\mathbf{A} = \mathbf{P}^{\mathbf{T}}\Lambda$.

Reciprocal eigenvectors Post-multiplying this last equation by the right eigenvector matrix **X** gives $\mathbf{P}^{\mathbf{T}}\mathbf{A}\,\mathbf{X} = \mathbf{P}^{\mathbf{T}}\Lambda\mathbf{X}$, which summarises the n sets of equations $\mathbf{P}_i^{\mathbf{T}}\mathbf{AX}_i = \mathbf{P}_i^{\mathbf{T}}\ \lambda_i\ \mathbf{X}_i = \lambda_i\mathbf{P}_i^{\mathbf{T}}\ \mathbf{X}_i = k_i\lambda_i$, where k_i is a scalar formed from the $(1 \times n)$ by $(n \times 1)$ vector product $\mathbf{P}_i^{\mathbf{T}}\ \mathbf{X}_i$. With both $\mathbf{P}_i$ and $\mathbf{X}_i$ being scaled arbitrarily, re-scaling the left eigenvectors such that $\mathbf{W}_i = (1/k_i)\ \mathbf{P}_i$, gives $\mathbf{W}_i^{\mathbf{T}}\ \mathbf{X}_j = \Leftarrow \delta ij = 1$, if $i = j$, and $= 0$ otherwise. In matrix form $\mathbf{W}^{\mathbf{T}}\,\mathbf{X} = \mathbf{U}$, the unit matrix. The re-scaled left eigenvectors $\mathbf{W}_i^{\mathbf{T}}$ are said to be the reciprocal eigenvectors corresponding to the right eigenvectors $\mathbf{X}_i$.

(iii) *Eigenvalue sensitivity analysis* The change in the numerical value of λ_i with a change in any matrix A element δa_{rs} is to a first approximation given by $\delta\lambda_i = (w_r)_i\ (x_s)_i\ \delta a_{rs}$ where $(w_r)_i$ is the r-th element of the reciprocal eignvector $\mathbf{W}_i$ corresponding to λ_i and $(\mathrm{x}_s)_i$ is the s-th element of the associated right eigenvector $\mathbf{X}_i$. In more compact form the sensitivity coefficients $\delta\lambda_i/\delta a_{rs}$ or condition numbers of all n eigenvalues with respect to all elements of matrix A are expressible by the 1-term dyads $\mathbf{S}_i = \mathbf{W}_i\,\mathbf{X}_i^{\mathbf{T}}$, $i = 1 \ldots n$.

$$
\mathbf{S}_i = \Leftarrow \begin{bmatrix}
\delta\lambda_i/\delta a_{11} & \delta\lambda_i/\delta a_{12} & \ldots & \delta\lambda_i/\delta a_{1n} \\
\delta\lambda_i/\delta a_{21} & \delta\lambda_i/\delta a_{22} & \ldots & \delta\lambda_i/\delta a_{2n} \\
\ldots & \ldots & \ldots & \ldots \\
\ldots & & \ldots & \\
\delta\lambda_i/\delta a_{n1} & \delta\lambda_i/\delta a_{n2} & \ldots & \delta\lambda_i/\delta a_{nn}
\end{bmatrix}
$$

The matrix $\mathbf{S}_i$ is known as i-th *eigenvalue sensitivity matrix*.

(iv) *Matrix functions* Transposed eigenvalue sensitivity matrices appear also in the dyadic expansion of a matrix and in matrix functions, thus $\mathbf{A} = \Sigma_i\lambda_i\mathbf{X}_i\ \mathbf{W}_i^{\mathbf{T}} = \Sigma_i\lambda_i\mathbf{S}_i^{\mathbf{T}}$, $i = 1 \ldots n$. Likewise $[\mathbf{A}]^2 > [\Sigma_i\lambda_i\ \ \mathbf{S}_i^{\mathbf{T}}]^2 = \Sigma_i\lambda_i^2\ \ \mathbf{S}_i^{\mathbf{T}}$ or in general $[\mathbf{A}]^{\mathrm{p}} = \Sigma_i\lambda_i^{\mathrm{p}}\ \mathbf{S}_i^{\mathbf{T}}$; thus, for example, $[\mathbf{A}]^{-1} = \Leftarrow \Sigma_i\ \lambda_i^{-1}\ \mathbf{S}_i^{\mathbf{T}}$.

1.2.10.6 Norms

(i) *Vector norms* A scalar measure of the magnitude of a vector $\boldsymbol{X}$ with elements $x_1, x_2 \ldots x_n$, is provided by a *norm*, the general family of norms being defined by $\|\boldsymbol{X}\| = [\Sigma_i\ |x_i|^p]^{1/p}$. The usual norms are found from the values of p.
If $p = 1$, $\|\boldsymbol{X}\|$ is the sum of the magnitudes of the elements,
$p = 2$, $\|\boldsymbol{X}\|$ is *Euclidean norm* or square root of the sum of the squares of the magnitudes of the elements,
$p =$ infinity, $\|\boldsymbol{X}\|$ is the *infinity norm* or magnitude of the largest element.

(ii) *Matrix norms* Several norms for matrices have also been defined, for matrix **A** two being the *Euclidean norm*, $\|\mathbf{A}\|_{\mathrm{E}} = [\Sigma_r\Sigma_s|\mathrm{a}_{rs}|^2]^{1/2}$, $r = 1, 2 \ldots m$; $s = 1, 2 \ldots n$, and the *absolute norm*, $\|\mathbf{A}\| = \max_{r,s}|\ a_{rs}|$.

1.3 Physical quantities

Engineering processes involve energy associated with physical materials to convert, transport or radiate energy. As energy has several natural forms, and as materials differ profoundly in their physical characteristics, separate technologies have been devised around specific processes; and materials may have to be considered macroscopically in bulk, or in microstructure (molecular, atomic and subatomic) in accordance with the applications or processes concerned.

1.3.1 Energy

Like 'force' and 'time', energy is a unifying concept invented to systematise physical phenomena. It almost defies precise definition, but may be described, as an aid to an intuitive appreciation.

Energy is the capacity for 'action' or *work*.

Work is the measure of the change in energy *state*.

State is the measure of the energy condition of a *system*.

System is the ordered arrangement of related physical entities or processes, represented by a *model*.

Mode is a description or mathematical formulation of the system to determine its *behaviour*.

Behaviour describes (verbally or mathematically) the energy processes involved in changes of state. Energy *storage* occurs if the work done on a system is recoverable in its original form. Energy *conversion* takes place when related changes of state concern energy in a different form, the process sometimes being reversible. Energy *dissipation* is an irreversible conversion into heat. Energy *transmission* and *radiation* are forms of energy transport in which there is a finite propagation time.

In a physical system there is an identifiable energy input W_i and output W_o. The system itself may store energy W_s and dissipate energy W. The energy conservation principle states that

$$W_i = W_s + W + W_o$$

Comparable statements can be made for energy changes Δw and for energy rates (i.e. powers), giving

$$\Delta w_i = \Delta w_s + \Delta w + \Delta w_o \text{ and } p_i = p_s + p + p_o$$

1.3.1.1 Analogues

In some cases the *mathematical* formulation of a system model resembles that of a model in a completely different physical system: the two systems are then analogues. Consider linear and rotary displacements in a simple mechanical system with the conditions in an electric circuit, with the following nomenclature:

f	force [N]	M	torque [N m]	v	voltage [V]
m	mass [kg]	J	inertia [kg m^2]	L	inductance [H]
r	friction [N s/m]	r	friction [N m s/rad]	R	resistance [Ω]
k	compliance [m/N]	k	compliance [rad/N m]	C	capacitance [F]
l	displacement [m]	θ	displacement [rad]	q	charge [C]
u	velocity [m/s]	ω	angular velocity [rad/s]	i	current [A]

The force necessary to maintain a uniform linear velocity u against a viscous frictional resistance r is $f = ur$; the power is $p = fu = u^2r$ and the energy expended over a distance l is $W = fut = u^2rt$, since $l = ut$. These are, respectively, the analogues of $v = iR$, $p = vi = i^2R$ and $W = vit = i^2Rt$ for the corresponding electrical system. For a constant angular velocity in a rotary mechanical system, $M = \omega r$, $p = M\omega = \omega^2 r$ and $W = \omega^2 rt$, since $\theta = \omega t$.

If a mass is given an acceleration $\mathrm{d}u/\mathrm{d}t$, the force required is $f = m(\mathrm{d}u/\mathrm{d}t)$ and the stored kinetic energy at velocity u_1 is $W = \frac{1}{2}mu_1^2$. For rotary acceleration, $M = J(\mathrm{d}\omega/\mathrm{d}t)$ and $W = \frac{1}{2}J\omega_1^2$. Analogously the application of a voltage v to a pure inductor L produces an increase of current at the rate $\mathrm{d}i/\mathrm{d}t$ such that $v = L(\mathrm{d}i/\mathrm{d}t)$ and the magnetic energy stored at current i_1 is $W = \frac{1}{2}Li^2$.

A mechanical element (such as a spring) of compliance k (which describes the displacement per unit force and is the inverse of the stiffness) has a displacement $l = kf$ when a force f is applied. At a final force f_1 the potential energy stored is $W = \frac{1}{2}kf_1^2$. For the rotary case, $\theta = kM$ and $W = \frac{1}{2}kM_1^2$. In the electric circuit with a pure capacitance C, to which a p.d. v is applied, the charge is $q = Cv$ and the electric energy stored at v_1 is $W = \frac{1}{2}Cv_1^2$.

Use is made of these correspondences in mechanical problems (e.g. of vibration) when the parameters can be considered to be 'lumped'. An ideal transformer, in which the primary m.m.f. in ampere-turns i_1N_1 is equal to the secondary m.m.f. i_2N_2 has as analogue the simple lever, in which a force f_1 at a point distant l_1 from the fulcrum corresponds to f_2 at l_2 such that $f_1l_1 = f_2l_2$.

A simple series circuit is described by the equation $v = L(\mathrm{d}i/\mathrm{d}t) + Ri + q/C$ or, with i written as $\mathrm{d}q/\mathrm{d}t$,

$$v = L(\mathrm{d}^2q/\mathrm{d}t^2) + R(\mathrm{d}q/\mathrm{d}t) + (1/C)q$$

A corresponding mechanical system of mass, compliance and viscous friction (proportional to velocity) in which for a displacement l the inertial force is $m(\mathrm{d}u/\mathrm{d}t)$, the compliance force is l/k and the friction force is ru, has a total force

$$f = m(\mathrm{d}^2l/\mathrm{d}t^2) + r(\mathrm{d}l/\mathrm{d}t) + (1/k)l$$

Thus the two systems are expressed in identical mathematical form.

1.3.1.2 Fields

Several physical problems are concerned with 'fields' having stream-line properties. The eddyless flow of a liquid, the current in a conducting medium, the flow of heat from a high- to a low-temperature region, are fields in which representative lines can be drawn to indicate at any point the direction of the flow there. Other lines, orthogonal to the flow lines, connect points in the field having equal potential. Along these equipotential lines there is no tendency for flow to take place.

Static electric fields between charged conductors (having equipotential surfaces) are of interest in problems of insulation stressing. Magnetic fields, which in air-gaps may be assumed to cross between high-permeability ferromagnetic surfaces that are substantially equipotentials, may be studied in the course of investigations into flux distribution in machines. All the fields mentioned above satisfy Laplacian equations of the form

$$(\partial^2V/\partial x^2) + (\partial^2V/\partial y^2) + (\partial^2V/\partial z^2) = 0$$

The solution for a physical field of given geometry will apply to other Laplacian fields of similar geometry, e.g.

System	*Potential*	*Flux*	*Medium*
current flow	voltage V	current I	conductivity σ
heat flow	temperature θ	heat q	thermal conductivity λ
electric field	voltage V	electric flux Q	permittivity ε
magnetic field	m.m.f. F	magnetic flux ϕ	permeability μ

The ratio I/V for the first system would give the effective conductance G; correspondingly for the other systems, $q/\theta\varsigma$ gives the thermal conductance, Q/V gives the capacitance and Φ/F gives the permeance, so that if measurements are made in one system the results are applicable to all the others.

It is usual to treat problems as two-dimensional where possible. Several field-mapping techniques have been devised, generally electrical because of the greater convenience and precision of electrical measurements. For two-dimensional problems, conductive methods include high-resistivity paper sheers, square-mesh 'nets' of resistors and electrolytic tanks. The tank is especially adaptable to three-dimensional cases of axial symmetry.

In the electrolytic *tank* a weak electrolyte, such as ordinary tap-water, provides the conducting medium. A scale model of the electrode system is set into the liquid. A low-voltage supply at some frequency between 50 Hz and 1 kHz is connected to the electrodes so that current flows through the electrolyte between them. A probe, adjustable in the horizontal plane and with its tip dipping vertically into the electrolyte, enables the potential field to be plotted. Electrode models are constructed from some suitable insulant (wood, paraffin wax, Bakelite, etc.), the electrode outlines being defined by a highly conductive material such as brass or copper. The metal is silver-plated to improve conductivity and reduce polarisation. Three-dimensional cases with axial symmetry are simulated by tilting the tank and using the surface of the electrolyte as a radial plane of the system.

The conducting-*sheet* analogue substitutes a sheet of resistive material (usually 'teledeltos' paper with silver-painted electrodes) for the electrolyte. The method is not readily adaptable to three-dimensional plots, but is quick and inexpensive in time and material.

The *mesh* or resistor-net analogue replaces a conductive continuum by a square mesh of equal resistors, the potential measurements being made at the nodes. Where the boundaries are simple, and where the 'grain size' is sufficiently small, good results are obtained. As there are no polarisation troubles, direct voltage supply can be used. If the resistors are made adjustable, the net can be adapted to cases of inhomogeneity, as when plotting a magnetic field in which permeability is dependent on flux density. Three-dimensional plots are made by arranging plane meshes in layers; the nodes are now the junctions of six instead of four resistors.

A stretched elastic membrane, depressed or elevated in appropriate regions, will accommodate itself smoothly to the differences in level: the height of the membrane everywhere can be shown to be in conformity with a two-dimensional Laplace equation. Using a rubber sheet as a membrane, the path of electrons in an electric field between electrodes in a vacuum can be investigated by the analogous paths of rolling bearing-balls. Many other useful analogues have been devised, some for the rapid solution of mathematical processes.

Recently considerable development has been made in point-by-point computer solutions for the more complicated field patterns in three-dimensional space.

1.3.2 Structure of matter

Material substances, whether solid, liquid or gaseous, are conceived as composed of very large numbers of *molecules*. A molecule is the smallest portion of any substance which cannot be further subdivided without losing its characteristic material properties. In all states of matter molecules are in a state of rapid continuous motion. In a *solid* the molecules are relatively closely 'packed' and the molecules, although rapidly moving, maintain a fixed mean position. Attractive forces between molecules account for the tendency of the solid to retain its shape. In a *liquid* the molecules are less closely packed and there is a weaker cohesion between them, so that they can wander about with some freedom within the liquid, which consequently takes up the shape of the vessel in which it is contained. The molecules in a *gas* are still more mobile, and are relatively far apart. The cohesive force is very small, and the gas is enabled freely to contract and expand. The usual effect of heat is to increase the intensity and speed of molecular activity so that 'collisions' between molecules occur more often; the average spaces between the molecules increase, so that the substance attempts to expand, producing internal pressure if the expansion is resisted.

Molecules are capable of further subdivision, but the resulting particles, called *atoms*, no longer have the same properties as the molecules from which they came. An atom is the smallest portion of matter than can enter into chemical combination or be chemically separated, but it cannot generally maintain a separate existence except in the few special cases where a single atom forms a molecule. A molecule may consist of one, two or more (sometimes many more) atoms of various kinds. A substance whose molecules are composed entirely of atoms of the same kind is called an *element*. Where atoms of two or more kinds are present, the molecule is that of a chemical *compound*. At present over 100 elements are recognised (*Table 1.14*: the atomic mass number A is relative to 1/12 of the mass of an element of carbon-12).

If the element symbols are arranged in a table in ascending order of atomic number, and in columns ('groups') and rows ('periods') with due regard to associated similarities, *Table 1.15* is obtained. Metallic elements are found on the left, non-metals on the right. Some of the correspondences that emerge are:

Group 1a: Alkali metals
(Li 3, Na 11, K 19, Rb 37, Cs 55, Fr 87)
2a: Alkaline earths
(Be 4, Mg 12, Ca 20, Sr 38, Ba 56, Ra 88)
1b: Copper group (Cu 29, Ag 47, Au 79)
6b: Chromium group (Cr 24, Mo 42, W 74)
7a: Halogens (F 9, Cl 17, Br 35, I 53, At 85)
0: Rare gases
(He 2, Ne 10, Ar 18, Kr 36, Xe 54, Rn 86)
3a–6a: Semiconductors
(B 5, Si 16, Ge 32, As 33, Sb 51, Te 52)

In some cases a horizontal relation obtains as in the transition series (Sc 21 . . . Ni 28) and the heavy-atom rare earth and actinide series. The explanation lies in the structure of the atom.

1.3.2.1 Atomic structure

The original Bohr model of the hydrogen atom was a central nucleus containing almost the whole mass of the atom, and a single *electron* orbiting around it. Electrons, as small particles of negative electric charge, were discovered at the end of the nineteenth century, bringing to light the complex structure of atoms. The hydrogen nucleus is a *proton*, a mass having a charge equal to that of an electron, but positive. Extended to all elements, each has a nucleus comprising mass particles, some (*protons*) with a positive charge, others (*neutrons*) with no charge. The atomic *mass number A* is the total number of protons and neutrons in the nucleus; the *atomic number Z* is the number of positive charges, and the normal number of orbital electrons. The nuclear structure is not known, and the forces that bind the protons against their mutual attraction are conjectural.

The hydrogen atom (*Figure 1.4*) has one proton ($Z=1$) and one electron in an orbit formerly called the K shell. Helium ($Z=2$) has two protons, the two electrons occupying the K shell which, by the Pauli exclusion principle, cannot have more than two. The next element in order is lithium ($Z=3$), the third electron in an outer L shell. With elements of increasing atomic number, the electrons are added to the L shell until it holds a maximum of 8, the surplus then occupying the M shell to a maximum of 18. The number of 'valence' electrons (those in the outermost shell) determines the physical and chemical properties of the element. Those with completed outer shells are 'stable'.

Isotopes An element is often found to be a mixture of atoms with the same chemical property but different atomic masses: these are isotopes. The isotopes of an element must have the same number of electrons and protons, but differ in the number of neutrons, accounting for the non-integral average mass numbers. For example, neon comprises 90.4% of mass number 20, with 0.6% of 21 and 9.0% of mass number 22, giving a resultant mass number of 20.18.

Energy states Atoms may be in various energy states. Thus, the filament of an incadescent lamp may emit light when excited by an electric current but not when the current is switched off. Heat energy is the kinetic energy of the atoms of a heated body. The more vigorous impact of atoms may not always shift the atom as a whole, but may shift an electron from one orbit to another of higher energy level within the atom. This position is not normally stable, and the electron gives up its momentarily acquired potential energy by falling back to its original level, releasing the energy as a light quantum or photon.

Ionisation Among the electrons of an atom, those of the outermost shell are unique in that, on account of all the electron charges on the shells between them and the nucleus, they are the most loosely bound and most easily *removable*. In a variety of ways it is possible so to excite an atom that one of the outer electrons is torn away, leaving the atom *ionised* or converted for the time into an *ion* with an effective positive charge due to the unbalanced electrical state it has acquired. Ionisation may occur due to impact by other fast-moving particles, by irradiation with rays of suitable wavelength and by the application of intense electric fields.

1.3.2.2 Wave mechanics

The fundamental laws of optics can be explained without regard to the nature of light as an electromagnetic wave phenomenon, and photoelectricity emphasises its nature as a stream or ray of corpuscles. The phenomena of diffraction or interference can only be explained on the wave concept. *Wave mechanics* correlates the two apparently conflicting ideas into a wider concept of 'waves of matter'. Electrons, atoms and even molecules participate in this duality, in that their effects appear sometimes as corpuscular, sometimes as of a wave nature. Streams of electrons behave in a corpuscular fashion in photoemission, but in certain circumstances show the diffraction effects familiar in wave action. Considerations of particle mechanics led de Broglie to write several theoretic papers (1922–1926) on the parallelism between the dynamics of a particle and geometrical optics, and suggested that it was necessary to admit that classical dynamics could not interpret phenomena involving energy quanta. Wave mechanics was established by Schrödinger in 1926 on de Broglie's conceptions.

When electrons interact with matter, they exhibit wave properties: in the free state they act like particles. Light has a similar duality, as already noted. The hypothesis of de Broglie is that a particle of mass m and velocity u has wave

Table 1.14 Elements (*Z*, atomic number; *A*, atomic mass; KLMNOPQ, electron shells)

Z	*Name and symbol*		*A*	*Shells*			
				K	L		
1	Hydrogen	H	1.008	1	—		
2	Helium	He	4.002	2	—		
3	Lithium	Li	6.94	2	1		
4	Beryllium	Be	9.02	2	2		
5	Boron	B	10.82	2	3		
6	Carbon	C	12	2	4		
7	Nitrogen	N	14.01	2	5		
8	Oxygen	O	16.00	2	6		
9	Fluorine	F	19.00	2	7		
10	Neon	Ne	20.18	2	8		
				KL	M	N	
11	Sodium	Na	22.99	10	1	—	
12	Magnesium	Mg	24.32	10	2	—	
13	Aluminium	Al	26.97	10	3	—	
14	Silicon	Si	28.06	10	4	—	
15	Phosphorus	P	31.02	10	5	—	
16	Sulphur	S	32.06	10	6	—	
17	Chlorine	Cl	35.46	10	7	—	
18	Argon	Ar	39.94	10	8	—	
19	Potassium	K	39.09	10	8	1	
20	Calcium	Ca	40.08	10	8	2	
21	Scandium	Sc	45.10	10	9	2	
22	Titanium	Ti	47.90	10	10	2	
23	Vanadium	V	0.95	10	11	2	
24	Chromium	Cr	52.01	10	13	1	
25	Manganese	Mn	54.93	10	13	2	
26	Iron	Fe	55.84	10	14	2	
27	Cobalt	Co	58.94	10	15	2	
28	Nickel	Ni	58.69	10	16	2	
29	Copper	Cu	63.57	10	18	1	
30	Zinc	Zn	65.38	10	18	2	
31	Gallium	Ga	69.72	10	18	3	
32	Germanium	Ge	72.60	10	18	4	
33	Arsenic	As	74.91	10	18	5	
34	Selenium	Se	78.96	10	18	6	
35	Bromine	Br	79.91	10	18	7	
36	Krypton	Kr	83.70	10	18	8	
				KLM	N	O	
37	Rubidium	Rb	85.44	28	8	1	
38	Strontium	Sr	87.63	28	8	2	
39	Yttrium	Y	88.92	28	9	2	
40	Zirconium	Zr	91.22	28	10	2	
41	Niobium	Nb	92.91	28	12	1	
42	Molybdenum	Mo	96.0	28	13	1	
43	Technetium	Tc	99.0	28	14	1	
44	Ruthenium	Ru	101.7	28	15	1	
45	Rhodium	Rh	102.9	28	16	1	
46	Palladium	Pd	106.7	28	18	—	
47	Silver	Ag	107.9	28	18	1	
48	Cadmium	Cd	112.4	28	18	2	
49	Indium	In	114.8	28	18	3	
50	Tin	Sn	118.7	28	18	4	
51	Antimony	Sb	121.8	28	18	5	
52	Tellurium	Te	127.6	28	18	6	
53	Iodine	I	126.9	28	18	7	
54	Xenon	Xe	131.3	28	18	8	
				KLM	N	O	P
55	Caesium	Cs	132.9	28	18	8	1
56	Barium	Ba	137.4	28	18	8	2

cont'd

Table 1.14 (*continued*)

Z	*Name and symbol*		*A*	*Shells*			
				KLM	N	O	P
57	Lanthanum	La	138.9	28	18	9	2
58	Cerium	Ce	140.1	28	19	9	2
59	Praseodymium	Pr	140.9	28	21	8	2
60	Neodymium	Nd	144.3	28	22	8	2
61	Promethium	Pm	147.0	28	23	8	2
62	Samarium	Sm	150.4	28	24	8	2
63	Europium	Eu	152.0	28	25	8	2
64	Gadolinium	Gd	157.3	28	25	9	2
65	Terbium	Tb	159.2	28	27	8	2
66	Dysprosium	Dy	162.5	28	28	8	2
67	Holmium	Ho	163.5	28	29	8	2
68	Erbium	Er	167.6	28	30	8	2
69	Thulium	Tm	169.4	28	31	8	2
70	Ytterbium	Yb	173.0	28	32	8	2
71	Lutecium	Lu	175.0	28	32	9	2
72	Hafnium	Hf	178.6	28	32	10	2
73	Tantalum	Ta	181.4	28	32	11	2
74	Tungsten	W	184.0	28	32	12	2
75	Rhenium	Re	186.3	28	32	13	2
76	Osmium	Os	191.5	28	32	14	2
77	Iridium	Ir	193.1	28	32	15	2
78	Platinum	Pt	195.2	28	32	17	1
79	Gold	Au	197.2	28	32	18	1
80	Mercury	Hg	200.6	28	32	18	2
81	Thallium	Tl	204.4	28	32	18	3
82	Lead	Pb	207.2	28	32	18	4
83	Bismuth	Bi	209.0	28	32	18	5
84	Polonium	Po	210.0	28	32	18	6
85	Astatine	At	211.0	28	32	18	7
86	Radon	Rn	222.0	28	32	18	8
				KLMN	O	P	Q
87	Francium	Fr	223.0	60	18	8	1
88	Radium	Ra	226.0	60	18	8	2
89	Actinium	Ac	227.0	60	18	9	2
90	Thorium	Th	232.0	60	18	10	2
91	Protoactinium	Pa	231.0	60	20	9	2
92	Uranium	U	238.0	60	21	9	2
93	Neptunium	Np	237.0	60	22	9	2
94	Plutonium	Pu	239.0	60	24	8	2
95	Americium	Am	243.0	60	25	8	2
96	Curium	Cm	247.0	60	25	9	2
97	Berkelium	Bk	247.0	60	26	9	2
98	Californium	Cf	251.0	60	28	8	2
99	Einsteinium	Es	254.0	60	29	8	2
100	Fermium	Fm	257.0	60	30	8	2
101	Mendelevium	Md	257.0	60	31	8	2
102	Nobelium	No	254.0	60	32	8	2
103	Lawrencium	Lr	256.0	60	32	9	2
104	Kurchatovium	Ku	—				
105	Hahnium	Ha	—				

properties with a wavelength $\lambda = h/mu$, where h is the Planck constant, $h = 6.626 \times 10^{-34}$ J s. The mass m is relativistically affected by the velocity.

When electron waves are associated with an atom, only certain fixed-energy states are possible. The electron can be raised from one state to another if it is provided, by some external stimulus such as a photon, with the necessary energy difference Δw in the form of an electromagnetic wave of wavelength $\lambda = hc/\Delta w$, where c is the velocity of free space radiation (3×10^8 m/s). Similarly, if an electron falls from a state of higher to one of lower energy, it emits energy Δw as radiation. When electrons are raised in energy level, the atom is *excited*, but not ionised.

1.3.2.3 Electrons in atoms

Consider the hydrogen atom. Its single electron is not located at a fixed point, but can be anywhere in a region near the nucleus with some probability. The particular region is a kind of shell or cloud, of radius depending on the electron's energy state.

With a nucleus of atomic number Z, the Z electrons can have several possible configurations. There is a certain radial pattern of electron probability cloud distribution (or shell pattern). Each electron state gives rise to a cloud pattern, characterised by a definite energy level, and described by the series of quantum numbers n, l, m_l and m_s. The number $n(=1, 2, 3, \ldots)$ is a measure of the energy level; $l(=0, 1, 2, \ldots)$ is concerned with angular momentum; m_l is a measure of the component of angular momentum in the direction of an applied magnetic field; and m_s arises from the electron spin. It is customary to condense the nomenclature so that electron states corresponding to $l = 0$, 1, 2 and 3 are described by the letters s, p, d and f and a numerical prefix gives the value of n. Thus boron has 2 electrons at level 1 with $l = 0$, two at level 2 with $l = 0$, and one at level 3 with $l = 1$: this information is conveyed by the description $(1s)^2(2s)^2(2p)^1$.

The energy of an atom as a whole can vary according to the electron arrangement. The most stable state is that of minimum energy, and states of higher energy content are *excited*. By Pauli's *exclusion principle* the maximum possible number of electrons in states 1, 2, 3, 4, ..., n are 2, 8, 18, 32, ..., $2n^2$, respectively. Thus, only 2 electrons can occupy the 1s state (or K shell) and the remainder must, even for the normal minimum-energy condition, occupy other states. Hydrogen and helium, the first two elements, have, respectively, 1 and 2 electrons in the 1-quantum (K) shell; the next, lithium, has its third electron in the 2-quantum (L) shell. The passage from lithium to neon results in the filling up of this shell to its full complement of 8 electrons. During the process, the electrons first enter the 2s subgroup, then fill the 2p subgroup until it has 6 electrons, the maximum allowable by the exclusion principle (see *Table 1.14*).

Very briefly, the effect of the electron-shell filling is as follows. Elements in the same chemical family have the same number of electrons in the subshell that is incompletely filled. The rare gases (He, Ne, Ar, Kr, Xe) have no uncompleted shells. Alkali metals (e.g. Na) have shells containing a single electron. The alkaline earths have two electrons in uncompleted shells. The good conductors (Ag, Cu, Au) have a single electron in the uppermost quantum state. An irregularity in the ordered sequence of filling (which holds consistently from H to Ar) begins at potassium (K) and continues to Ni, becoming again regular with Cu, and beginning a new irregularity with Rb.

The electron of a hydrogen atom, normally at level 1, can be raised to level 2 by endowing it with a particular quantity of energy most readily expressed as 10.2 eV. (1 eV = 1 electron-volt = 1.6×10^{-19} J is the energy acquired by a free electron falling through a potential difference of 1 V, which accelerates it and gives it kinetic energy.) The 10.2 V is the *first excitation potential* for the hydrogen atom. If the electron is given an energy of 13.6 eV, it is freed from the atom, and 13.6 V is the *ionisation potential*. Other atoms have different potentials in accordance with their atomic arrangement.

Table 1.15 Elements: periodic table

Periods	*Groups*																	
	1a	2a	3b	4b	5b	6b	7b	8b	8b	8b	1b	2b	3a	4a	5a	6a	7a	0
I	1 H				Metals									Non-metals				2 He
II	3 Li	4 Be											5 B	6 C	7 N	8 O	9 F	10 Ne
III	11 Na	12 Mg			Transitions								13 Al	14 Si	15 P	16 S	17 Cl	18 Ar
IV	19 K	20 Ca	21 Sc	22 Ti	23 V	24 Cr	25 Mn	26 Fe	27 Co	28 Ni	29 Cu	30 Zn	31 Ga	32 Ge	33 As	34 Se	35 Br	36 Kr
V	37 Rb	38 Sr	39 Y	40 Zr	41 Nb	42 Mo	43 Tc	44 Ru	45 Rh	46 Pd	47 Ag	48 Cd	49 In	50 Sn	51 Sb	52 Te	53 I	54 Xe
VI	55 Cs	56 Ba	57 La	72 Hf	73 Ta	74 W	75 Re	76 Os	77 Ir	78 Pt	79 Au	80 Hg	81 Tl	82 Pb	83 Bi	84 Po	85 At	86 Rn
VII	87 Fr	88 Ra	89 Ac															
		Rare earths			58 Ce	59 Pr	60 Nd	61 Pm	62 Sm	63 Eu	64 Gd	65 Tb	66 Dy	67 Ho	68 Er	69 Tm	70 Yb	71 Lu
		Actinides			90 Th	91 Pa	92 U	93 Np	94 Pu	95 Am	96 Cm	97 Bk	98 Cf	99 Es	100 Fm	101 Md	102 No	103 Lr

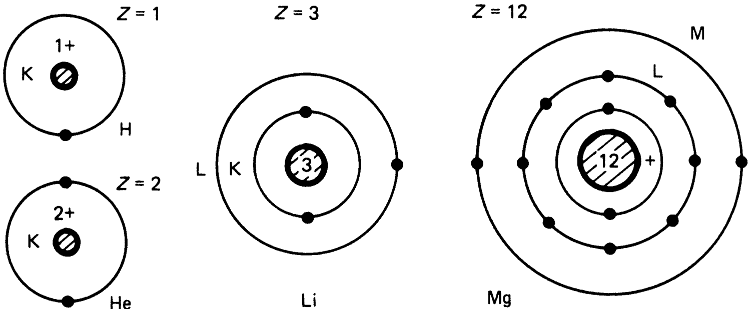


Figure 1.4 Atomic structure

1.3.2.4 Electrons in metals

An approximation to the behaviour of metals assumes that the atoms lose their valency electrons, which are free to wander in the ionic lattice of the material to form what is called an electron gas. The sharp energy levels of the free atom are broadened into wide bands by the proximity of others. The potential within the metal is assumed to be smoothed out, and there is a sharp rise of potential at the surface which prevents the electrons from escaping: there is a potential-energy step at the surface which the electrons cannot normally overcome: it is of the order of 10 eV. If this is called W, then the energy of an electron wandering within the metals is $-W+\frac{1}{2}mu^2$.

The electrons are regarded as undergoing continual collisions on account of the thermal vibrations of the lattice, and on Fermi-Dirac statistical theory it is justifiable to treat the energy states (which are in accordance with Pauli's principle) as forming an energy continuum. At very low temperatures the ordinary classical theory would suggest that electron energies spread over an almost zero range, but the exclusion principle makes this impossible and even at absolute zero of temperature the energies form a continuum, and physical properties will depend on how the electrons are distributed over the upper levels of this energy range.

Conductivity The interaction of free electrons with the thermal vibrations of the ionic lattice (called 'collisions' for brevity) causes them to 'rebound' with a velocity of random direction but small compared with their average velocities as particles of an electron gas. Just as a difference of electric potential causes a drift in the general motion, so a difference of temperature between two parts of a metal carries energy from the hot region to the cold, accounting for thermal conduction and for its association with electrical conductivity. The free electron theory, however, is inadequate to explain the dependence of conductivity on crystal axes in the metal.

At absolute zero of temperature (zero K = −273°C) the atoms cease to vibrate, and free electrons can pass through the lattice with little hindrance. At temperatures over the range 0.3–10 K (and usually round about 5 K) the resistance of certain metals, e.g. Zn, Al, Sn, Hg and Cu, becomes substantially zero. This phenomenon, known as *superconductivity*, has not been satisfactorily explained.

Superconductivity is destroyed by moderate magnetic fields. It can also be destroyed if the current is large enough to produce at the surface the same critical value of magnetic field. It follows that during the superconductivity phase the current must be almost purely superficial, with a depth of penetration of the order of 10 μm.

Table 1.16 Physical properties of metals

Approximate general properties at normal temperatures:

δς	density [kg/m^3]	k	thermal conductivity [W/(m K)]
E	elastic modulus [GPa]	T_m	melting point [K]
e	linear expansivity [µm/(m K)]	ρς	resistivity [nΩ m]
c	specific heat capacity [kJ/(kg K)]	ας	resistance–temperature coefficient [mΩ/(Ω K)]

Metal	δς	E	e	c	k	T_m	ρς	ας
Pure metals								
4 Beryllium	1840	300	120	1700	170	1560	33	9.0
11 Sodium	970	—	71	710	130	370	47	5.5
12 Magnesium	1740	44	26	1020	170	920	46	3.8
13 Aluminium	2700	70	24	900	220	930	27	4.2
19 Potassium	860	—	83	750	130	340	67	5.4
20 Calcium	1550	—	22	650	96	1120	43	4.2
24 Chromium	7100	25	8.5	450	43	2170	130	3.0
26 Iron	7860	220	12	450	75	1810	105	6.5
27 Cobalt	8800	210	13	420	70	1770	65	6.2
28 Nickel	8900	200	13	450	70	1730	78	6.5
29 Copper	8930	120	16	390	390	1360	17	4.3
30 Zinc	7100	93	26	390	110	690	62	4.1
42 Molybdenum	10 200	—	5	260	140	2890	56	4.3
47 Silver	10 500	79	19	230	420	1230	16	3.9
48 Cadmium	8640	60	32	230	92	590	75	4.0
50 Tin	7300	55	27	230	65	500	115	4.3
73 Tantalum	16 600	190	6.5	140	54	3270	155	3.1
74 Tungsten	19 300	360	4	130	170	3650	55	4.9
78 Platinum	21 500	165	9	130	70	2050	106	3.9
79 Gold	19 300	80	14	130	300	1340	23	3.6
80 Mercury	13 550	—	180	140	10	230	960	0.9
82 Lead	11 300	15	29	130	35	600	210	4.1
83 Bismuth	9800	32	13	120	9	540	1190	4.3
92 Uranium	18 700	13	—	120	—	1410	220	2.1
Alloys								
Brass (60 Cu, 40 Zn)	8500	100	21	380	120	1170	60	2.0
Bronze (90 Cu, 10 Sn)	8900	100	19	380	46	1280	—	—
Constantan	8900	110	15	410	22	1540	450	0.05
Invar (64 Fe, 36 Ni)	8100	145	2	500	16	1720	100	2.0
Iron, soft (0.2 C)	7600	220	12	460	60	1800	140	—
Iron cast (3.5 C, 2.5 Si)	7300	100	12	460	60	1450	—	—
Manganin	8500	130	16	410	22	1270	430	0.02
Steel (0.85 C)	7800	200	12	480	50	1630	180	—

Electron emission A metal may be regarded as a potential 'well' of depth $-V$ relative to its surface, so that an electron in the lowest energy state has (at absolute zero temperature) the energy $W = Ve$ (of the order 10 eV): other electrons occupy levels up to a height ε^* (5–8 eV) from the bottom of the 'well'. Before an electron can escape from the surface it must be endowed with an energy not less than $\phi = W - \varepsilon^*$, called the *work function*.

Emission occurs by *surface irradiation* (e.g. with light) of frequency v if the energy quantum hv of the radiation is at least equal to ϕ. The threshold of photoelectric emission is therefore with radiation at a frequency not less than $v = \phi/h$.

Emission takes place at *high temperatures* if, put simply, the kinetic energy of electrons normal to the surface is great enough to jump the potential step W. This leads to an expression for the emission current i in terms of temperature T, a constant A and the thermionic work function ϕ:

$$i = AT^2 \exp(-\phi/kT)$$

Electron emission is also the result of the application of a *high electric field intensity* (of the order 1–10 GV/m) to a metal surface; also when the surface is bombarded with electrons or ions of sufficient kinetic energy, giving the effect of *secondary* emission.

Crystals When atoms are brought together to form a crystal, their individual sharp and well-defined energy levels merge into energy *bands*. These bands may overlap, or there may be gaps in the energy levels available, depending on the lattice spacing and interatomic bonding. Conduction can take place only by electron migration into an empty or partly filled band; filled bands are not available. If an electron acquires a small amount of energy from the externally applied electric field, and can move into an available empty level, it can then contribute to the conduction process.

1.3.2.5 Insulators

In this case the 'distance' (or energy increase Δw in electron-volts) is too large for moderate electric applied fields to endow electrons with sufficient energy, so the material remains an insulator. High temperatures, however, may

Table 1.17 Physical properties of non-metals

Approximate general properties:

δς	density [kg/m^3]	T_m	melting point [K]
e	linear expansivity [μm/(m K)]	ρς	resistivity [MΩ m]
c	specific heat capacity [kJ/(kg K)]	ϵ_r	relative permittivity [−]
k	thermal conductivity [W/(m K)]		

Material	δς	e	c	k	T_m	ρς	ες
Asbestos (packed)	580	—	0.84	0.19	—	—	3
Bakelite	1300	30	0.92	0.20	—	0.1	7
Concrete (dry)	2000	10	0.92	1.70	—	—	—
Diamond	3510	1.3	0.49	165	4000	10^7	—
Glass	2500	8	0.84	0.93	—	10^6	8
Graphite	2250	2	0.69	160	3800	10^{-11}	—
Marble	2700	12	0.88	3	—	10^3	8.5
Mica	2800	3	0.88	0.5	—	10^8	7
Nylon	1140	100	1.7	0.3	—	—	—
Paper	900	—	—	0.18	—	10^4	2
Paraffin wax	890	110	2.9	0.26	—	10^9	2
Perspex	1200	80	1.5	1.9	—	10^{14}	3
Polythene	930	180	2.2	0.3	—	—	2.3
Porcelain	2400	3.5	0.8	1.0	1900	10^6	6
Quartz (fused)	2200	0.4	0.75	0.22	2000	10^{14}	3.8
Rubber	1250	—	1.5	0.15	—	10^7	3
Silicon	2300	7	0.75	—	1690	0.1	2.7

Table 1.18 Physical properties of liquids

Average values at 20°C (293 K):

δς	density [kg/m^3]	k	thermal conductivity [W/(m K)]
v	viscosity [mPa s]	T_m	melting point [K]
e	cubic expansivity [10^{-3}/K]	T_b	boiling point [K]
c	specific heat capacity [kJ/(kg K)]	ες	relative permittivity [−]

Liquid		δς	v	e	c	k	T_m	T_b	ες
Acetone	$(CH_3)_2CO$	792	0.3	1.43	2.2	0.18	178	329	22
Benzine	C_6H_6	881	0.7	1.15	1.7	0.14	279	353	2.3
Carbon disulphide	CS_2	1260	0.4	1.22	1.0	0.14	161	319	2.6
Carbon tetrachloride	CCl_4	1600	1.0	1.22	0.8	0.10	250	350	2.2
Ether	$(C_2H_5)_2O$	716	0.2	1.62	2.3	0.14	157	308	4.3
Glycerol	$C_3H_5(OH)_3$	1270	1500	0.50	2.4	0.28	291	563	56
Methanol	CH_3OH	793	0.6	1.20	1.2	0.21	175	338	32
Oil	—	850	85	0.75	1.6	0.17	—	—	3.0
Sulphuric acid	H_2SO_4	1850	28	0.56	1.4	—	284	599	—
Turpentine	$C_{10}H_{16}$	840	1.5	0.10	1.8	0.15	263	453	2.3
Water	H_2O	1000	1.0	0.18	4.2	0.60	273	373	81

result in sufficient thermal agitation to permit electrons to 'jump the gap'.

1.3.2.6 Semiconductors

Intrinsic semiconductors (i.e. materials between the good conductors and the good insulators) have a small spacing of about 1 eV between their permitted bands, which affords a low conductivity, strongly dependent on temperature and of the order of one-millionth that of a conductor.

Impurity semiconductors have their low conductivity raised by the presence of minute quantities of foreign atoms (e.g. 1 in 10^8) or by deformations in the crystal structure. The impurities 'donate' electrons of energy level that can be raised into a conduction band (n-type); or they can attract an electron from a filled band to leave a 'hole', or electron deficiency, the movement of which corresponds to the movement of a positive charge (p-type).

1.3.2.7 Magnetism

Modern magnetic theory is very complex, with ramifications in several branches of physics. Magnetic phenomena are associated with moving charges. Electrons, considered as particles, are assumed to possess an axial spin, which gives them the effect of a minute current turn or of a small permanent magnet, called a Bohr *magneton*. The gyroscopic effect of electron spin develops a precession when a magnetic field is applied. If the precession effect exceeds the spin effect, the external applied magnetic field produces less

Table 1.19 Physical properties of gases

Values at 0°C (273 K) and atmospheric pressure:

$\delta\varsigma$	density [kg/m^3]	k	thermal conductivity [m W/(m K)]
v	viscosity [μPa s]	T_m	melting point [K]
c_p	specific heat capacity [kJ/(kg K)]	T_b	boiling point [K]
c_p/c_v	ratio between specific heat capacity at constant pressure and at constant volume		

Gas		$\delta\varsigma$	v	c_p	c_p/c_v	k	T_m	T_b
Air	—	1.293	17.0	1.00	1.40	24	—	—
Ammonia	NH_3	0.771	9.3	2.06	1.32	22	195	240
Carbon dioxide	CO_2	1.977	13.9	0.82	1.31	14	216*	194
Carbon monoxide	CO	1.250	16.4	1.05	1.40	23	68	81
Chlorine	Cl_2	3.214	12.3	0.49	1.36	7.6	171	239
Deuterium	D	0.180	—	—	1.73	—	18	23
Ethane	C_2H_6	1.356	8.6	1.72	1.22	18	89	184
Fluorine	F_2	1.695	—	0.75	—	—	50	85
Helium	He	0.178	18.6	5.1	1.66	144	1.0	4.3
Hydrogen	H_2	0.090	8.5	14.3	1.41	174	14	20
Hydrogen chloride	HCl	1.639	13.8	0.81	1.41	—	161	189
Krypton	Kr	3.740	23.3	—	1.68	8.7	116	121
Methane	CH_4	0.717	10.2	2.21	1.31	30	90	112
Neon	Ne	0.900	29.8	1.03	1.64	46	24	27
Nitrogen	N_2	1.251	16.7	1.04	1.40	24	63	77
Oxygen	O_2	1.429	19.4	0.92	1.40	25	55	90
Ozone	O_3	2.220	—	—	1.29	—	80	161
Propane	C_3H_8	2.020	7.5	1.53	1.13	15	83	231
Sulphur dioxide	SO_2	2.926	11.7	0.64	1.27	8.4	200	263
Xenon	Xe	5.890	22.6	—	1.66	5.2	161	165

*At pressure of 5 atm.

magnetisation than it would in free space, and the material of which the electron is a constituent part is *diamagnetic*. If the spin effect exceeds that due to precession, the material is *paramagnetic*. The spin effect may, in certain cases, be very large, and high magnetisations are produced by an external field: such materials are *ferromagnetic*.

An iron atom has, in the $n = 4$ shell (N), electrons that give it conductive properties. The K, L and N shells have equal numbers of electrons possessing opposite spin directions, so cancelling. But shell M contains 9 electrons spinning in one direction and 5 in the other, leaving 4 net magnetons. Cobalt has 3, and nickel 2. In a solid metal further cancellation occurs and the average number of unbalanced magnetons is: Fe, 2.2; Co, 1.7; Ni, 0.6.

In an iron crystal the magnetic axes of the atoms are aligned, unless upset by excessive thermal agitation. (At 770°C for Fe, the Curie point, the directions become random and ferromagnetism is lost.) A single Fe crystal magnetises most easily along a cube edge of the structure. It does not exhibit spontaneous magnetisation like a permanent magnet, however, because a crystal is divided into a large number of *domains* in which the various magnetic directions of the atoms form closed paths. But if a crystal is exposed to an external applied magnetic field, (a) the electron spin axes remain initially unchanged, but those domains having axes in the favourable direction grow at the expense of the others (domain wall displacement); and (b) for higher field intensities the spin axes orientate into the direction of the applied field.

If wall movement makes a domain acquire more internal energy, then the movement will relax again when the external field is removed. But if wall movement results in loss of energy, the movement is non-reversible—i.e. it needs external force to reverse it. This accounts for hysteresis and remanence phenomena.

The closed-circuit self-magnetisation of a domain gives it a mechanical strain. When the magnetisation directions of individual domains are changed by an external field, the strain directions alter too, so that an assembly of domains will tend to lengthen or shorten. Thus, readjustments in the crystal lattice occur, with deformations (e.g. 20 parts in 10^6) in one direction. This is the phenomenon of *magnetostriction*.

The practical art of magnetics consists in control of magnetic properties by alloying, heat treatment and mechanical working to produce variants of crystal structure and consequent magnetic characteristics.

Table 1.20 Characteristic temperatures

Temperature T [kelvin] corresponds to $\theta_c = T - 273.15$ [degree Celsius] and to $\theta_f = \theta_c\,(9/5) - 32$ [degree Fahrenheit].

Condition	T	θ_c	θ_f
Absolute zero	**0**	−273.15	−459.7
Boiling point of oxygen	90.18	−182.97	−297.3
Zero of Fahrenheit scale	255.4	−17.78	**0**
Melting point of ice	273.15	**0**	32.0
Triple point of water	273.16	0.01	32.02
Maximum density of water	277.13	3.98	39.16
‘Normal’ ambient	293.15	20	68
Boiling point of water	373.15	100	212
Boiling point of sulphur	717.8	444.6	832
Freezing point of silver	1234	962	1762
Freezing point of gold	1336	1064	1945

Table 1.21 General physical constants (approximate values, to five significant figures)

Quantity	*Symbol*	*Numerical value*	*Unit*
Acceleration of free fall (standard)	g_n	9.8066	m/s^2
Atmospheric pressure (standard)	p_0	1.0132×10^{5}	Pa
Atomic mass unit	u	1.6606×10^{-27}	kg
Avogadro constant	N_A	6.0220×10^{23}	mol^{-1}
Bohr magneton	μ_B	9.2741×10^{-24}	J/T, A m^2
Boltzmann constant	k	1.3807×10^{-23}	J/K
Electron			
charge	$-e$	1.6022×10^{-19}	C
mass	m_e	9.1095×10^{-31}	kg
charge/mass ratio	e/m_e	1.7588×10^{11}	C/kg
Faraday constant	F	9.6485×10^{4}	C/mol
Free space			
electric constant	ε_0	8.8542×10^{-12}	F/m
intrinsic impedance	Z_0	376.7	Ω
magnetic constant	μ_0	$4\pi \times 10^{-7}$	H/m
speed of electromagnetic waves	c	2.9979×10^{8}	m/s
Gravitational constant	G	6.6732×10^{-11}	N m^2/kg^2
Ideal molar gas constant	R	8.3144	J/(mol K)
Molar volume at s.t.p.	V_m	2.2414×10^{-2}	m^3/mol
Neutron rest mass	m_n	1.6748×10^{-27}	kg
Planck constant	h	6.6262×10^{-34}	J s
normalised	$h/2\pi$	1.0546×10^{-34}	J s
Proton			
charge	$+e$	1.6022×10^{-19}	C
rest mass	m_p	1.6726×10^{-27}	kg
charge/mass ratio	e/m_p	0.9579×10^{8}	C/kg
Radiation constants	c_1	3.7418×10^{-16}	W m^2
	c_2	1.4388×10^{-2}	m K
Rydberg constant	R_H	1.0968×10^{7}	m^{-1}
Stefan-Boltzmann constant	σ	5.6703×10^{-8}	J/(m^2 K^4)
Wien constant	k_w	2.8978×10^{-3}	m K

1.4 Physical properties

The nature, characteristics and properties of materials arise from their atomic and molecular structure. Tables of approximate values for the physical properties of metals, non-metals, liquids and gases are appended, together with some characteristic temperatures and the numerical values of general physical constants.

1.5 Electricity

In the following paragraphs electrical phenomena are described in terms of the effects of electric charge, at a level adequate for the purpose of simple explanation.

In general, charges may be at rest, or in motion, or in acceleration. *At rest*, charges have around them an electric (or *electrostatic*) field of force. *In motion* they constitute a current, which is associated with a magnetic (or *electrodynamic*) field of force additional to the electric field. In *acceleration*, a third field component is developed which results in energy propagation by *electromagnetic waves*.

1.5.1 Charges at rest

Figure 1.5 shows two bodies in air, charged by applying between them a potential difference, or (having been in close contact) by forcibly separating them. Work must have been done in a physical sense to produce on one an excess and on the other a deficiency of electrons, so that the system is a repository of potential energy. (The work done in separating charges is measured by the product of the charges separated and the difference of electrical potential that results.) Observation of the system shows certain effects of interest: (1) there is a difference of electric potential between the bodies depending on the amount of charge and the geometry of the system; (2) there is a mechanical force of attraction between the bodies. These effects are deemed to be manifestations of the *electric field* between the bodies, described as a special state of space and depicted by *lines of force* which express in a pictorial way the strength and direction of the force effects. The lines stretch between positive and negative elements of charge through the medium (in this case, air) which separates the two charged bodies. The electric field is only a concept—for the lines have no real existence—used to calculate various effects produced when charges are separated by any method which results in excess and deficiency states of atoms by electron transfer. Electrons and protons, or electrons and positively ionised atoms, attract each other, and the stability of the atom may be considered due to the balance of these attractions and dynamic forces such as electron spin. Electrons are repelled by electrons and protons by protons, these forces being summarised in the rules, formulated experimentally long before our present knowledge of atomic structure, that 'like charges repel and unlike charges attract one another'.

1.5.2 Charges in motion

In substances called *conductors*, the outer shell electrons can be more or less freely interchanged between atoms.

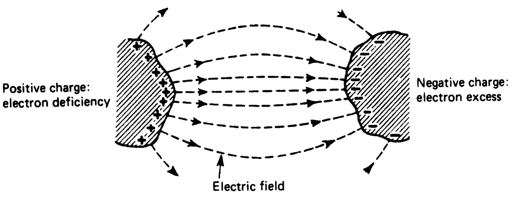

Positive charge: electron deficiency
Negative charge: electron excess
Electric field

Figure 1.5 Charged conductors and their electric field

In copper, for example, the molecules are held together comparatively rigidly in the form of a 'lattice'—which gives the piece of copper its permanent shape—through the interstices of which outer electrons from the atoms can be interchanged within the confines of the surface of the piece, producing a random movement of free electrons called an 'electron atmosphere'. Such electrons are responsible for the phenomenon of electrical conductivity.

In other substances called *insulators*, all the electrons are more or less firmly bound to their parent atoms, so that little or no relative interchange of electron charges is possible. There is no marked line of demarcation between conductors and insulators, but the copper group metals, in the order silver, copper, gold, are outstanding in the series of conductors.

1.5.2.1 Conduction

Conduction is the name given to the movement of electrons, or ions, or both, giving rise to the phenomena described by the term *electric current*. The effects of a current include a redistribution of charges, heating of conductors, chemical changes in liquid solutions, magnetic effects, and many subsidiary phenomena.

If at a specified point on a conductor (*Figure 1.6*) n_1 carriers of electric charge (they can be water-drops, ions, dust particles, etc.) each with a positive charge e_1 arrive per second, and n_2 carriers (such as electrons) each with a negative charge e_2 arrive in the opposite direction per second, the total rate of passing of charge is $n_1e_1 + n_2e_2$, which is the charge per second or *current*. A study of conduction concerns the kind of carriers and their behaviour under given conditions. Since an electric field exerts mechanical forces on charges, the application of an electric field (i.e. a potential difference) between two points on a conductor will cause the movement of charges to occur, i.e. a current to flow, so long as the electric field is maintained.

The discontinuous particle nature of current flow is an observable factor. The current carried by a number of electricity carriers will vary slightly from instant to instant with the number of carriers passing a given point in a conductor. Since the electron charge is 1.6×10^{-19} C, and the passage of one coulomb per second (a rate of flow of *one ampere*) corresponds to $10^{19}/1.6 = 6.3 \times 10^{18}$ electron charges per second, it follows that the discontinuity will be observed only when the flow comprises the very rapid movement of a few electrons. This may happen in gaseous conductors, but in metallic conductors the flow is the very slow drift (measurable in mm/s) of an immense number of electrons.

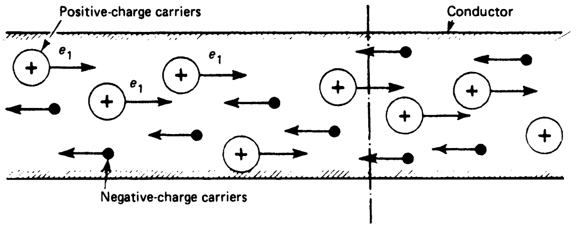


Figure 1.6 Conduction by charge carriers

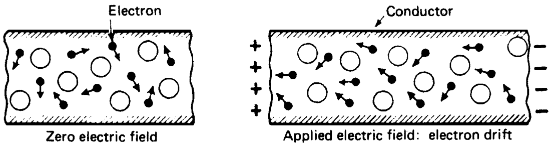


Figure 1.7 Electronic conduction in metals

A current may be the result of a two-way movement of positive and negative particles. Conventionally the direction of current flow is taken as the same as that of the positive charges and against that of the negative ones.

1.5.2.2 Metals

Reference has been made above to the 'electron atmosphere' of electrons in random motion within a lattice of comparatively rigid molecular structure in the case of copper, which is typical of the class of good metallic conductors. The random electronic motion, which intensifies with rise in temperature, merges into an average shift of charge of almost (but not quite) zero continuously (*Figure 1.7*). When an electric field is applied along the length of a conductor (as by maintaining a potential difference across its ends), the electrons have a *drift* towards the positive end superimposed upon their random digressions. The drift is slow, but such great numbers of electrons may be involved that very large currents, entirely due to electron drift, can be produced by this means. In their passage the electrons are impeded by the molecular lattice, the collisions producing heat and the opposition called *resistance*. The conventional direction of current flow is actually opposite to that of the drift of charge, which is exclusively electronic.

1.5.2.3 Liquids

Liquids are classified according to whether they are *non-electrolytes* (non-conducting) or *electrolytes* (conducting). In the former the substances in solution break up into electrically balanced groups, whereas in the latter the substances form ions, each a part of a single molecule with either a positive or a negative charge. Thus, common salt, NaCl, in a weak aqueous solution breaks up into sodium and chlorine ions. The sodium ion Na^{+} is a sodium atom less one electron; the chlorine ion Cl^{-} is a chlorine atom with one electron more than normal. The ions attach themselves to groups of water molecules. When an electric field is applied, the sets of ions move in opposite directions, and since they are much more massive than electrons, the conductivity produced is markedly inferior to that in metals. Chemical actions take place in the liquid and at the electrodes when current passes. Faraday's Electrolysis Law states that the mass of an ion deposited at an electrode by electrolyte action is proportional to the quantity of electricity which passes and to the *chemical equivalent* of the ion.

1.5.2.4 Gases

Gaseous conduction is strongly affected by the pressure of the gas. At pressures corresponding to a few centimetres of mercury gauge, conduction takes place by the movement of positive and negative ions. Some degree of ionisation is

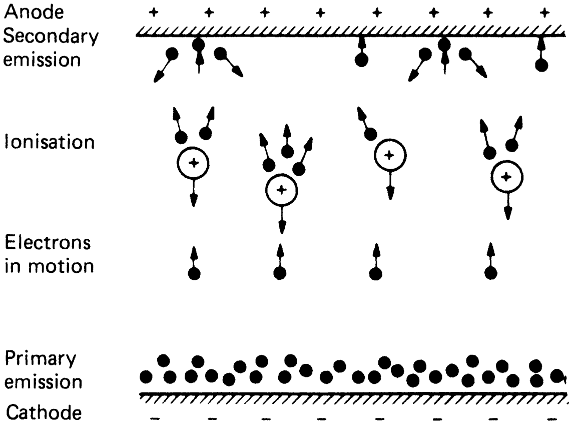


Figure 1.8 Conduction in low-pressure gas

always present due to stray radiations (light, etc.). The electrons produced attach themselves to gas atoms and the sets of positive and negative ions drift in opposite directions. At very low gas pressures the electrons produced by ionisation have a much longer free path before they collide with a molecule, and so have scope to attain high velocities. Their motional energy may be enough to *shockionise* neutral atoms, resulting in a great enrichment of the electron stream and an increased current flow. The current may build up to high values if the effect becomes cumulative, and eventually conduction may be effected through a *spark* or *arc*.

In a *vacuum* conduction can be considered as purely electronic, in that any electrons present (there can be no *molecular* matter present in a perfect vacuum) are moved in accordance with the force exerted on them by an applied electric field. The number of electrons is small, and although high speeds may be reached, the conduction is generally measurable only in milli- or microamperes.

Some of the effects are illustrated in *Figure 1.8*, representing part of a vessel containing a gas or vapour at low pressure. At the bottom is an electrode, the *cathode*, from the surface of which electrons are emitted, generally by heating the cathode material. At the top is a second electrode, the *anode*, and an electric field is established between the electrodes. The field causes electrons emitted from the cathode to move upward. In their passage to the anode these electrons will encounter gas molecules. If conditions are suitable, the gas atoms are ionised, becoming in effect positive charges associated with the nuclear mass. Thereafter the current is increased by the detached electrons moving upwards and by the positive ions moving more slowly downwards. In certain devices (such as the mercury arc rectifier) the impact of ions on the cathode surface maintains its emission. The impact of electrons on the anode may be energetic enough to cause the *secondary emission* of electrons from the anode surface. If the gas molecules are excluded and a vacuum is established, the conduction becomes purely electronic.

1.5.2.5 Insulators

If an electric field is applied to a perfect insulator, whether solid, liquid or gaseous, the electric field affects the atoms by producing a kind of 'stretching' or 'rotation' which displaces the electrical centres of negative and positive in opposite directions. This polarisation of the dielectric insulating material may be considered as taking place in the manner indicated in *Figure 1.9*. Before the electric field is applied, the atoms of the insulator are neutral and unstrained; as the potential difference is raised the electric field exerts opposite mechanical forces on the negative and positive charges and the atoms become more and more highly strained (*Figure 1.9*(*a*)). On the left face the atoms will all present their negative charges at the surface: on the right face, their positive charges. These surface polarisations are such as to account for the effect known as *permittivity*. The small displacement of the atomic electric charges constitutes a *polarisation current*. *Figure 1.9*(*b*) shows that, for excessive electric field strength, conduction can take place, resulting in insulation breakdown.

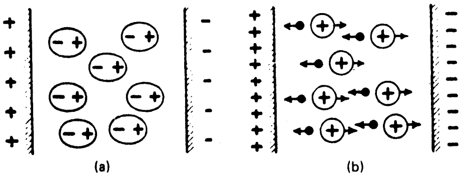


Figure 1.9 Polarisation and breakdown in insulator

The electrical properties of metallic conductors and of insulating materials are listed in *Tables 1.22* and *1.23*.

1.5.2.6 Convection current

Charges can be moved mechanically, on belts, water-drops, dust and mist particles, and by beams of high-speed electrons (as in a cathode ray oscilloscope). Such movement, independent of an electric field, is termed a *convection current*.

1.5.3 Charges in acceleration

Reference has been made to the emission of energy (photons) when an electron falls from an energy level to a lower one. Radiation has both a particle and a wave nature, the latter associated with energy propagation through empty space and through transparent media.

1.5.3.1 Maxwell equations

Faraday postulated the concept of the field to account for 'action at a distance' between charges and between magnets. Maxwell (1873) systematised this concept in the form of electromagnetic field equations. These refer to media *in bulk*. They naturally have no direct relation to the electronic nature of conduction, but deal with the fluxes of electric, magnetic and conduction fields, their flux densities, and the bulk material properties (permittivity ε, permeability μςand conductivity σ) of the media in which the fields exist. To the work of Faraday. Ampère and Gauss, Maxwell added the concept of displacement current.

Displacement current Around an electric field that changes with time there is evidence of a magnetic field. By analogy with the magnetic field around a conduction current, the rate of change of an electric field may be represented by the presence of a *displacement current*. The concept is applicable to an electric circuit containing a capacitor: there is a conduction current i_c in the external circuit but not between the electrodes of the capacitor. The capacitor, however, must be acquiring or losing charge and its electric field must be changing. If the rate of change is represented by a displacement current $i_d = i_c$, not only is the magnetic field accounted for, but also there now exists a 'continuity' of current around the circuit.

Displacement current is present in any material medium, conducting or insulating, whenever there is present an

Table 1.22 Electrical properties of conductors

Typical approximate values at 293 K (20 °C):

g conductivity relative to I.S.A.C. [%]
ρ resistivity [nΩ m]
α resistance–temperature coefficient [mΩ/(Ω K)]

Material		g	ρ	α
International standard				
annealed copper (ISAC)		100	17.2	3.93
Copper				
annealed		99	17.3	3.90
hard-drawn		97	17.7	3.85
Brass (60/40)				
cast		23	75	1.6
rolled		19	90	1.6
Bronze		48	36	1.65
Phosphor-bronze		29–14	6–12	1.0
Cadmium-copper, hard-drawn		82–93	21–18	4.0
Copper-clad steel, hard-drawn		30–40	57–43	3.75
Aluminium				
cast		66	26	3.90
hard-drawn		62	28	3.90
duralumin		36	47	—
Iron				
wrought		16	107	5.5
cast				
grey		2.5	700	—
white		1.7	1000	2.0
malleable		5.9	300	—
nomag		1.1	1600	4.5
Steel				
0.1% C		8.6	200	4.2
0.4% C		11	160	4.2
core				
1% Si		10	170	—
2% Si		4.9	350	—
4% Si		3.1	550	—
wire				
galvanised		12	140	4.4
45 ton		10	170	3.4
80 ton		8	215	3.4
Resistance alloys*				
80 Ni, 20 Cr	(1)	1.65	1090	0.1
59 Ni, 16 Cr, 25 Fe	(2)	1.62	1100	0.2
37 Ni, 18 Cr, 2 Si, 43 Fe	(3)	1.89	1080	0.26
45 Ni, 54 Cu	(4)	3.6	490	0.04
20 Ni, 80 Cu	(5)	6.6	260	0.29
15 Ni, 62 Cu, 22 Zn	(6)	5.0	340	0.25
4 Ni, 84 Cu, 12 Mn	(7)	3.6	480	0.0
Gold		73	23.6	3.0
Lead		7.8	220	4.0
Mercury		1.8	955	0.7
Molybdenum		30	57	4.0
Nickel		12.6	136	5.0
Platinum		14.7	117	3.9
Silver				
annealed		109	15.8	4.0
hard-drawn		98.5	17.5	4.0
Tantalum		11.1	155	3.1
Tungsten		31	56	4.5
Zinc		28	62	4.0

*Resistance alloys: (1) furnaces, radiant elements; (2) electric irons, tubular heaters; (3) furnace elements; (4) control resistors; (5) cupro; (6) German silver, platinoid; (7) Manganin.

electric field that changes with time. There is a displacement current along a copper conductor carrying an alternating current, but the conduction current is vastly greater even at very high frequencies. In poor conductors and in insulating materials the displacement current is comparable to (or greater than) the conduction current if the frequency is high enough. In free space and in a perfect insulator only displacement current is concerned.

Equations The following symbols are used, the SI unit of each appended. The permeability and permittivity are absolute values ($\mu = \mu_r\mu_0$, $\varepsilon = \varepsilon_r\varepsilon_0$). Potentials and fluxes are scalar quantities: field strength and flux density, also surface and path-length elements, are vectorial.

Field	*Electric*	*Magnetic*	*Conduction*
Potential	V [V]	F [A]	V [V]
Field strength	E [V/m]	H [A/m]	E [V/m]
Flux	Q [C]	ϕ [Wb]	I [A]
Flux density	D [C/m^2]	B [T]	J [A/m^2]
Material property	ε [F/m]	μ [H/m]	σ [S/m]

The total electric flux emerging from a charge $+Q$ or entering a charge $-Q$ is equal to Q. The integral of the electric flux density D over a closed surface s enveloping the charge is

$$\oint_s D \cdot ds = Q \qquad (1.1)$$

If the surface has no enclosed charge, the integral is zero. This is the Gauss law.

The magnetomotive force F, or the line integral of the magnetic field strength H around a closed path l, is equal to the current enclosed, i.e.

$$\oint_l H \cdot dl = F = i_c + i_d \qquad (1.2)$$

This is the Ampère law with the addition of displacement current.

The Faraday law states that, around any closed path l encircling a magnetic flux ϕ that changes with time, there is an electric field, and the line integral of the electric field strength E around the path is

$$\oint_l E \cdot dl = e = -(\mathrm{d}\phi/\mathrm{d}t) \qquad (1.3)$$

Magnetic flux is a solenoidal quantity, i.e. it comprises a structure of closed loops; over any closed surface s in a magnetic field as much flux leaves the surface as enters it. The surface integral of the flux density B is therefore always zero, i.e.

$$\oint_s B \cdot ds = 0 \qquad (1.4)$$

To these four laws are added the *constitutive equations*, which relate the flux densities to the properties of the media in which the fields are established. The first two are, respectively, electric and magnetic field relations; the third relates conduction current density to the voltage gradient in a conducting medium; the fourth is a statement of the displacement current density resulting from a time rate of change of the electric flux density. The relations are

$$D = \varepsilon E; \quad B = \mu H; \quad J_c = \sigma E; \quad J_d = \partial D/\partial t$$

Table 1.23 Electrical properties of insulating materials

Typical approximate values (see also Section 1.4):

ε_r	relative permittivity	
E	electric strength	[MV/m]
$\tan\delta$	loss tangent	
θ	maximum working temperature	[°C]
k	thermal conductivity	[mW/(m K)]
G	density	[kg/m^3]

Material	ε_r	E	$\tan\delta$	θ	k	G
Air at n.t.p.	1.0	3	—	—	25	1.3
Alcohol	26	—	—	—	180	790
Asbestos	2	2	—	400	80	3000
paper	2	2	—	250	250	1200
Bakelite moulding	4	6	0.03	130	—	1600
paper	5	15	0.03	100	270	1300
Bitumen						
pure	2.7	1.6	—	50	150	1200
vulcanised	4.5	5	—	100	200	1250
Cellulose film	5.8	28	—	—	—	800
Cotton fabric						
dry	—	0.5	—	95	80	—
impregnated	—	2	—	95	250	—
Ebonite	2.8	50	0.005	80	150	1400
Fabric tape, impregnated	5	17	0.1	95	240	—
Glass						
flint	6.6	6	—	—	1100	4500
crown	4.8	6	0.02	—	600	2200
toughened	5.3	9	0.003	—	—	—
Gutta-percha	4.5	—	0.02	—	200	980
Marble	7	2	0.03	—	2600	2700
Mica	6	40	0.02	750	600	2800
Micanite	—	15	—	125	150	2200
Oil						
transformer	2.3	—	—	85	160	870
castor	4.7	8	—	—	—	970
Paper						
dry	2.2	5	0.007	90	130	820
impregnated	3.2	15	0.06	90	140	1100
Porcelain	5.7	15	0.008	1000	1000	2400
Pressboard	6.2	7	—	95	170	1100
Quartz						
fused	3.5	13	0.002	1000	1200	2200
crystalline	4.4	—	—	—	—	2700
Rubber						
pure	2.6	18	0.005	50	100	930
vulcanised	4	10	0.01	70	250	1500
moulding	4	10	—	70	—	—
Resin	3	—	—	—	—	1100
Shellac	3	11	—	75	250	1000
paper	5.5	11	0.05	80	—	1350
Silica, fused	3.6	14	—	—	—	—
Silk	—	—	—	95	60	1200
Slate	—	0.5	—	—	2000	2800
Steatite	—	0.6	—	1500	2000	2600
Sulphur	4	—	0.0003	100	220	2000
Water	70	—	—	—	570	1000
Wax (paraffin)	2.2	12	0.0003	35	270	860

In electrotechnology concerned with direct or low-frequency currents, the Maxwell equations are rarely used in the form given above. Equation (1.2), for example, appears as the number of amperes (or ampere-turns) required to produce in an area *a* the specified magnetic flux $\phi = Ba = \mu Ha$. Equation (1.3) in the form $e = -(\mathrm{d}\phi/\mathrm{d}t)$ gives the e.m.f. in a transformer primary or secondary turn. The concept of the 'magnetic circuit' embodies Equation (1.4). But when dealing with such field phenomena as the eddy currents in massive conductors, radio propagation or the transfer of energy along a transmission line, the Maxwell equations are the basis of analysis.

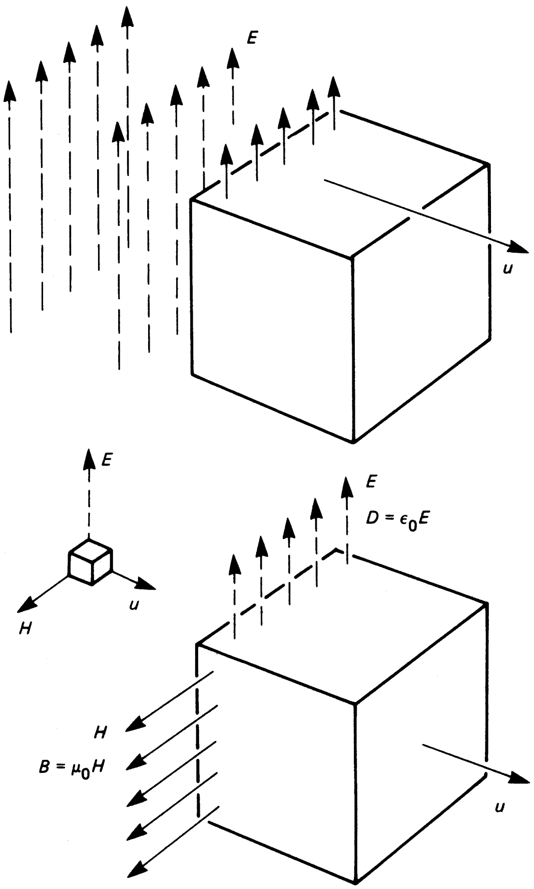


Figure 1.10 Electromagnetic wave propagation

1.5.3.2 Electromagnetic wave

The local 'induction' field of a charge at rest surrounds it in a predictable pattern. Let the position of the charge be suddenly displaced. The field pattern also moves, but because of the finite rate of propagation there will be a region in which the original field has not yet been supplanted by the new. At the instantaneous boundary the electric field pattern may be pictured as 'kinked', giving a transverse electric field component that travels away from the charge. Energy is propagated, because the transverse electric field is accompanied by an associated transverse magnetic field in accordance with the Ampère law.

Consider a unit cube of *free space* (*Figure 1.10*) approached by a transverse electric field of strength E at a velocity u in the specified direction. As E enters the cube, it produces therein an electric flux, of density $D=\varepsilon_0 E$ increasing at the rate Du. This is a displacement current which produces a magnetic field of strength H and flux density $B=\Leftarrow\mu_0 H$ increasing at the rate Bu. Then the E and H waves are mutually dependent:

$$Du = \varepsilon_0 Eu = H \qquad (1.5)\Leftarrow$$

$$Bu = \mu_0 Hu = E \qquad (1.6)\Leftarrow$$

Multiplication and division of (1.5) and (1.6) give

$$u = 1/\sqrt{(\varepsilon_0\mu_0)} = c \simeq 3 \times 10^8 \text{ m/s}$$

$$E/H = \sqrt{(\mu_0/\varepsilon_0)} = Z_0 \simeq 377\,\Omega$$

The velocity of propagation in free space is thus fixed; the ratio E/H is also fixed, and is called the intrinsic impedance. Further, $\frac{1}{2}\varepsilon_0 E^2 = \frac{1}{2}\mu_0 H^2$ [J/m^3], showing that the electric and magnetic energy densities are equal.

Propagation is normally maintained by charge acceleration which results from a high-frequency alternating current (e.g. in an aerial), so that waves of E and H of sinusoidal distribution are propagated with a wavelength dependent on the frequency (*Figure 1.11*). There is a fixed relation between the directions of E, H and the energy flow. The rate at which energy passes a fixed point is EH[W/m^2], and the direction of E is taken as that of the wave polarisation.

Plane wave transmission in a perfect homogeneous loss-free insulator takes place as in free space, except that ε_0 is replaced by $\varepsilon = \varepsilon_r\varepsilon_0$, where ε_r is the relative permittivity of the medium: the result is that both the propagation velocity and the intrinsic impedance are reduced.

When a plane wave from free space enters a material with conducting properties, it is subject to attenuation by reason of the I^2R loss. In the limit, a perfect conductor presents to the incident wave a complete barrier, reflecting the wave as a perfect mirror. A wave incident upon a general medium is partly reflected, and partly transmitted with attenuation and phase-change.

Table 1.24 gives the wavelength and frequency of free space electromagnetic waves with an indication of their technological range and of the physical origin concerned.

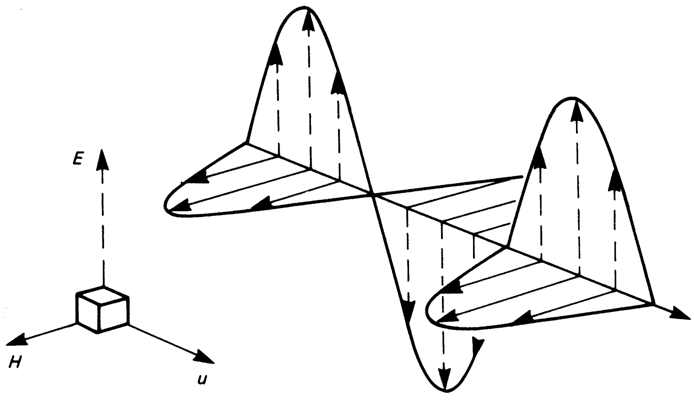


Figure 1.11 Electromagnetic wave

Table 1.24 Electromagnetic wave spectrum

Free space properties:

Electric constant	$\varepsilon_0 = 8.854\times10^{-12}$	F/m	Intrinsic impedance	$Z_0 = 376.8$	Ω
Magnetic constant	$\mu_0 = 4\pi\times10^{-7}$	H/m	Velocity	$c = 2.9979\times10^{8}$	m/s

The product of wavelength λ[m] and frequency f [Hz] is $f\lambda = c \simeq 3\times10^{8}$ [m/s].

λ	*f*	*Range*			*Origin*
	100 EHz	γ-rays			Transition between states of nuclei
10 pm					
	10	X-rays			Transition between states of inner electrons
100					
	1 EHz				
1 nm					Transition between states of outer electrons
	100	Ultraviolet			
10					
	10				
100					
	1 PHz		400 nm		
		Light	450	Vi	
1 μm			500	Bl	
	100				Molecular vibration
10			550	Gn	
	10		600	Yl	
100			650	Or	
	1 THz		700 nm	Rd	
1 mm					
	100				
10			microwave		
	10		radar		Vibration of conduction electrons
100					
	1 GHz		television		
1 m			frequency-modulated		
	100	Radio	short-wave		
10					
	10		medium-wave		
100					Alternating currents
	1 MHz		long-wave		
1 km					
	100				
10					
	10				
100					
	1 kHz				
1 Mm					
	100				
10		Power system			
	10 Hz				

2

Electrotechnology

M G Say PhD, MSc, CEng, ACGI, DIC, FIEE, FRSE
Heriot-Watt University
(Sections 2.1–2.6)

G R Jones PhD, DSc, CEng, FIEE, MInst P
University of Liverpool
(Section 2.7)

Contents

- 2.1 Nomenclature 2/3
  - 2.1.1 Circuit phenomena 2/3
  - 2.1.2 Electrotechnical terms 2/4
- 2.2 Thermal effects 2/6
  - 2.2.1 Resistance 2/6
  - 2.2.2 Heating and cooling 2/8
- 2.3 Electrochemical effects 2/10
  - 2.3.1 Electrolysis 2/10
  - 2.3.2 Cells 2/11
- 2.4 Magnetic field effects 2/12
  - 2.4.1 Magnetic circuit 2/12
  - 2.4.2 Magnetomechanical effects 2/14
  - 2.4.3 Electromagnetic induction 2/15
  - 2.4.4 Inductance 2/17
- 2.5 Electric field effects 2/19
  - 2.5.1 Electrostatics 2/20
  - 2.5.2 Capacitance 2/20
  - 2.5.3 Dielectric breakdown 2/22
  - 2.5.4 Electromechanical effects 2/22
- 2.6 Electromagnetic field effects 2/23
  - 2.6.1 Movement of charged particles 2/23
  - 2.6.2 Free space propagation 2/23
  - 2.6.3 Transmission line propagation 2/23
- 2.7 Electrical discharges 2/25
  - 2.7.1 Introduction 2/25
  - 2.7.2 Types of discharge 2/26
  - 2.7.3 Discharge–network interaction 2/27
  - 2.7.4 Discharge applications 2/28

Electrotechnology concerns the electrophysical and allied principles applied to practical electrical engineering. A completely general approach is not feasible, and many separate *ad hoc* technologies have been developed using simplified and delimited areas adequate for particular applications.

In establishing a technology it is necessary to consider whether the relevant applications can be dealt with (a) in *macroscopic* terms of physical qualities of materials in bulk (as with metallic conduction or static magnetic fields); or (b) in *microscopic* terms involving the microstructure of materials as an essential feature (as in domain theory); or (c) in molecular, atomic or *subatomic* terms (as in nuclear reaction and semi-conduction). There is no rigid line of demarcation, and certain technologies must cope with two or more such subdivisions at once. Electrotechnology thus tends to become an assembly of more or less discrete (and sometimes apparently unrelated) areas in which methods of treatment differ widely.

To a considerable extent (but not completely), the items of plant with which technical electrical engineering deals—generators, motors, feeders, capacitors, etc.—can be represented by *equivalent circuits* or *networks* energised by an electrical *source*.

For the great majority of cases within the purview of ‘heavy electrical engineering’ (that is, generation, transmission and utilisation for power purposes, as distinct from telecommunications), a *source* of electrical energy is considered to produce a *current* in a conducting *circuit* by reason of an *electromotive force* acting against a property of the circuit called *impedance*. The behaviour of the circuit is described in terms of the energy fed into the circuit by the source, and the nature of the conversion, dissipation or storage of this energy in the several circuit components.

Electrical phenomena, however, are only in part associated with conducting circuits. The generalised basis is one of magnetic and electrical fields in free space or in material media. The fundamental starting point is the conception contained in Maxwell’s electromagnetic equations (Section 1.5.3), and in this respect the voltage and currents in a circuit are only representative of the fundamental field phenomena within a restricted range. Fortunately, this range embraces very nearly the whole of ‘heavy’ electrical engineering practice. The necessity for a more comprehensive viewpoint makes itself apparent in connection with problems of long-line transmission; and when the technique of ultra-high-frequency work is reached, it is necessary to give up the familiar circuit ideas in favour of a whole-hearted application of field principles.

2.1 Nomenclature

2.1.1 Circuit phenomena

Figure 2.1 shows in a simplified form a hypothetical circuit with a variety of electrical energy sources and a representative selection of devices in which the energy received from the source is converted into other forms, or stored, or both. The forms of variation of the current or voltage are shown in *Figure 2.2*. In an actual circuit the current may change in a quite arbitrary fashion as indicated at (*a*): it may rise or fall, or reverse its direction, depending on chance or control. Such random variation is inconveniently difficult to deal with, and engineers prefer to simplify the conditions as much as possible. For example (*Figure 2.2*(*b*)), the current may be assumed to be rigidly constant, in which case it is termed a *direct current*. If the current be deemed to reverse cyclically according to a sine function, it becomes a *sinusoidal alternating current* (*c*). Less simple waveforms, such as (*d*), may be dealt with by application of Fourier’s theorem, thus making it possible to calculate a great range of practical cases—such as those involving rectifiers—in which the sinusoidal waveform assumption is inapplicable. The cases shown in (*b*), (*c*) and (*d*) are known as *steady states*, the current (or voltage) being assumed established for a considerable time before the circuit is investigated. But since the electric circuit is capable of storing energy, a change in the circuit may alter the conditions so as to cause a redistribution of circuit energy. This occurs with a circulation of *transient* current. An example of a simple oscillatory transient is shown in *Figure 2.2*(*e*).

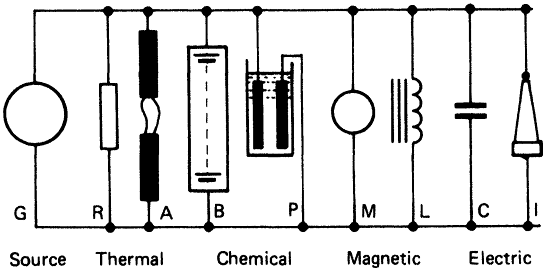


Figure 2.1 Typical circuit devices. G, source generator; R, resistor; A, arc; B, battery; P, plating bath; M, motor; L, inductor; C, capacitor; I, insulator

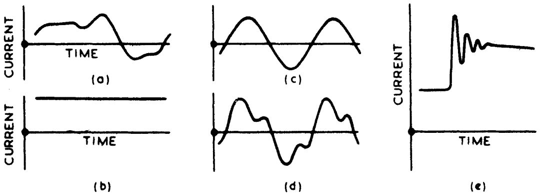


Figure 2.2 Modes of current (or voltage) variation

The calculation of circuits in which direct currents flow is comparatively straightforward. For sine wave alternating current circuits an algebra has been developed by means of which problems can be reduced to a technique very similar to that of d.c. circuits. Where non-sinusoidal waveforms are concerned, the treatment is based on the analysis of the current and voltage waves into fundamental and harmonic sine waves, the standard sine wave method being applied to the fundamental and to each of the harmonics. In the case of transients, a more searching investigation may be necessary, but there are a number of common modes in which transients usually occur, and (so long as the circuit is relatively simple) it may be possible to select the appropriate mode by inspection.

Circuit *parameters*—resistance, inductance and capacitance—may or may not be constant. If they are not, approximation, linearising or step-by-step computation is necessary.

2.1.1.1 Electromotive-force sources

Any device that develops an electromotive force (e.m.f.) capable of sustaining a current in an electric circuit must be associated with some mode of energy conversion into the electrical from some different form. The modes are (1) mechanical/electromagnetic, (2) mechanical/electrostatic, (3) chemical, (4) thermal, and (5) photoelectric.

2.1.2 Electrotechnical terms

The following list includes the chief terms in common use. The symbols and units employed are given in *Table 2.1*.

Admittance: The ratio between current and voltage in r.m.s. terms for sinusoidally varying quantities.

Admittance operator: The ratio between current and voltage in operational terms.

Table 2.1 Electrotechnical symbols and units

Quantity	*Quantity symbol*	*Unit name*	*Unit symbol*
Admittance	Y	siemens	S
Ampere-turn	—	ampere-turn	A-t
Angular frequency	$\omega = 2\pi f$	radian/second	rad/s
Capacitance	C	farad	F
Charge	Q	coulomb	C
Conductance	G	siemens	S
Conductivity	γ, $\sigma\psi$	siemens/metre	S/m
Current	I	ampere	A
Current density, linear	A	ampere/metre	A/m
Current density, surface	J	ampere/metre-square	A/m^2
Electric field strength	E	volt/metre	V/m
Electric flux	Q	coulomb	C
Electric flux density	D	coulomb/metre-square	C/m^2
Electric space constant	ϵ_0	farad/metre	F/m
Electromotive force	E	volt	V
Force	F, f	newton	N
Frequency	f	hertz	Hz
Impedance	Z	ohm	Ω
Inductance, mutual	L_{jk}, M	henry	H
Inductance, self-	L	henry	H
Linkage		weber-turn	Wb-t
Loss angle	$\delta\psi$	radian	rad
Magnetic field strength	H	ampere/metre	A/m
Magnetic flux	Φ	weber	Wb
Magnetic flux density	B	tesla	T
Magnetic space constant	$\mu_0 = 4\pi/10^7$	henry/metre	H/m
Magnetomotive force	F	ampere (-turn)	A, A-t
Period	T	second	s
Permeability, absolute	$\mu = \mu_r\mu_0$	henry/metre	H/m
Permeability, free space	μ_0	henry/metre	H/m
Permeability, relative	μ_r	—	—
Permeance	Λ	weber/ampere (-turn)	Wb/A, Wb/A-t
Permittivity, absolute	$\epsilon = \epsilon_r\epsilon_0$	farad/metre	F/m
Permittivity, free space	ϵ_0	farad/metre	F/m
Permittivity relative	ϵ_r	—	—
Phase angle	$\phi\psi$	radian	rad
Potential	V	volt	V
Potential difference	V, U	volt	V
Potential gradient	E	volt/metre	V/m
Power, active	P	watt	W
Power, apparent	S	volt-ampere	V-A
Power, reactive	Q	var	var
Quantity	Q	coulomb	C
Reactance	X	ohm	Ω
Reluctance	S	ampere (-turn)/weber	A/Wb, A-t/Wb
Resistance	R	ohm	Ω
Resistivity	$\rho\psi$	ohm-metre	Ω-m
Susceptance	B	siemens	S
Time constant	$\tau\psi$	second	s
Voltage	V	volt	V
Voltage gradient	E	volt/metre	V/m

Ampere-turns: The product of the number of turns of a circuit and the current flowing in them.

Angular frequency: The number of periods per second of a periodically varying quantity multiplied by 2π.

Capacitance: The property of a conducting body by virtue of which an electric charge has to be imparted to it to produce a difference of electrical potential between it and the surrounding bodies. The ratio between the charge on a conductor and its potential when all neighbouring conductors are at zero (earth) potential. The ratio between the charge on each electrode of a capacitor and the potential difference between them.

Capacitor: A device having capacitance as a chief property.

Charge: The excess of positive or negative electricity on a body or in space.

Coercive force: The demagnetising force required to reduce to zero the remanent flux density in a magnetic body.

Coercivity: The value of the coercive force when the initial magnetisation has the saturation value.

Complexor: A non-vectorial quantity expressible in terms of a complex number.

Conductance: For steady direct currents, the reciprocal of the resistance. For sinusoidal alternating currents, the resistance divided by the square of the impedance.

Conductivity: The reciprocal of resistivity.

Core loss (iron loss): The loss in a magnetic body subject to changing magnetisation, resulting from hysteresis and eddy current effects.

Current: The flow or transport of electric charges along a path or around a circuit.

Current density: The current per unit area of a conductor, or per unit width of an extended conductor.

Diamagnetic: Having a permittivity less than that of free space.

Dielectric loss: The loss in an insulating body, resulting from hysteresis, conduction and absorption.

Displacement current: The current equivalent of the rate of change of electric flux with time.

Eddy current: The current electromagnetically induced in a conductor lying in a changing magnetic field.

Electric field: The energetic state of the space between two oppositely charged conductors.

Electric field strength: The mechanical force per unit charge on a very small charge placed in an electric field. The negative voltage gradient.

Electric flux: The electric field, equal to the charge, between oppositely charged conductors.

Electric flux density: The electric flux per unit area.

Electric space constant: The permittivity of free space.

Electric strength: The property of an insulating material which enables it to withstand electric stress; or the stress that it can withstand without breakdown.

Electric stress: The electric field intensity, which tends to break down the insulating property of an insulating material.

Electromagnetic field: A travelling field having electric field and magnetic field components and a speed of propagation depending on the electrical properties of the ambient medium.

Electromagnetic induction: The production of an electromotive force in a circuit by a change of magnetic linkage through the circuit. The e.m.f. so produced is an *induced e.m.f.*, and any current that may result therefrom is an *induced current*.

Electromotive force (*e.m.f.*): That quality which tends to cause a movement of charges around a circuit. The direction is that of the movement of positive charges. E.m.f. is measured by the amount of energy developed by transfer of unit positive charge. The term is applied to sources that convert electrical energy to or from some other kind (chemical, mechanical, thermal, etc.).

Ferromagnetic: Having a permeability much greater than that of free space, and varying with the magnetic flux density.

Force: The cause of the mechanical displacement, motion, acceleration and deformation of massive bodies.

Frequency: The number of repetitions of a cyclically time-varying quantity in unit time.

Hysteresis: The phenomenon by which an effect in a body depends not only on the present cause, but also on the previous state of the body. In *magnetisation* a flux density produced by a given magnetic field intensity depends on the previous magnetisation history. A comparable effect occurs in the *electrification* of insulating materials. In cyclic changes hysteresis is the cause of energy loss.

Immitance: A circuit property that can be either impedance or admittance.

Impedance: The ratio between voltage and current in r.m.s. terms for sinusoidally varying quantities.

Impedance operator: The ratio between voltage and current in operational terms.

Inductance: The property of a circuit by virtue of which the passage of a current sets up magnetic linkage and stores magnetic energy. If the linkage of a circuit arises from the current in another circuit, the property is called *mutual inductance*.

Inductor: A device having inductance as a chief property.

Insulation resistance: The resistance under prescribed conditions between two conductors or conducting systems normally separated by an insulating medium.

I^2R loss (copper loss): The loss (converted into heat) due to the passage of a current through the resistance of a conductor.

Line of flux (line of force): A line drawn in a field to represent the direction of the flux at any point.

Linkage: The summation of the products of elements of magnetic flux and the number of turns of the circuit they embrace in a given direction.

Loss angle: The phase angle by which the current in a reactor fails to lead (or lag) the voltage by $\frac{1}{2}\pi$ rad under sinusoidal conditions. The tangent of this angle is called the *loss tangent*.

Magnetic circuit: The closed path followed by a magnetic flux.

Magnetic field: The energetic state of the space surrounding an electric current.

Magnetic field strength: The cause at any point of a magnetic circuit of the magnetic flux density there.

Magnetic flux: The magnetic field, equal to the summation of flux density and area, around a current. A phenomenon in the neighbourhood of currents or magnets. The magnetic flux through any area is the surface integral of the magnetic flux density through the surface. Unit magnetic flux is that flux, the removal of which from a circuit of unit resistance causes unit charge to flow in the circuit; or in an open turn produces a voltage–time integral of unity.

Magnetic flux density: The magnetic flux per unit area at a point in a magnetic field, the area being oriented to give a maximum value to the flux. The normal to the area is the direction of the flux at the point. The direction of the current produced in the electric circuit on removal of the flux, and the positive direction of the flux, have the relation of a right-handed screw.

Magnetic leakage: That part of a magnetic flux which follows such a path as to make it ineffective for the purpose desired.

Magnetic potential difference: A difference between the magnetic states existing at two points which produces a magnetic field between them. It is equal to the line integral of the magnetic field intensity between the points, except in the presence of electric currents.

Magnetic space constant: The permeability of free space.

Magnetising force: The same as magnetic field strength.

Magnetomotive force: Along any path, the line integral of the magnetic field strength along that path. If the path is closed, the line integral is equal to the total magnetising current in ampere-turns.

Paramagnetic: Having a permeability greater than that of free space.

Period: The time taken by one complete cycle of a waveform.

Permeability: The ratio of the magnetic flux density in a medium or material at a point to the magnetic field strength at the point. The *absolute permeability* is the product of the *relative permeability* and the *permeability of free space* (magnetic space constant).

Permeance: The ratio between the magnetic flux in a magnetic circuit and the magnetomotive force. The reciprocal of reluctance.

Permittivity: The ratio between the electric flux density in a medium or material at a point and the electric field strength at the point. The *absolute permittivity* is the product of the *relative permittivity* and the *permittivity of free space* (electric space constant).

Phase angle: The angle between the phasors that represent two alternating quantities of sinusoidal waveform and the same frequency.

Phasor: A sinusoidally varying quantity represented in the form of a complex number.

Polarisation: The change of the electrical state of an insulating material under the influence of an electric field, such that each small element becomes an electric dipole or doublet.

Potential: The electrical state at a point with respect to potential zero (normally taken as that of the earth). It is measured by the work done in transferring unit charge from pontential zero to the point.

Potential difference: A difference between the electrical states existing at two points tending to cause a movement of positive charges from one point to the other. It is measured by the work done in transferring unit charge from one point to the other.

Potential gradient: The potential difference per unit length in the direction in which it is a maximum.

Power: The rate of transfer, storage, conversion or dissipation of energy. In sinusoidal alternating current circuits the *active power* is the mean rate of energy conversion; the *reactive power* is the peak rate of circulation of stored energy; the *apparent power* is the product of r.m.s. values of voltage and current.

Power factor: The ratio between active power and apparent power. In sinusoidal alternating current circuits the power factor is cos ϕ, where ϕ is the phase angle between voltage and current waveforms.

Quantity: The product of the current and the time during which it flows.

Reactance: In sinusoidal alternating current circuits, the quantity ωL or $1/\omega C$, where L is the inductance, C is the capacitance and ω is the angular frequency.

Reactor: A device having reactance as a chief property; it may be an inductor or a capacitor. A *nuclear reactor* is a device in which energy is generated by a process of nuclear fission.

Reluctance: The ratio between the magnetomotive force acting around a magnetic circuit and the resulting magnetic flux. The reciprocal of permeance.

Remanence: The remanent flux density obtained when the initial magnetisation reaches the saturation value for the material.

Remanent flux density: The magnetic flux density remaining in a material when, after initial magnetisation, the magnetising force is reduced to zero.

Residual magnetism: The magnetism remaining in a material after the magnetising force has been removed.

Resistance: That property of a material by virtue of which it resists the flow of charge through it, causing a dissipation of energy as heat. It is equal to the constant potential difference divided by the current produced thereby when the material has no e.m.f. acting within it.

Resistivity: The resistance between opposite faces of a unit cube of a given material.

Resistor: A device having resistance as a chief property.

Susceptance: The reciprocal of reactance.

Time constant: The characteristic time describing the duration of a transient phenomenon.

Voltage: The same as potential difference.

Voltage gradient: The same as potential gradient.

Waveform: The graph of successive instantaneous values of a time-varying physical quantity.

2.2 Thermal effects

2.2.1 Resistance

That property of an electric circuit which determines for a given current the rate at which electrical energy is converted into heat is termed *resistance*. A device whose chief property is resistance is a *resistor*, or, if variable, a *rheostat*. A current I flowing in a resistance R develops heat at the rate

$$P = I^2R \text{ joule/second or watts}$$

a relation expressing *Joule's law*.

2.2.1.1 Voltage applied to a resistor

In the absence of any energy storage effects (a physically unrealisable condition), the current in a resistor of value R is I when the voltage across it is V, in accordance with the relation $I = V/R$. If a *steady* p.d. V be suddenly applied to a resistor R, the current instantaneously assumes the value given, and energy is expended at the rate $P = I^2R$ watts, continuously. No transient occurs. If a constant frequency, constant amplitude *sine wave* voltage v is applied, the current i is at every instant given by $i = v/R$, and in consequence the current has also a sine waveform, provided that the resistance is linear. The instantaneous rate of energy dissipation depends on the instantaneous current: it is $p = vi = i^2R$. Should the applied voltage be non-sinusoidal, the current has (under the restriction mentioned) an exactly similar waveform. The three cases are illustrated in *Figure 2.3*.

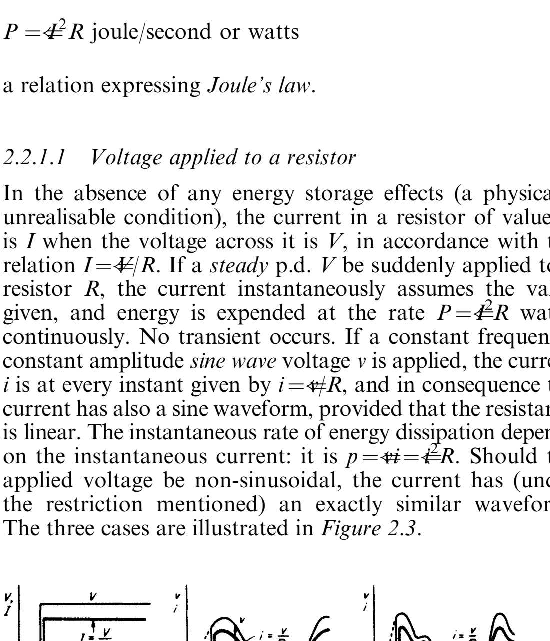


Figure 2.3 Voltage applied to a pure resistor

In the case of alternating waveform, the average rate of energy dissipation is given by $P = I^2R$, where I is the *root-mean-square* current value.

2.2.1.2 Voltage–current characteristics

For a given resistor R carrying a constant current I, the p.d. is $V = IR$. The ratio $R = V/I$ may or may not be invariable. In some cases it is sufficient to assume a degree of constancy, and calculation is generally made on this assumption. Where the variations of resistance are too great to make the assumption reasonably valid, it is necessary to resort to less simple analysis or to graphical methods.

A constant resistance is manifested by a constant ratio between the voltage across it and the current through it, and by a straight-line graphical relation between I and V (*Figure 2.4*(*a*)), where $R = V/I = \cot\theta$. This case is typical of metallic resistance wires at constant temperature.

Certain circuits exhibit *non-linear* current–voltage relations (*Figure 2.4*(*b*)). The non-linearity may be *symmetrical* or *asymmetrical*, in accordance with whether the conduction characteristics are the same or different for the two current-flow directions. Rectifiers are an important class of non-linear, asymmetrical resistors.

A hypothetical device having the current–voltage characteristic shown in *Figure 2.4*(*c*) has, at an operating condition represented by the point P, a current I_d and a p.d. V_d. The ratio $R_d = V_d/I_d$ is its *d.c. resistance* for the given condition. If a small alternating voltage Δv_a be applied under the same condition (i.e. superimposed on the p.d. V_d), the current will fluctuate by Δi_a and the ratio $r_a = \Delta v_a/\Delta i_a$ is the *a.c.* or *incremental resistance* at P. The d.c. resistance is also obtainable from $R_d = \cot\theta$, and the a.c. resistance from $r_a = \cot\alpha$. In the region of which Q is a representative point, the a.c. resistance is *negative*, indicating that the device is capable of giving a small output of a.c. power, derived from its greater d.c. input. It remains in sum an energy dissipator, but some of the energy is returnable under suitable conditions of operation.

2.2.1.3 D.c. or ohmic resistance: linear resistors

The d.c. or ohmic resistance of linear resistors (a category confined principally to metallic conductors) is a function of the dimensions of the conducting path and of the *resistivity* of the material from which the conductor is made. A wire of length l, cross-section a and resistivity ρ has, at constant given temperature, a resistance

$$R = \rho l/a \text{ ohms}$$

where ρ, l and a are in a consistent system of dimensions (e.g. l in metres, a in square metres, ρ in ohms per 1 m length and 1 m^2 cross-section—generally contracted to ohm-metres). The expression above, though widely applicable, is true only on the assumption that the current is uniformly distributed over the cross-section of the conductor and flows in paths parallel to the boundary walls. If this assumption is inadmissible, it is necessary to resort to integration or the use of current-flow lines. *Figure 2.5* summarises the expressions for the resistance of certain arrangements and shapes of conductors.

Resistivity The resistivity of conductors depends on their composition, physical condition (e.g. dampness in the case of non-metals), alloying, manufacturing and heat treatment, chemical purity, mechanical working and ageing. The *resistance-temperature coefficient* describes the rate of change of resistivity with temperature. It is practically 0.004 Ω/°C at 20°C for copper. Most pure metals have a resistivity that rises with temperature. Some alloys have a very small coefficient. Carbon is notable in that its resistivity decreases markedly with temperature rise, while uranium dioxide has a resistivity which falls in the ratio 50:1 over a range of a few hundred degrees. *Table 2.2* lists the resistivity ρ and the resistance-temperature coefficient α for a number of representative materials. The effect of temperature is assessed in accordance with the expressions

$$R_1 = R_0(1 + \alpha\theta_1); \quad R_2/R_1 = (1 + \alpha\theta_2)/(1 + \alpha\theta_1);$$

or

$$R_2 = R_1[1 + \alpha(\theta_2 - \theta_1)]$$

where R_0, R_1 and R_2 are the resistances at temperatures 0, θ_1 and θ_2, and α is the resistance-temperature coefficient at 0°C.

2.2.1.4 Liquid conductors

The variations of resistance of a given aqueous solution of an electrolyte with temperature follow the approximate rule:

$$R_\theta = R_0/(1 + 0.03\theta)$$

where θ is the temperature in degrees Celsius. The conductivity (or reciprocal of resistivity) varies widely with the percentage strength of the solution. For low concentrations the variation is that given in *Table 2.2*.

2.2.1.5 Frequency effects

The resistance of a given conductor is affected by the frequency of the current carried by it. The simplest example is that of an isolated wire of circular cross-section. The inductance of the central parts of the conductor is greater than that of the outside skin because of the additional flux linkages due to the internal magnetic flux lines. The impedance of the central parts is consequently greater, and the current flows mainly at and near the surface of the conductor, where the impedance is least. The useful cross-section of the conductor is less than the actual area, and the effective resistance is consequently higher. This is called the *skin effect*. An analogous phenomenon, the *proximity effect*, is due to mutual inductance between conductors arranged closely parallel to one another.

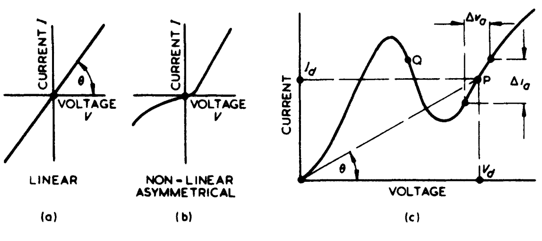


Figure 2.4 Current–voltage characteristics

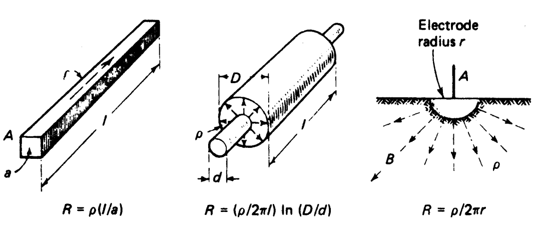


Figure 2.5 Resistance in particular cases

Table 2.2 Conductivity of aqueous solutions* (mS/cm)

Concentration (%)	*a*	*b*	*c*	*d*	*e*	*f*	*g*	*h, j*	*k*	*l, m*
1	40	18	12	10	10	8	6	4	3	3
2	72	35	23	20	20	16	12	8	6	6
3	102	51	34	30	30	24	18	12	9	8
4	130	65	44	39	39	32	23	16	11	10
5		79	55	48	47	39	28	20	13	11
7.5		110	79	69	67	54	39	29	18	16
10			99	90	85	69		31	22	20

*(a) NaOH, caustic soda; (b) NH_4Cl, sal ammoniac; (c) NaCl, common salt; (d) $NaNO_3$, Chilean saltpetre; (e) $CaCl_2$, calcium chloride; (f) $ZnCl_2$, zinc chloride; (g) $NaHCO_3$, baking soda; (h) Na_2CO_3, soda ash; (j) Na_2SO_4, Glauber's salt; (k) $Al_2(SO_4)_3K_2SO_4$, alum; (l) $CuSO_4$, blue vitriol; (m) $ZnSO_4$, white vitriol.

The effects depend on conductor size, frequency f of the current, resistivity ρ and permeability μ of the material. For a circular conductor of diameter d the increase of effective resistance is proportional to $d^2 f\mu/\rho$. At power frequencies and for small conductors the effect is negligible. It may, however, be necessary to investigate the skin and proximity effects in the case of large conductors such as bus-bars.

2.2.1.6 *Non-linear resistors*

Prominent among non-linear resistors are electric arcs; also silicon carbide and similar materials.

Arcs An electric arc constitutes a conductor of somewhat vague dimensions utilising electronic and ionic conduction in a gas. It is strongly affected by physical conditions of temperature, gas pressure and cooling. In air at normal pressure a d.c. arc between copper electrodes has the voltage–current relation given approximately by $V = 30 + 10/I + l[1 + 3/I]10^3$ for a current I in an arc length l metres. The expression is roughly equivalent to 10 V/cm for large currents and high voltages. The current density varies between 1 and 1000 A/mm^2, being greater for large currents because of the *pinch* effect. See also Section 2.7.

Silicon carbide Conducting pieces of this material have a current–voltage relation expressed approximately by $I = KV^x$, where x is usually between 3 and 5. For rising voltage the current increases very rapidly, making silicon carbide devices suitable for circuit protection and the discharge of excess transmission-line surge energy.

2.2.2 Heating and cooling

The *heating* of any body such as a resistor or a conducting circuit having inherent resistance is a function of the losses within it that are developed as heat. (This includes core and dielectric as well as ohmic I^2R losses, but the effective value of R may be extended to cover such additional losses.) The *cooling* is a function of the facilities for heat dissipation to outside media such as air, oil or solids, by radiation, conduction and convection.

2.2.2.1 *Rapid heating*

If the time of heating is short, the cooling may be ignored, the temperature reached being dependent only on the rate of development of heat and the thermal capacity. If p is the heat development per second in joules (i.e. the power in watts), G the mass of the heated body in kilograms and c its specific heat in joules per kilogram per kelvin, then

$$Gc \cdot d\theta = p \cdot dt, \quad \text{giving} \quad \theta = (1/Gc)\int p \cdot dt$$

For steady heating, the temperature rise is p/Gc in Kelvin per second.

Standard annealed copper is frequently used for the windings and connections of electrical equipment. Its density is $G = 8900$ kg/m^3 and its resistivity at 20°C is 0.017 μΩ-m; at 75°C it is 0.021 μΩ-m. A conductor worked at a current density J (in amperes per square metre) has a specific loss (watts per kilogram) of $\rho J^2/8900$. If $J = 2.75$ MA/m^2 (or 2.75 A/mm^2), the specific loss at 75°C is 17.8 W/kg, and its rate of self-heating is $17.8/375 = 0.048$°C/s.

2.2.2.2 *Continuous heating*

Under prolonged steady heating a body will reach a temperature rise above the ambient medium of $\theta_m = p/A\lambda$, where A is the cooling surface area and λ the specific heat dissipation (joules per second per square metre of surface per degree Celsius temperature rise above ambient). The expression is based on the assumption, roughly true for moderate temperature rises, that the rate of heat emission is proportional to the temperature rise. The specific heat dissipation λ is compounded of the effects of *radiation*, *conduction* and *convection*.

Radiation The heat radiated by a surface depends on the absolute temperature T (given by $T = \theta + 273$, where θ is the Celsius temperature), and on its character (surface smoothness or roughness, colour, etc.). The Stefan law of heat radiation is

$$p_r = 5.7eT^4 \times 10^{-8} \text{ watts per square metre}$$

where e is the coefficient of radiant emission, always less than unity, except for the perfect 'black body' surface, for which $e = 1$. The radiation from a body is independent of the temperature of the medium in which it is situated. The process of radiation of a body to an exterior surface is accompanied by a re-absorption of part of the energy when re-radiated by that surface. For a small spherical radiating body inside a large and/or black spherical cavity, the radiated power is given by the Stefan-Boltzmann law:

$$p_r = 5.7e_1[T_1^4 - T_2^4]10^{-8} \text{ watts per square metre}$$

where T_1 and e_1 refer to the body and T_2 to the cavity.

The emission of radiant heat from a perfect black body surface is independent of the roughness or corrugation of the surface. If $e < 1$, however, there is some increase of radiation if the surface is rough.

Conduction The conduction of heat is a function of the thermal or temperature gradient and the thermal resistivity, the latter being defined as the temperature difference in degrees Celsius across a path of unit length and unit section required for the continuous transmission of 1 W. Thus, the heat conducted per unit area along a path of length x in a material of thermal resistivity ρ for a temperature difference of θ is

$$p_{\mathrm{d}} = \theta/\rho x \text{ watts}$$

Resistivities for metals are very low. For insulating materials such as paper, $\rho = 5$–10; for *still* air, $\rho = 20$ W per °C per m and per m^2, approximately.

Convection Convection currents in liquids and gases (e.g. oil and air) are always produced near a heated surface unless baffled. Convection adds greatly to heat dissipation, especially if artificially stimulated (as in force cooling by fans). Experiment shows that a rough surface dissipates heat by convection more readily than a smooth one, and that high fluid speeds are essential to obtain *turbulence* as opposed to *stream-line* flow, the former being much more efficacious.

Convection is physically a very complex phenomenon, as it depends on small changes in buoyancy resulting from temperature rise due to heating. Formulae for dissipation of heat by convection have a strongly empirical basis, the form and orientation of the convection surfaces having considerable influence.

Cooling coefficient For electrical purposes the empirically derived coefficient of emission λ, or its reciprocal $1/\lambda$, are employed for calculations on cooling and temperature rise of wires, resistors, machines and similar plant.

2.2.2.3 *Measurement of temperature rise*

The temperature rise of a device developing heat can be measured (a) by a thermometer placed in contact with the surface whose temperature is required, (b) by resistance-temperature detectors or thermocouples on the surface of, or embedded in, the device, or (c) by the measurement of resistance (in the case of conducting circuits), using the known resistance-temperature coefficient. These methods measure different temperatures, and do not give merely alternative estimates of the same thing.

2.2.2.4 *Heating and cooling cycles*

In some cases a device (such as a machine or one of its parts) developing internal heat may be considered as sufficiently homogeneous to apply the exponential law. Suppose the device to have a temperature rise θ after the lapse of a time t. In an element of time dt a small temperature rise dθ takes place. The heat developed is $p\cdot\mathrm{d}t$, the heat stored is $Gh\cdot\mathrm{d}\theta$, and the heat dissipated is $A\theta\lambda\cdot\mathrm{d}t$. Since the heat stored and dissipated together equal the total heat produced,

$$Gh \cdot \mathrm{d}\theta + A\theta\lambda \cdot \mathrm{d}t = p \cdot \mathrm{d}t$$

the solution of which is

$$\theta = \theta_{\mathrm{m}}[1 - \exp(-t/\tau)]$$

where θ_{m} is the final steady temperature rise, calculated from $\theta_{\mathrm{m}} = p/A\lambda$, and $\tau = Gh/A\lambda$ is called the *heating time constant*. For the lapse of time t equal to the time constant

$$\theta = \theta_{\mathrm{m}}[1 - \exp(-1)] = 0.632\theta_{\mathrm{m}}$$

When a heated body cools owing to a reduction or cessation of internal heat production, the temperature–time relation is the exponential function

$$\theta = \theta_{\mathrm{m}} \exp(-t/\tau_1)$$

where τ_1 is the *cooling time constant*, not necessarily the same as that for heating conditions.

Both heating and cooling as described are examples of *thermal transients*, and the laws governing them are closely analogous to those concerned with transient electric currents, in which exponential time relations also occur.

2.2.2.5 *Fusing currents*

For a given diameter d, the heat developed by a wire carrying a current I is inversely proportional to d^3, because an increase of diameter reduces the current density in proportion to the increase of area, and the emitting surface is increased in proportion to the diameter. The temperature rise is consequently proportional to I^2/d^3. If the temperature is raised to the fusing or melting point, $\theta = aI^2/d^3$ and the fusing current is

$$I = \surd(\theta d^3/a) = kd^{3/2}$$

This is *Preece's law*, from which an estimate may be made of the fusing current of a wire of given diameter, provided that k is known. The exponent 3/2 and the value of k are both much affected by enclosure, conduction of heat by terminals, and similar physical conditions.

It is obvious that any rule regarding suitable current densities giving a value regardless of the diameter is likely to be uneconomically low for small wires and excessive for large ones. Further, the effects of length and enclosure make a direct application of Preece's law unreliable. For small wires the exponent x in the term d^x may be 1.25–1.5, and for larger wires it may exceed 1.5.

2.2.2.6 *Thermo-e.m.f.s*

An effect known as the *thermoelectric effect* or *Seebeck effect* is that by which an e.m.f. is developed due to a difference of temperature between two junctions of dissimilar conductors in the same circuit. The *Thomson effect* or *Kelvin effect* is (a) that an e.m.f. is developed due to a difference of temperature between two parts of the same conductor, and (b) that an absorption or liberation of heat takes place when a current flows from a hotter to a colder part of the same material. The *Peltier effect* describes the liberation or absorption of heat at a joint where current passes from one material to another, whereby the joint becomes heated or cooled.

In *Figure 2.6*(*a*)–(*c*) the symbols are absolute temperature T, thermo-e.m.f. E and rate of heat production or absorption Q. The *Seebeck coefficient* (*a*) is the e.m.f. per degree difference between hot and cold junctions:

$$\alpha_{\mathrm{S}} = \Delta E/\Delta T$$

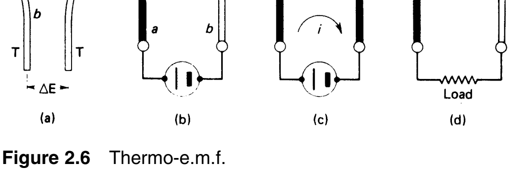


Figure 2.6 Thermo-e.m.f.

Typical e.m.f.s for a number of common junctions are given in *Table 2.3*. In the Peltier effect (*b*) a rate of heat generation (reversible, and distinct from the irreversible I^2R heat) results from the passage of a current *i* through the different conductors A and B. The *Peltier coefficient* is

$$\alpha_P = Q/i$$

The Thomson effect (*c*) concerns the rate of reversible heat when a current *i* flows through a length of homogeneous conductor across which there is a temperature difference. The *Thomson coefficient* is

$$\alpha_T = \Delta Q/i \cdot \Delta T$$

The relation between the Seebeck and Peltier coefficients is important: it is

$$\alpha_S = \alpha_P/T$$

The Seebeck coefficient is the more easily measured, but the Peltier coefficient determines the cooling effect of a thermoelectric refrigerator.

Table 2.3 Thermocouple e.m.f.s (mV): cold junction at 0 °C

Hot-junction temperature (°C)	*Platinum/* $^{87}Pt - {}^{13}Rh$	*Chromel/ Alumel*	*Iron/ Eureka*	*Copper/ Eureka*
100	0.65	4.1	5	4
200	1.46	8.1	11	9
400	3.40	16.4	22	21

2.2.2.7 *Thermoelectric devices*

If a current flows through a thermocouple (*Figure 2.6*(*b*)), with one junction in thermal contact with a heat sink, the other removes heat from a source. The couple must comprise conductors with positive and negative Seebeck coefficients, respectively. The arrangement is a *refrigerator* with the practical advantages of simplicity and silence.

A heat source applied to a junction develops an e.m.f. that will circulate a current in an external load (*Figure 2.6*(*d*)). If semiconductors of low thermal conductivity are used in place of metals for the couple elements, a better efficiency is obtainable because heat loss by conduction is reduced. The couples in *Figure 2.7*(*a*) of a thermoelectric power generator are constructed with p- and n-type materials. The efficiency, limited by Carnot cycle considerations, does not at present exceed 10 %.

A thermocouple generator in which one element is an electron stream or plasma is the *thermionic generator*, in effect a diode with flat cathode and anode very close together. By virtue of their kinetic energy, electrons emitted from the cathode reach the anode against a small, negative anode potential, providing current for an external circuit (*Figure 2.7*(*b*)). The work function of the anode material must be less than for the cathode. The device is a heat engine operating over the cathode–anode temperature fall, with electrons providing the 'working fluid'.

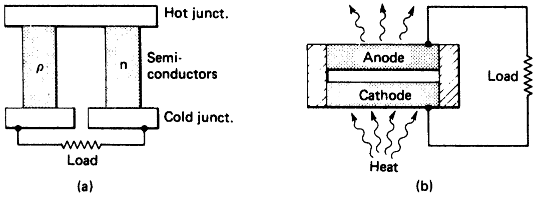


Figure 2.7 Thermoelectric devices

Outputs of 2 kW/m^2 at an efficiency of 25 % may be reached when the device has been fully developed and the space charge effects overcome. Cathode heating by solar energy is a possibility.

2.3 Electrochemical effects

2.3.1 Electrolysis

If a liquid conductor undergoes chemical changes when a current is passed through it, the effect is ascribed to the movement of constituent parts of the molecules of the liquid *electrolyte*, called *ions*, which have a positive or negative electric charge. Positive ions move towards the negative electrode (*cathode*) and negative ions to the positive electrode (*anode*). The ionic movement is the reason for the current conduction. Ions reaching the electrodes have their charges neutralised and may be subject to chemical change. Hydrogen and metal ions are electropositive: non-metals of the chlorine family (Cl, Br, I and F) and acid radicals (such as SO_4 and NO_3) form negative ions in solution. As examples, hydrochloric acid, HCl, forms H^+ and Cl^- ions; sulphuric acid forms $2H^+$ and SO_4^{--} ions; and sodium hydroxide, NaOH, yields Na^+ and OH^- ions. The products of electrolysis depend on the nature of the electrolyte. Basic solutions of sodium or similar hydroxides produce H_2 and O_2 gases at the cathode and anode, respectively. Acid solutions give products depending on the nature of the electrodes. Solutions of metal salts with appropriate electrodes result in electrodeposition.

The mass of the ion of an element of radical deposited on, dissolved from or set free at either electrode is proportional to the quantity of electricity passed through the electrolytic cell and to the ionic weight of the material, and inversely proportional to the valency of the ion; whence the mass m_e in kilogram-equivalents is the product (*z* in kilogram-equivalents per coulomb) × Q in coulombs). The value of *z* is a natural constant 0.001 036. Representative figures (for convenience in milligrams per coulomb) are given in *Table 2.4*.

To pass a current through an electrolyte, a p.d. must be applied to the electrodes to overcome the drop in resistance of the electrolyte, and to overcome the e.m.f. of *polarisation*. The latter is due to a drop across a thin film of gas, or through a strong ionic concentration, at an electrode.

Every chemical reaction may be represented as two electrode reactions. The algebraic p.d. between the two is a measure of the reactivity. A highly negative p.d. represents a spontaneous reaction that might be utilised to generate a current. A high positive p.d. represents a reaction requiring an external applied p.d. to maintain it.

2.3.1.1 *Uses of electrolysis*

Ores of copper, zinc and cadmium may be electrolytically treated with sulphuric acid to deposit the metal. Copper may be deposited by use of a low voltage at the cathode, while oxygen is emitted at the anode. Electrorefining by deposition may be employed with copper, nickel, tin, silver, etc., produced by smelting or electrowinning, by using the impure metal as anode, which is dissolved away and redeposited on the cathode, leaving at the bottom of the cell the impurities in the form of sludge. Electroplating is similar to electrorefining except that pure metal or alloy is used as the anode.

Table 2.4 Electrochemical equivalents z (mg/C)

Element	*Valency*	*z*	*Element*	*Valency*	*z*
H	1	0.010 45	Zn	2	0.338 76
Li	1	0.071 92	As	3	0.258 76
Be	2	0.046 74	Se	4	0.204 56
O	2	0.082 90	Br	1	0.828 15
F	1	0.196 89	Sr	2	0.454 04
Na	1	0.238 31	Pd	4	0.276 42
Mg	2	0.126 01	Ag	1	1.117 93
Al	3	0.093 16	Cd	2	0.582 44
Si	4	0.072 69	Sn	2	0.615 03
S	2	0.166 11	Sn	4	0.307 51
S	4	0.083 06	Sb	3	0.420 59
S	6	0.055 37	Te	4	0.330 60
Cl	1	0.367 43	I	1	1.315 23
K	1	0.405 14	Cs	1	1.377 31
Ca	2	0.207 67	Ba	2	0.711 71
Ti	4	0.124 09	Ce	3	0.484 04
V	5	0.105 60	Ta	5	0.374 88
Cr	3	0.179 65	W	6	0.317 65
Cr	6	0.089 83	Pt	4	0.505 78
Mn	2	0.284 61	Au	1	2.043 52
Fe	1	0.578 65	Au	3	0.681 17
Fe	2	0.289 33	Hg	1	2.078 86
Fe	3	0.192 88	Hg	2	1.039 42
Co	2	0.305 39	Tl	1	2.118 03
Ni	2	0.304 09	Pb	2	1.073 63
Cu	1	0.658 76	Bi	3	0.721 93
Cu	2	0.329 38	Th	4	0.601 35

2.3.2 Cells

2.3.2.1 Primary cells

An elementary cell comprising electrodes of copper (positive) and zinc (negative) in sulphuric acid develops a p.d. between copper and zinc. If a circuit is completed between the electrodes, a current will flow, which acts in the electrolyte to decompose the acid, and causes a production of hydrogen gas round the copper, setting up an e.m.f. of polarisation in opposition to the original cell e.m.f. The latter therefore falls considerably. In practical primary cells the effect is avoided by the use of a *depolariser*. The most widely used primary cell is the Leclanché. It comprises a zinc and a carbon electrode in a solution of ammonium chloride, NH_4Cl. When current flows, zinc chloride, ZnCl, is formed, releasing electrical energy. The NH_4 positive ions travel to the carbon electrode (positive), which is packed in a mixture of manganese dioxide and carbon as depolariser. The NH_4 ions are split up into NH_3 (ammonia gas) and H, which is oxidised by the MnO_2 to become water. The removal of the hydrogen prevents polarisation, provided that the current taken from the cell is small and intermittent.

The *wet* form of Leclanché cell is not portable. The *dry* cell has a paste electrolyte and is suitable for continuous moderate discharge rates. It is exhausted by use or by ageing and drying up of the paste. The 'shelf life' is limited. The *inert* cell is very similar in construction to the dry cell, but is assembled in the dry state, and is activated when required by moistening the active materials. In each case the cell e.m.f. is about 1.5 V.

2.3.2.2 Standard cell

The Weston normal cell has a positive element of mercury, a negative element of cadmium, and an electrolyte of cadmium sulphate with mercurous sulphate as depolariser. The open-circuit e.m.f. at 20°C is about 1.018 30 V, and the e.m.f./temperature coefficient is of the order of – 0.04 mV/°C.

2.3.2.3 Secondary cells

In the *lead–acid* storage cell or accumulator, lead peroxide reacts with sulphuric acid to produce a positive charge at the anode. At the cathode metallic lead reacts with the acid to produce a negative charge. The lead at both electrodes combines with the sulphate ions to produce the poorly soluble lead sulphate. The action is described as

$$\text{Charged}\qquad \underset{\text{Brown}}{\overset{+\Leftarrow}{PbO_2}}\ +\Leftarrow\underset{\text{Strong acid}}{2H_2SO_4}\ +\Leftarrow\underset{\text{Grey}}{\overset{-\Leftarrow}{Pb}}\qquad \text{Discharged}$$

$$=\Leftarrow\underset{\text{Sulphurate}}{\overset{+\Leftarrow}{PbSO_4}}\ +\Leftarrow\underset{\text{Weak acid}}{2H_2O}\ +\Leftarrow\underset{\text{Sulphate}}{\overset{-\Leftarrow}{PbSO_4}}$$

Both electrode reactions are reversible, so that the initial conditions may be restored by means of a 'charging current'.

In the *alkaline* cell, nickel hydrate replaces lead peroxide at the anode, and either iron or cadmium replaces lead at the cathode. The electrolyte is potassium hydroxide. The reactions are complex, but the following gives a general indication:

Charged Discharged

$$2Ni(\overset{+\Leftarrow}{OH})_3 + KOH + \overset{-\Leftarrow}{Fe} = \Leftarrow 2Ni(\overset{+\Leftarrow}{\Leftarrow OH})_2 + KOH + \Leftarrow Fe(\overset{-\Leftarrow}{OH})_2$$
$$\text{or } 2Ni(OH)_3 + KOH + Cd = \Leftarrow 2Ni(OH)_3 + KOH + Cd(OH)_2$$

2.3.2.4 Fuel cell

Whereas a storage battery cell contains all the substances in the electrochemical oxidation–reduction reactions involved and has, therefore, a limited capacity, a fuel cell is supplied with its reactants externally and operates continuously as long as it is supplied with fuel. A practical fuel cell for direct conversion into electrical energy is the hydrogen–oxygen cell (*Figure 2.8*). Microporous electrodes serve to bring the gases into intimate contact with the electrolyte (potassium hydroxide) and to provide the cell terminals. The hydrogen and oxygen reactants are fed continuously into the cell from externally, and electrical energy is available on demand.

At the fuel (H_2) electrode, H_2 molecules split into hydrogen atoms in the presence of a catalyst, and these combine with $OH^{-\Leftarrow}$ions from the electrolyte, forming H_2O and releasing electrons *e*. At the oxygen electrode, the oxygen molecules (O_2) combine (also in the presence of a catalyst) with water molecules from the electrolyte and with pairs of electrons arriving at the electrode through the external load from the fuel electrode. Perhydroxyl ions (O_2H^-) and hydroxyl ions (OH^-) are produced: the latter enter the electrolyte, while the more resistant $O_2H^{-\Leftarrow}$ions, with special catalysts, can be

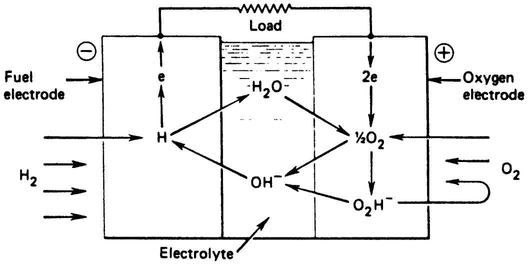


Figure 2.8 Fuel cell

reduced to OH^- ions and oxygen. The overall process can be summarised as:

Fuel electrode	$H_2 + 2OH^- = 2OH_2O + 2e$
Oxygen electrode	$\frac{1}{2}O_2 + H_2O + 2e = 2OH^{-\Leftarrow}$
Net reaction	$H_2 + \frac{1}{2}O_2 \rightarrow 2e \text{ flow} \rightarrow H_2O$

In a complete reaction 2 kg hydrogen and 16 kg oxygen combine chemically (not explosively) to form 18 kg of water with the release of 400 MJ of electrical energy. For each kiloamp/hour the cell produces 0.33 1 of water, which must not be allowed unduly to weaken the electrolyte. The open-circuit e.m.f. is 1.1 V, while the terminal voltage is about 0.9 V, with a delivery of 1 kA/m^2 of plate area.

2.4 Magnetic field effects

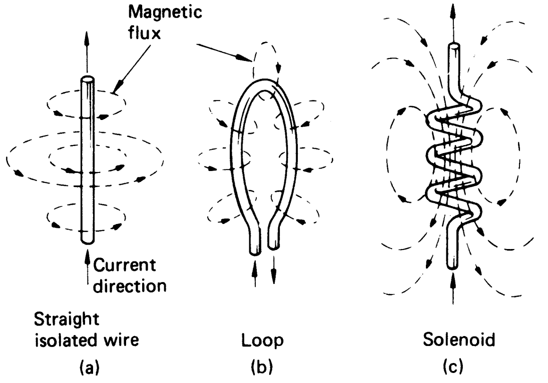
The space surrounding permanent magnets and electric circuits carrying currents attains a peculiar state in which a number of phenomena occur. The state is described by saying that the space is threaded by a *magnetic field of flux*. The field is mapped by an arrangement of *lines of induction* giving the strength and direction of the flux. *Figure 2.9* gives a rough indication of the flux pattern for three simple cases of magnetic field due to a current. The diagrams show the conventions of polarity, direction of flux and direction of current adopted. Magnetic lines of induction form closed loops in a *magnetic circuit* linked by the circuit current wholly or in part.

2.4.1 Magnetic circuit

By analogy with the electric circuit, the magnetic flux produced by a given current in a magnetic circuit is found from the magnetomotive force (m.m.f.) and the circuit reluctance. The m.m.f. produced by a coil of N turns carrying a current I is $F = NI$ ampere-turns. This is expended over any closed path linking the current I. At a given point in a magnetic field in free space the m.m.f. per unit length or magnetising force H gives rise to a magnetic flux density $B_0 = \mu_0 H$, where $\mu_0 = 4\pi/10^7$. If the medium in which the field exists has a relative permeability μ_r, the flux density established is

$$B = \mu_r B_0 = \mu_r \mu_0 H = \mu H$$

The summation of $H{\cdot}\mathrm{d}l$ round any path linking an N-turn circuit carrying current I is the total m.m.f. F. If the distribution of H is known, the magnetic flux density B or B_0 can be found for all points in the field, and a knowledge of the area a of the magnetic path gives $\Phi = Ba$, the total magnetic flux.



Straight isolated wire (a) Loop (b) Solenoid (c)

Figure 2.9 Magnetic fields

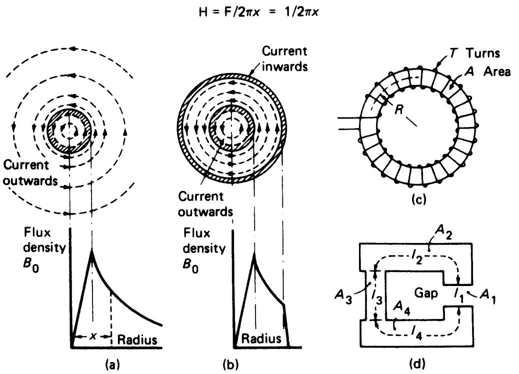


Figure 2.10 Magnetic circuits

Only in a few cases of great geometrical simplicity can the flux due to a given system of currents be found precisely. Among these are the following.

Long straight isolated wire (*Figure 2.10*(*a*)): This is not strictly a realisable case, but the results are useful. Assume a current of 1 A. The m.m.f. around any closed linking path is therefore 1 A-t. Experiment shows that the magnetic field is symmetrical about, and concentric with, the axis of the wire. Around a closed path of radius x metres there will be a uniform distribution of m.m.f. so that

$$H = F/2\pi x = 1/2\pi x(\text{A-t/m})\Leftarrow$$

Consequently, in free space the flux density (T) at radius x is

$$B_0 = \mu_0 H = \mu_0/2\pi x$$

In a medium of constant permeability $\mu = \mu_r\mu_0$ the flux density is $B = \mu_r B_0$. There will be magnetic flux following closed circular paths within the cross-section of the wire itself: at any radius x the m.m.f. is $F = (x/r)^2$ because the circular path links only that part of the (uniformly distributed) current within the path. The magnetising force is $H = F/2\pi x = x/2\pi r^2$ and the corresponding flux density in a non-magnetic conductor is

$$B_0 = \mu_0 H = \mu_0 x/2\pi r^2$$

and μ_r times as much if the conductor material has a relative permeability μ_r. The expressions above are for a conductor current of 1 A.

Concentric conductors (*Figure 2.10*(*b*)): Here only the inner conductor contributes the magnetic flux in the space between the conductors and in itself, because all such flux can link only the inner current. The flux distribution is found exactly as in the previous case, but can now be summed in defined limits. If the outer conductor is sufficiently thin radially, the flux in the interconductor space, per metre axial length of the system, is

$$\Phi = \Leftarrow\int_r^R \frac{\mu_0}{2\pi x}\,\mathrm{d}x = \Leftarrow\frac{\mu_0}{2\pi\psi}\ln\frac{R}{r}$$

Toroid (*Figure 2.10*(*c*)): This represents the closest approach to a perfectly symmetrical magnetic circuit, in which the m.m.f. is distributed evenly round the magnetic path and the m.m.f. per metre H corresponds at all points exactly to the flux density existing at those points. The magnetic flux is therefore wholly confined to the path. Let the mean radius of the toroid be R and its cross-sectional area be A. Then, with N uniformly distributed turns carrying a current I and a toroid core of permeability μ,

$$F = NI; \quad H = F/2\pi R; \quad B = \mu H; \quad \Phi = \mu FA/2\pi R$$

This applies approximately to a long solenoid of length l, replacing R by $l/2\pi$. The permeability will usually be μ_0.

Composite magnetic circuit containing iron (*Figure 2.10(d)*): For simplicity practical composite magnetic circuits are arbitrarily divided into parts along which the flux density is deemed constant. For *each* part

$$F = Hl = Bl/\mu = BlA/\mu A = \Phi S$$

where $S = l/\mu A$ is the *reluctance*. Its reciprocal $\Lambda = 1/S = \mu A/l$ is the *permeance*. The expression $F = \Phi S$ resembles $E = IR$ for a simple d.c. circuit and is therefore sometimes called the *magnetic Ohm's law*.

The total excitation for the magnetic circuit is

$$F = H_1 l_1 + H_2 l_2 + H_3 l_3 + \cdots$$

for a series of parts of length l_1, l_2 ..., along which magnetic field intensities of H_1, H_2... (A-t/m) are necessary. For free space, air and non-magnetic materials, $\mu_r = 1$ and $B_0 = \mu_0 H$, so that $H = B_0/\mu_0 \simeq 800\,000\ B_0$. This means that an excitation $F = 800\,000$ A-t is required to establish unit magnetic flux density (1 T) over a length $l = 1$ m. For ferromagnetic materials it is usual to employ *B–H* graphs (*magnetisation curves*) for the determination of the excitation required, because such materials exhibit a *saturation* phenomenon. Typical *B–H* curves are given in *Figures 2.11* and *2.12*.

2.4.1.1 Permeability

Certain *diamagnetic* materials have a relative permeability slightly less than that of vacuum. Thus, bismuth has $\mu_r = 0.9999$. Other materials have μ_r slightly greater than unity: these are called *paramagnetic*. Iron, nickel, cobalt, steels, Heusler alloy (61% Cu, 27% Mn, 13% Al) and a number of others of great metallurgical interest have *ferromagnetic* properties, in which the flux density is not directly proportional to the magnetising force but which under suitable conditions are strongly magnetic. The more usual constructional materials employed in the magnetic circuits of electrical machinery may have peak permeabilities in the neighbourhood of 5000–10 000. A group of nickel–iron alloys, including *mumetal* (73% Ni, 22% Fe, 5% Cu), *permalloy* 'C' (77.4% Ni, 13.3% Fe, 3.7% Mo, 5% Cu) and *hypernik* (50% Ni, 50% Fe), show much higher permeabilities at low densities (*Figure 2.12*). Permeabilities depend on exact chemical composition, heat treatment, mechanical stress and temperature conditions, as well as on the flux density. Values of μ_r exceeding 5×10^5 can be achieved.

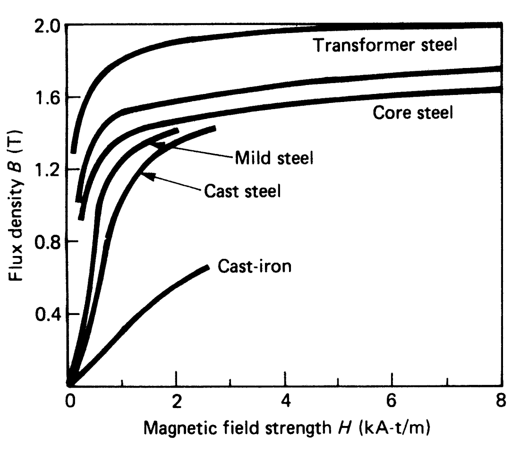


Figure 2.11 Magnetisation curves

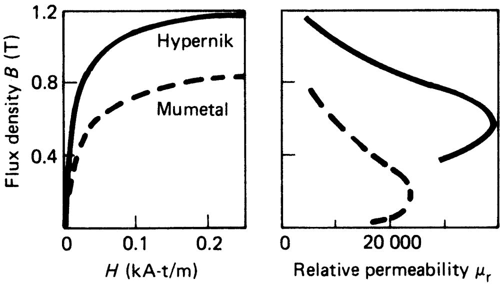


Figure 2.12 Magnetisation and permeability curves

2.4.1.2 Core losses

A ferromagnetic core subjected to cycles of magnetisation, whether alternating (reversing), rotating or pulsating, exhibits *hysteresis*. *Figure 2.13* shows the cycle *B–H* relation typical of this phenomenon. The significant quantities *remanent flux* density and *coercive force* are also shown. The area of the *hysteresis loop* figure is a measure of the energy loss in the cycle per unit volume of material. An empirical expression for the *hysteresis loss* in a core taken through f cycles of magnetisation per second is

$$p_h = k_h f B_m^x \text{ watts per unit mass or volume}$$

Here B_m is the maximum induction reached and k_h is the hysteretic constant depending on the molecular quality and structure of the core metal. The exponent x may lie between 1.5 and 2.3. It is often taken as 2.

A further cause of loss in the same circumstances is the *eddy current loss*, due to the I^2R losses of induced currents. It can be shown to be

$$p_e = k_e t^2 F^2 B^2 \text{ watts per unit mass or volume}$$

the constant k_e depending on the resistivity of the metal and t being its thickness, the material being laminated to decrease the induced e.m.f. per lamina and to increase the resistance of the path in which the eddy currents flow. In practice, curves of loss per kilogram or per cubic metre for various flux densities are employed, the curves being constructed from the results of

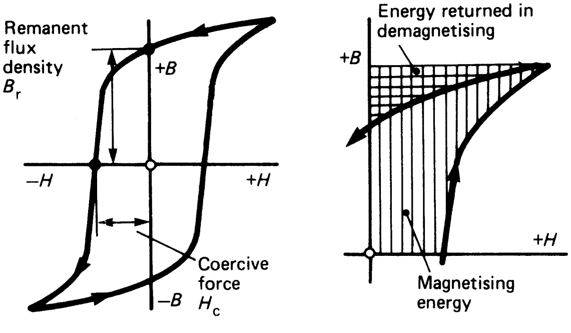


Figure 2.13 Hysteresis

careful tests. It should be noted that hysteresis loss is dependent on the maximum flux density B_m, while the eddy current loss is a function of r.m.s. induced current and e.m.f., and therefore of the r.m.s. flux density B, and not the maximum density B_m.

2.4.1.3 *Permanent magnets*

Permanent magnets are made from heat-treated alloys, or from ferrites and rare earths, to give the material a large hysteresis loop. *Figure 2.14* shows the demagnetisation B/H quadrant of the loop of a typical material. In use, a magnet produces magnetic energy in the remainder of the magnetic circuit derived from a measure of self-demagnetisation: consequently, the working point of the magnet is on the loop between the coercive force/zero flux point and the zero force/remanent flux point. Different parts of the magnet will work at different points on the loop, owing to leakage, and the conditions become much more complex if the reluctance of the external magnetic circuit fluctuates.

In designing a magnet it is necessary to allow for leakage by use of an m.m.f. allowance F_a (normally not more than 1.25) and a flux allowance Φ_a, which may be anything from 2 to 20, being greater for a high ratio between gap length and gap section.

If H_g is the field strength in gap, l_g the gap length, A_g the gap section, B_m the working density in the magnet, H_m the working demagnetising field strength in the magnet, l_m the magnet length and A_m the magnet section, then it may be shown that

$$H_g = \sqrt{[(B_m H_m / F_a \Phi_a)(A_m l_m / A_g l_g)]}$$

i.e. it is greatest when $B_m H_m$ is a maximum. This occurs for a working point at $(BH)_{max}$. The magnet length and section must be proportioned to suit the alloy and the gap dimensions to secure the required condition. The section is $A_m = (H_g A_g \Phi_a / B_m)$ and the length $l_m = (H_g l_g F_a / H_m)$. To calculate these the B_m and H_m values at the $(BH)_{max}$ point must be known. Alternatively, if the three points corresponding to the remanent flux density B_r, the $(BH)_{max}$ and the coercive force H_c are given, the working values can be calculated from

$$B_m = \sqrt{[(BH)_{max}(B_r/H_c)]} \quad \text{and} \quad H_m = \sqrt{[(BH)_{max}(H_c/B_r)]}$$

2.4.2 Magnetomechanical effects

Mechanical forces are developed in magnetic field systems in such a way that the resulting movement increases the flux linkage with the electric circuit, or lowers the m.m.f. required for a given flux. In the former case an increase of linkage requires more energy from the circuit, making mechanical energy available; in the latter, stored magnetic energy is released in mechanical form.

Systems in which magnetomechanical forces are developed are shown diagrammatically in *Figure 2.15*.

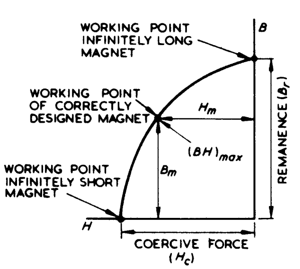
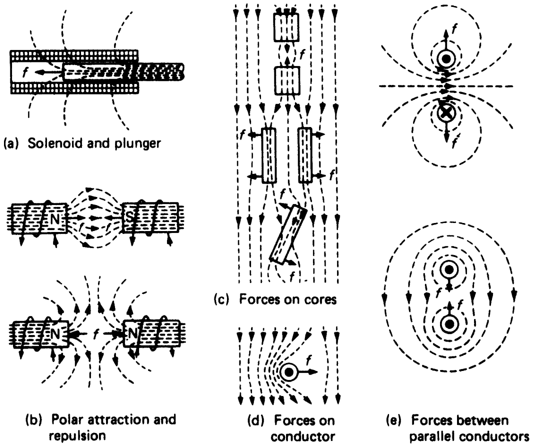

WORKING POINT INFINITELY LONG MAGNET
B
WORKING POINT OF CORRECTLY DESIGNED MAGNET
H_m
$(BH)_{max}$
WORKING POINT INFINITELY SHORT MAGNET
B_m
REMANENCE (B_r)
H
COERCIVE FORCE (H_c)

Figure 2.14 Ideal permanent magnet conditions



(a) Solenoid and plunger

(b) Polar attraction and repulsion

(c) Forces on cores

(d) Forces on conductor

(e) Forces between parallel conductors

Figure 2.15 Magnetomechanical forces

(*a*) *Solenoid and magnetic core*: The core is drawn into the solenoid, increasing the magnetic flux and, in consequence, the circuit's flux linkages.

(*b*) *Attraction or repulsion of magnetised surfaces*: The attraction in (*b*) increases the flux by reducing the reluctance of the intergap. Repulsion gives more space for the opposing fluxes and again reduces the reluctance.

(*c*) *Forces on magnetic cores in a magnetic field*: Cores in line with the field attract each other, cores side by side repel each other, and a core out of line with the general field direction experiences a force tending to align it.

(*d*) *Electromagnetic force on current-carrying conductor*: A current-carrying conductor lying in an externally produced (or 'main') field tends to move so as to increase the flux on that side where its own field has the same direction as the main field.

(*e*) *Electromagnetic force between current-carrying conductors*: Two parallel conductors carrying currents in opposite directions repel, for in moving apart they provide a greater area for the flux between them. If carrying currents in the same direction, they attract, tending to provide a shorter path for the common flux.

It is worth noting as a guide to the behaviour of magnetic field problems (although not a physical explanation of that behaviour) that the forces observed are in directions such as would cause the flux lines to shorten their length and to expand laterally, as if they were stretched elastic threads.

The calculation of the force developed in the cases is based on the movement of the force-system by an amount dl and the amount of mechanical energy dW thereby absorbed or released. Then the force is

$$f = dW/dl$$

With energy in joules and displacement in metres, the force is given in newtons. The calculation is only directly possible in a few simple cases, as below. The references are to the diagrams in *Figure 2.15*.

Case (*c*): The energy stored per cubic metre of a medium in which a magnetising force H produces a density B is $\frac{1}{2}BH = \frac{1}{2}B^2/\mu_r\mu_0$ joules. At an iron surface in air, a movement of the surface into a space originally occupied by air results, for a constant density B, in a reduction of the energy per cubic metre from $B^2/2\mu_0$ to $B^2/2\mu_r\mu_0$, since for air $\mu_r = 1$. The force must therefore be

$$f = \Leftarrow\frac{B^2}{2\mu_0}\left(1 - \frac{1}{\mu_r}\right)\Big]\simeq\Leftarrow\frac{B^2}{2\mu_0} \text{ newtons per square metre}$$

the latter expression being sufficiently close when $\mu_r \gg 1$.

Case (*d*): This case is of particular importance, as it is the basic principle of normal motors and generators, and of moving-coil permanent magnet instruments. Consideration of the mechanical energy gives, for a current I of length l lying perpendicular to the main field, a force

$$f = BlI$$

where B is the flux density. The force, as indicated in *Figure 2.15*(*d*), is at right angles to B and to I. Suppose the conductor to be moved in the direction of the force (either with it or against it): the work done in a displacement of x is

$$fx = W = BlIx = \Phi I \text{ joules}$$

where $\Phi = Blx$ is the total flux cut across by the conductor.

2.4.3 Electromagnetic induction

A magnetic field is a store of energy. When it is increased or decreased, the amount of stored energy increases or decreases. Where the energy is obtained from, or restored to, an associated electric circuit, the energy delivered or received is in the form of a current flowing by reason of an *induced e.m.f.* for a time (specified by the conditions), these three being the essential associated quantities determining electrical energy. The relative directions of e.m.f. and current depend on the direction of energy flow. This is described by *Lenz's law* (*Figure 2.16*), which states that the direction of the e.m.f. induced by a change of linked magnetic field is such as would oppose the change if allowed to produce a current in the associated circuit.

Faraday's law states that the e.m.f. induced in a circuit by the linked magnetic field is proportional to the rate of change of flux linkage   with time. The flux linkage is the summation of products of magnetic flux with the number of turns of the circuit linked by it. Then

$$e = -\mathrm{d}\psi/\mathrm{d}t = -\Sigma N(\mathrm{d}\Phi/\mathrm{d}t)\Leftarrow$$

the negative sign being indicative of the direction of the e.m.f. as specified in Lenz's law.

Consider a circuit of N turns linked completely with a flux Φ. The linkage $= \Phi N$ may change in a variety of ways.

(1) Supposing the flux is constant in value, the circuit may move through the flux (relative motion of flux and circuit: the *motional* or *generator effect*).
(2) Supposing the coil is stationary with reference to the magnetic path of the flux, the latter may vary in magnitude (flux pulsation: the *pulsational* or *transformer effect*).
(3) Both changes may occur simultaneously (movement of coil through varying flux: combination of the effects in (1) and (2)).

The generator effect is associated with conversion of energy between the electrical and mechanical form, using an intermediate magnetic form; the transformer effect concerns the conversion of electrical energy into or from magnetic energy.

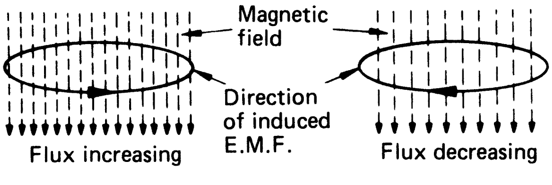


Figure 2.16 Faraday–Lenz law

2.4.3.1 Generator effect

In simplified terms applicable to heteropolar rotating electrical machines (*Figure 2.17*) the instantaneous e.m.f. due to rate of change of linkage resulting from the motion at speed u of an N-turn full-pitch coil of effective length l is $e = 2NBlu$, where B is the flux density in which the coil sides move at the instant considered.

On this expression as a basis, well-known formulae for the motional e.m.f.s of machines can be derived. For example, consider the arrangement (right) in *Figure 2.17*, where the flux density B is considered to be uniform: let the coil rotate at angular velocity ω_r rad/s corresponding to a speed $n = \omega_r/2\pi\psi$rev/s. Then the peripheral speed of the coil is $u = \omega_r R$ if its radius is R. Let the coil occupy a position perpendicular to the flux axis when time $t = 0$. At $t = \theta/\omega_r$ it will be in a position making the angle θ. Its rate of moving across the flux is $\omega_r R \sin\theta$, and the instantaneous coil e.m.f. is

$$e = \omega_r RNBl \sin\theta\psi\!= \omega_r N\Phi_m \sin\theta\psi$$

where $\Phi_m = 2BlR$ is the maximum flux embraced, i.e. at $\theta = 0$. The e.m.f. is thus a sine function of frequency $\omega_r/2\pi\psi$ and the values peak:

$$e_m = \omega_r N\Phi_m \quad \text{r.m.s.;} \quad E = (1/\sqrt{2})\omega_r N\Phi_m$$

The same result is obtained by a direct application of the Faraday law. At $t = 0$ the linked flux is Φ_m; at $t = \theta/\omega_r$ it is $\Phi_m \cos\theta = \Phi_m \cos\omega_r t$. The instantaneous e.m.f. is

$$e = -\mathrm{d}\psi/\mathrm{d}t = -N\Phi_m \mathrm{d}(\cos\omega_r t)/\mathrm{d}t = \omega_r N\Phi_m \sin\theta\psi$$

as before.

2.4.3.2 Transformer effect

The practical case concerns a coil of N turns embracing a varying flux Φ. If the flux changes sinusoidally with time it can be expressed as

$$\Phi = \Phi_m \cos\omega t = \Phi_m \cos 2\pi f t$$

where Φ_m is its time maximum value, f is its frequency, and $\omega = 2\pi f$ is its angular frequency. The instantaneous e.m.f. in the coil is

$$e = -N(d\Phi/\mathrm{d}t) = \omega N\Phi_m \sin\omega t$$

This relation forms the basis of the e.m.f. induced in transformers and induction motors. The e.m.f. and flux relationship is that of *Figures 2.16* and *2.21*(*c*).

2.4.3.3 Calculation of induced e.m.f.

The two methods of calculating electromagnetically induced e.m.f.s are: (1) the change-of-flux law, and (2) the flux-cutting law.

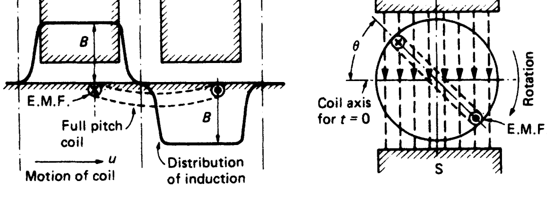


Figure 2.17 Motional e.m.f.

Flux change: This law has the basic form

$$e = -N(\mathrm{d}\Phi/\mathrm{d}t)\Leftarrow$$

and is applicable where a circuit of constant shape links a changing magnetic flux.

Flux cutting: Where a conductor of length l moves at speed u at right angles to a uniform magnetic field of density B, the e.m.f. induced in the conductor is

$$e = Blu$$

This can be applied to the motion of conductors in constant magnetic fields and when sliding contacts are involved.

Linkage change: Where coils move in changing fluxes, and both flux-pulsation and flux-cutting processes occur, the general expression.

$$e = -\mathrm{d}\psi/\mathrm{d}t$$

must be used, with variation of the linkage expressed as the result of both processes.

2.4.3.4 *Constant linkage principle*

The linkage of a *closed* circuit cannot be changed instantaneously, because this would imply an instantaneous change of associated magnetic energy, i.e. the momentary appearance of infinite power. It can be shown that the linkage of a closed circuit of *zero resistance* and no internal source cannot be changed at all. The latter concept is embodied in the following theorem.

Constant linkage theorem The linkage of a closed passive circuit of zero resistance is a constant. External attempts to change the linkage are opposed by induced currents that effectively prevent any net change of linkage.

The theorem is very helpful in dealing with transients in highly inductive circuits such as those of transformers, synchronous generators, etc.

2.4.3.5 *Ideal transformer*

An ideal transformer comprises two resistanceless coils embracing a common magnetic circuit of infinite permeability and zero core loss (*Figure 2.18*). The coils produce no leakage flux: i.e. the whole flux of the magnetic circuit completely links both coils. When the primary coil is energised by an alternating voltage V_1, a corresponding flux of peak value Φ_m is developed, inducing in the N_1-turn primary coil an e.m.f. $E_1 = V_1$. At the same time an e.m.f. E_2 is induced in the N_2-turn secondary coil. If the terminals of the secondary coil are connected to a load taking a current I_2, the primary coil must accept a balancing current I_1 such that $I_1N_1 = I_2N_2$, as the core requires zero excitation. The operating conditions are therefore

$$N_1/N_2 = E_1/E_2 = I_2/I_1; \quad \text{and} \quad E_1I_1 = E_2I_2$$

The secondary load impedance $Z_2 = E_2/I_2$ is reflected into the primary to give the impedance $Z_1 = E_1/I_1$ such that

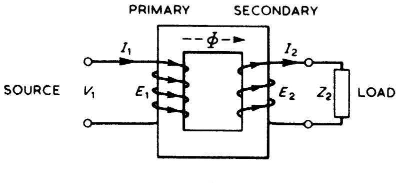


Figure 2.18 Ideal transformer

$$Z_1 = (N_1/N_2)^2Z_2$$

A practical power transformer differs from the ideal in that its core is not infinitely permeable and demands an excitation $N_1I_0 = N_1I_1 - N_2I_2$; the primary and secondary coils have both resistance and magnetic leakage; and core losses occur. By treating these effects separately, a practical transformer may be considered as an ideal transformer connected into an external network to account for the defects.

2.4.3.6 *Electromagnetic machines*

An electromagnetic machine links an electrical energy system to a mechanical one, by providing a reversible means of energy flow between them in the common or 'mutual' magnetic flux linking stator and rotor. Energy is stored in the field and released as work. A current-carrying conductor in the field is subjected to a mechanical force and, in moving, does work and generates a counter e.m.f. Thus the force–motion product is converted to or from the voltage–current product representing electrical power.

The energy-rate balance equations relating the mechanical power p_e, and the energy stored in the magnetic field w_f, are:

Motor: $p_e = p_m = \mathrm{d}w_f/\mathrm{d}t$

Generator: $p_m = p_e + \mathrm{d}w_f/\mathrm{d}t$

The mechanical power term must account for changes in stored kinetic energy, which occur whenever the speed of the machine and its coupled mechanical loads alter.

Reluctance motors The force between magnetised surfaces (*Figure 2.15*(*b*)) can be applied to rotary machines (*Figure 2.19*(*a*)). The armature tends to align itself with the field axis, developing a *reluctance torque*. The principle is applied to miniature rotating-contact d.c. motors and synchronous clock motors.

Machines with armature windings Consider a machine rotating with constant angular velocity ω_r and developing a torque M. The mechanical power is $p_m = M\omega_r$: the electrical power is $p_e = ei$, where e is the counter e.m.f. due to the reaction of the mutual magnetic field. Then $ei = M\omega_r + \mathrm{d}w_f/\mathrm{d}t$ at every instant. If the armature conductor a in *Figure 2.19*(*b*) is running in a non-time-varying flux of local density B, the e.m.f. is entirely rotational and equal to $e_r = Blu = Bl\omega_rR$. The tangential force on the conductor is $f = Bli$ and the torque is $M = BliR$. Thus, $e_ri = M\omega_r$ because $\mathrm{d}w_f/\mathrm{d}t = 0$. This case applies to constant flux (d.c., three-phase synchronous and induction) machines.

If the armature in *Figure 2.19*(*b*) is given two conductors a and b they can be connected to form a turn. Provided the turn is of full pitch, the torques will always be additive. More turns in series form a winding. The total flux in the machine results from the m.m.f.s of all current-carrying conductors, whether on stator or rotor, but the torque arises from that component of the total flux at right angles to the m.m.f. axis of the armature winding.

Armature windings (*Figure 2.19*(*c*)) may be of the commutator or phase (tapped) types. The former is closed on itself, and current is led into and out of the winding by fixed brushes which include between them a constant number of conductors in each armature current path. The armature m.m.f. coincides always with the brush axis. Phase windings have separate external connections. If the winding is on the rotor, its current and m.m.f. rotate with it and the external connections must be made through slip-rings. Two (or three) such windings with two-phase (or three-phase) currents can produce a resultant m.m.f. that rotates with respect to the windings.

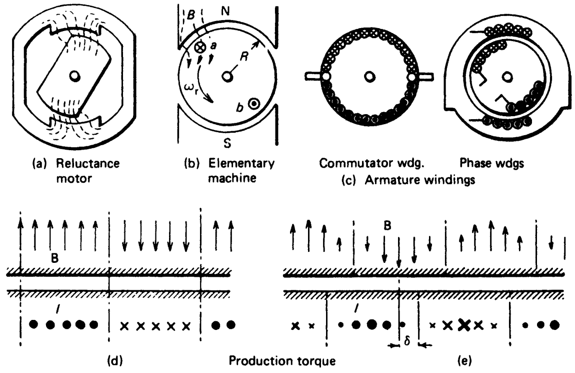


Figure 2.19 Electromagnetic machines

Torque Figure 2.19(d) shows a commutator winding arranged for maximum torque: i.e. the m.m.f. axis of the winding is displaced electrically $\pi/2$ from the field pole centres. If the armature has a radius R and a core length l, the flux has a constant uniform density B, and there are Z conductors in the $2p$ pole pitches each carrying the current I, the torque is $BRlIZ/2p$. This applies to a d.c. machine. It also gives the mean torque of a single-phase commutator machine if B and I are r.m.s. values and the factor cos $\phi\psi$is introduced for any time phase angle between them.

The torque of a phase winding can be derived from *Figure 2.19(e)*. The flux density is assumed to be distributed sinusoidally, and reckoned from the pole centre to be $B_m \cos\alpha$. The current in the phase winding produces the m.m.f. F_a, having an axis displaced by angle $\delta\psi$from the pole centre. The total torque is then

$$M = \pi B_m F_a l R\ \sin\delta\psi = \tfrac{1}{2}\pi\Phi F_a \sin\delta\psi$$

per pole pair. This case applies directly to the three-phase synchronous and induction machines.

Types of machine For unidirectional torque, the axes of the pole centres and armature m.m.f. must remain fixed relative to one another. Maximum torque is obtained if these axes are at right angles. The machine is technically better if the field flux and armature m.m.f. do not fluctuate with time (i.e. they are d.c. values): if they do alternate, it is preferable that they be co-phasal.

Workable machines can be built with (1) concentrated ('field') or (2) phase windings on one member, with (A) commutator or (B) phase windings on the other. It is basically immaterial which function is assigned to stator and which to rotor, but for practical convenience a commutator winding normally rotates. The list of chief types below gives the type of winding (1, 2, A, B) and current supply (d or a), with the stator first:

D.C. machine, 1d/Ad: The arrangement is that of *Figure 2.19(d)*. A commutator and brushes are necessary for the rotor.

Single-phase commutator machine, 1a/Aa: The physical arrangement is the same as that of the d.c. machine. The field flux alternates, so that the rotor m.m.f. must also alternate at the same frequency and preferably in time phase. Series connection of stator and rotor gives this condition.

Synchronous machine, Ba/1d: The rotor carries a concentrated d.c. winding, so the rotor m.m.f. must rotate with it at corresponding (synchronous) speed, requiring a.c. (normally three-phase) supply. The machine may be inverted (1d/Ba).

Induction machine, 2a/Ba (Figure 2.19(e)): The polyphase stator winding produces a rotating field of angular velocity ω_1. The rotor runs with a slip s, i.e. at a speed $\omega_1(1-s)$. The torque is maintained unidirectional by currents induced in the rotor winding at frequency $s\omega_1$. With d.c. supplied to the rotor (2a/Bd) the rotor m.m.f. is fixed relatively to the windings and unidirectional torque is maintained only at synchronous speed ($s=0$).

All electromagnetic machines are variants of the above.

2.4.3.7 Magnetohydrodynamic generator

Magnetohydrodynamics (m.h.d.) concerns the interaction between a conducting fluid in motion and a magnetic field. If a fast-moving gas at high temperature (and therefore ionised) passes across a magnetic field, an electric field is developed across the gaseous stream exactly as if it were a metallic conductor, in accordance with Faraday's law. The electric field gives rise to a p.d. between electrodes flanking the stream, and a current may be made to flow in an external circuit connected to the electrodes. The m.h.d. generator offers a direct conversion between heat and electrical energy.

2.4.3.8 Hall effect

If a flat conductor carrying a current I is placed in a magnetic field of density B in a direction normal to it (*Figure 2.20*), then an electric field is set up across the width of the conductor. This is the Hall effect, the generation of an e.m.f. by the movement of conduction electrons through the magnetic field. The Hall e.m.f. (normally a few microvolts) is picked off by tappings applied to the conductor edges, for the measurement of I or for indication of high-frequency powers.

2.4.4 Inductance

The e.m.f. induced in an electric circuit by change of flux linkage may be the result of changing the circuit's own current. A magnetic field always links a current-carrying circuit, and the linkage is (under certain restrictions) proportional to the current. When the current changes, the linkage also changes and an e.m.f. called the *e.m.f. of self-induction* is induced. If the linkage due to a current i in the circuit is $= \Phi N = Li$, the e.m.f. induced by a change of current is

$$e = -\mathrm{d}\psi/\mathrm{d}t) = -N(\mathrm{d}\Phi/\mathrm{d}t) = -L(\mathrm{d}i/\mathrm{d}t)\Leftarrow$$

L is a coefficient giving the linkage per ampere: it is called the *coefficient of self-induction*, or, more usually, the *inductance*. The unit is the henry, and in consequence of its relation to linkage, induced e.m.f., and stored magnetic energy, it can be defined as follows.

A circuit has unit inductance (1 H) if: (a) the energy stored in the associated magnetic field is $\frac{1}{2}$ J when the current is 1 A; (b) the induced e.m.f. is 1 V when the current is changed at

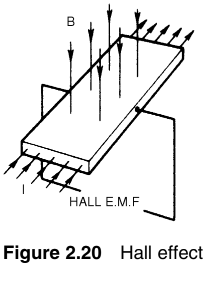


Figure 2.20 Hall effect

the rate 1 A/s; or (c) the flux linkage is 1 Wb-t when the current is 1 A.

2.4.4.1 Voltage applied to an inductor

Let an inductor (i.e. an inductive coil or circuit) devoid of resistance and capacitance be connected to a supply of constant potential difference V, and let the inductance be L. By definition (b) above, a current will be initiated, growing at such a rate that the e.m.f. induced will counterbalance the applied voltage V. The current must rise uniformly at V/L amperes per second, as shown in *Figure 2.21*(*a*), so long as the applied p.d. is maintained. Simultaneously the circuit develops a growing linked flux and stores a growing amount of magnetic energy. After a time t_1 the current reaches $I_1 = (V/L)t_1$, and has absorbed a store of energy at voltage V and average current $I_1/2$, i.e.

$$W_1 = V \cdot \tfrac{1}{2}I_1 \cdot t_1 = \tfrac{1}{2}VI_1t_1 = \tfrac{1}{2}LI_1^2 \text{ joules}$$

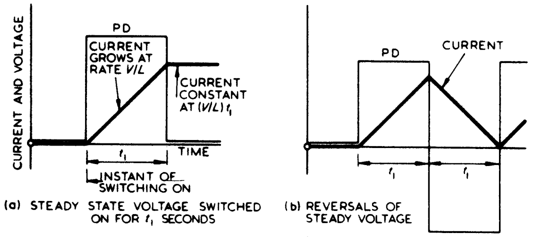
since $V = I_1L/t_1$. If now the supply is removed but the circuit remains closed, there is no way of converting the stored energy, which remains constant. The current therefore continues to circulate indefinitely at value I_1.

Suppose that V is applied for a time t_1, then reversed for an equal time interval, and so on, repeatedly. The resulting current is shown in *Figure 2.21*(*b*). During the first period t_1 the current rises uniformly to $I_1 = (V/L)t_1$ and the stored energy is then $\frac{1}{2}LI_1{}^2$. On reversing the applied voltage the current performs the same rate of change, but negatively so as to reduce the current magnitude. After t_1 it is zero and so is the stored energy, which has all been returned to the supply from which it came.

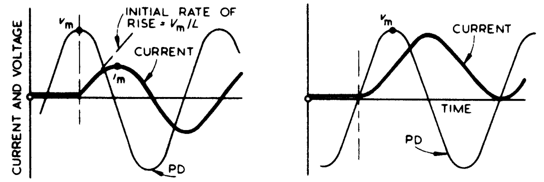
If the applied voltage is sinusoidal and alternates at frequency f, such that $v = v_m \cos 2\pi ft = v_m \cos \omega t$, and is switched on at instant $t = 0$ when $v = v_m$, the current begins to rise at rate v_m/L (*Figure 2.21*(*c*)); but the immediate reduction and subsequent reversal of the applied voltage require corresponding changes in the rate of rise or fall of the current. As $v = L(\mathrm{d}i/\mathrm{d}t)$ at every instant, the current is therefore

$$i = \Leftarrow \int \frac{v}{L}\,\mathrm{d}t = \Leftarrow \frac{v_m}{\omega L}\sin \omega t$$

The peak current reached is $i_m = v_m/\omega L$ and the r.m.s. current is $I = V/\omega L = V/X_L$, where $X_L = \omega L = 2\pi fL$ is the *inductive reactance*.



(a) STEADY STATE VOLTAGE SWITCHED ON FOR t_1 SECONDS (b) REVERSALS OF STEADY VOLTAGE



(c) ALTERNATING VOLTAGE SWITCHED ON AT A PEAK VALUE (d) ALTERNATING VOLTAGE SWITCHED ON AT A ZERO VALUE

Figure 2.21 Voltage applied to a pure inductor

Should the applied voltage be switched on at a voltage zero (*Figure 2.21*(*d*)), the application of the same argument results in a sine-shaped current, unidirectional but pulsating, reaching the peak value $2i_m = 2v_m/\omega L$, or twice that in the symmetrical case above. This is termed the *doubling effect*. Compare with *Figure 2.21*(*b*).

2.4.4.2 Calculation of inductance

To calculate inductance in a given case (a problem capable of reasonably exact solution only in cases of considerable geometrical simplicity), the approach is from the standpoint of definition (c). The calculation involves estimating the magnetic field produced by a current of 1 A, summing the linkage ΦN produced by this field with the circuit, and writing the inductance as $L = \Sigma\Phi N$. The cases illustrated in *Figure 2.10* and *2.22* give the following results.

Long straight isolated conductor (*Figure 2.10*(*a*)) The magnetising force in a circular path concentric with the conductor and of radius x is $H = F/2\pi x = 1/2\pi x$; this gives rise to a circuital flux density $B_0 = \mu_0 H = \mu_0/2\pi x$. Summing the linkage from the radius r of the conductor to a distance s gives

$$= (\mu_0/2\pi)\ln(s/r)\Leftarrow$$

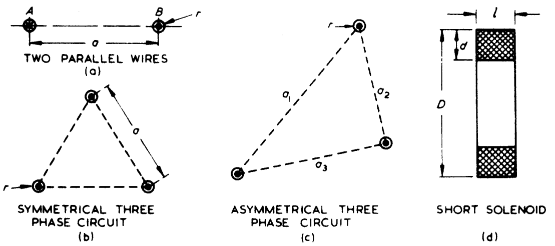
weber-turn per metre of conductor length. If s is infinite, so is the linkage and therefore the inductance: but in practice it is not possible so to isolate the conductor.

There is a magnetic flux following closed circular paths within the conductor, the density being $B_i = \mu x/2\pi r^2$ at radius x. The effective linkage is the product of the flux by that proportion of the conductor actually enclosed, giving $\mu/8\pi\psi$per metre length. It follows that the internal linkage produces a contribution $L_i = \mu/8\pi\psi$henry/metre, regardless of the conductor diameter on the assumption that the current is uniformly distributed. The absolute permeability $\mu\psi$ of the conductor material has a considerable effect on the internal inductance.

Concentric cylindrical conductors (*Figure 2.10*(*b*)) The inductance of a metre length of concentric cable carrying equal currents oppositely directed in the two parts is due to the flux in the space between the central and the tubular conductor set up by the inner current alone, since the current in the outer conductor cannot set up internal flux. Summing the linkages and adding the internal linkage of the inner conductor:

$$L = (\mu/8\pi) + (\mu_0/2\pi)\ln(R/r) \text{ henry/metre}$$

Parallel conductors (*Figure 2.22*) Between two conductors (*a*) carrying the same current in opposite directions, the linkage is found by summing the flux produced by conductor



TWO PARALLEL WIRES (a)

SYMMETRICAL THREE PHASE CIRCUIT (b)

ASYMMETRICAL THREE PHASE CIRCUIT (c)

SHORT SOLENOID (d)

Figure 2.22 Parallel conductors

A in the space a assuming conductor B to be absent, and doubling the result. Provided that $a \gg r$, this gives for the loop

$$L = (\mu/4\pi) + (\mu_0/\pi)\ln(a/r) \text{ henry/metre}$$

It is permissible to regard one-half of the linkage as associated with each conductor, to give for each the *line-to-neutral* inductance

$$L_0 = (\mu/8\pi) + (\mu_0/2\pi)\ln(a/r)\Leftarrow$$

A three-phase line (*b*) has the same line-to-neutral inductance if the conductors are symmetrically spaced. If, however, the spacing is asymmetric but the conductors are cyclically transposed (*c*), the expression applies with $a = \Leftarrow\sqrt[3]{(a_1 a_2 a_3)}$, the geometric mean spacing.

Toroid (*Figure 2.10*(*c*)) For a core of permeability μ the inductance is

$$L = \mu N^2 A/2\pi R = \mu N^2 A/l \text{ henry}$$

where $l = 2\pi R$ is the mean circumference and A is the effective cross-sectional area of the core.

Solenoid A solenoid having a ratio length/diameter of at least 20 has an inductance approximating to that of the toroid above. A short solenoidal coil of overall diameter D, length l, radial thickness d, and N turns has an inductance given approximately by

$$L = \frac{6.4\mu_0 N^2 D^2}{3.5D + 8l} \cdot \Leftarrow \frac{D - 2.25d}{D}$$

For the best ratio of inductance to resistance, $l = d$, giving a square winding cross-section, and $D = 4.7d$. Then $L = (0.8 \times 10^{-6})N^2 D$.

Inductor with ferromagnetic circuit (*Figure 2.10*(*d*)) Saturation makes it necessary to obtain the m.m.f. F for a series of magnetic circuit fluxes Φ. Then with an N-turn exciting coil the inductance is $L = N\Phi/I$, where $I = F/N$. Thus, L is a function of I, decreasing with increase of current. The variation can be mitigated by the inclusion in the magnetic circuit of an air gap to 'stiffen' the flux.

2.4.4.3 *Mutual inductance*

If two coils (primary and secondary) are so oriented that the flux developed by a current in one links the other, the two have mutual inductance. The pair have unit mutual inductance (1 H) if: (a) the energy stored in the common magnetic field is 1 J when the current in each circuit is 1 A; or (b) the e.m.f. induced in one is 1 V when the current in the other changes at the rate 1 A/s, or (c) the secondary linkage is 1 Wb-t when the primary current is 1 A.

Figure 2.23(*a*) shows two coils on a common magnetic circuit: it is assumed that all the flux due to a primary current also links the secondary (a condition approached in a transformer). Let 1 A in the two-turn primary produce 2 Wb. The primary self-inductance is consequently $L_1 = 2 \times 2 = 4$ Wb-t/A = 4 H. The secondary linkage is $4 \times 2 = 8$ Wb-t, so that the mutual inductance is $L_{21} = 8$ H. Let now a current of 1 A circulate instead in the secondary (*b*): it develops double the m.m.f. of that developed by the primary in (*a*) and twice the flux, i.e. 4 Wb. The self-inductance is $L_2 = 4 \times 4 = 16$ H. The primary linkage is $2 \times 4 = 8$, i.e. $L_{12} = 8$ H. Thus, $L_{12} = L_{21}$.

In *Figure 2.23*(*c*) the coils are connected in *series aiding*. With a current of 1 A, the total m.m.f. is $2 + 4 = 6$, the common flux is 6 Wb and the total inductance is $6 \times 6 = 36$ H, which can be shown to be

$$L = L_1 + L_2 + L_{12} + L_{21} = 4 + 16 + 2(8) = 36 \text{ H}$$

For the *series opposing* connection (*d*) the m.m.f.s oppose with a net value $4 - 2 = 2$, and the resulting flux is 2 Wb. The linkages oppose, amounting to $(4 \times 2) - (2 \times 2) = 4$. The total inductance is 4 H, obtained by

$$L = L_1 + L_2 - L_{12} - L_{21} = 4 + 16 - 2(8) = 4 \text{ H}$$

The example shows that $L_{12} = L_{21} = \sqrt{(L_1 L_2)}$. Normally the linkages are less complete, and the ratio $L_{12}/\sqrt{(L_1 L_2)} = k$, the *coefficient of coupling*.

2.4.4.4 *Connection of inductors*

In the absence of mutual inductance, the total inductance of a circuit consisting of inductors L_1, $L_2 \ldots$, is $L = L_1 + L_2 + \cdots$, if they are in series, and $L = 1/[(1/L_1) + (1/L_2) + \cdots]$ if they are in parallel. When there is mutual inductance, it is necessary to set up a circuit equation including the mutual inductance coefficients L_{12}, L_{13}, $L_{23} \ldots$. For two inductors the inductance, as already discussed, is

$$L = L_1 + L_2 \pm 2L_{12}$$

With coils associated on a common ferromagnetic circuit, L_{12} may differ from L_{21} because of saturation effects.

2.5 Electric field effects

When two conductors are separated by a dielectric medium and are maintained at a potential difference, an electric field exists between them. Consider two such conductors A and B (*Figure 2.24*): the application of a p.d. causes a transfer of conduction electrons from A to B, leaving A positive and making B negative, and setting up the electric field

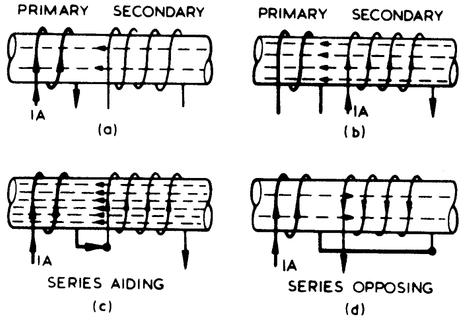
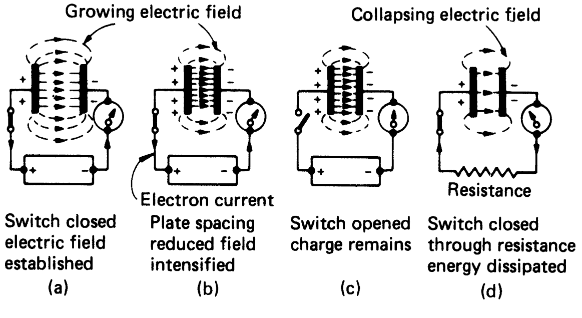


Figure 2.23 Mutual inductance



Switch closed electric field established (a)

Plate spacing reduced field intensified (b)

Switch opened charge remains (c)

Switch closed through resistance energy dissipated (d)

Figure 2.24 Capacitor charge and discharge

(*Figure 2.24*(*a*)). The positive charge on A prevents more than a given number of electrons leaving this conductor, depending on the p.d., the size and configuration, and the spacing. Similarly, the surplus of electrons on B repels others arriving, so that here, too, an equilibrium is established. If, as in (*b*), the spacing is reduced, a further electron transfer takes place until equilibrium is again reached; the charge and the electric field have been intensified by the increase in *capacitance*. If now the switch is opened, the charge, p.d. and field remain as a store of electric field energy (*c*). Let the supply be removed (*d*) and the switch closed through a resistor; the charges on the conductors are dissipated by an electron current from B to A, and the field energy is converted into heat in the resistor.

2.5.1 Electrostatics

The lines used to depict the pattern of an electric field begin on a positive charge and terminate on an equal negative one. Two similar point charges of q_1 and q_2 coulombs, spaced d metres apart in free space (or air) develop a force

$$f = q_1 q_2 / 4\pi\epsilon_0 d^2 \text{ newtons}$$

of repulsion if the charges have the same polarity, of attraction if they have opposite polarity, ϵ_0 being the electric space constant. The force on q_2 can be considered as due to its immersion in the electric field E_1 of q_1, i.e. $f = E_1 q_2$; whence

$$E_1 = q_2 / 4\pi\epsilon_0 d^2 \text{ volts per metre}$$

defining the electric field strength at distance d from a concentrated charge q_1. Thus, a unit charge (1 C) is that which repels a similar charge at unit distance (1 m) with a force of $1/4\pi\epsilon_0$ newton. Similarly, a field of unit strength (1 V/m) produces a mechanical force of 1 N on a unit charge placed in it.

To charge a system like that in *Figure 2.24*, a quantity of electricity has been moved under an applied electric force—i.e. work has been done measured by the charge transferred and the p.d. Unit p.d. (1 V) exists between two points in an electric field when unit work (1 J) is done in moving unit charge (1 C) between them. The two conductors are equipotential surfaces, and potential levels or equipotential lines can be drawn at right angles to the field lines (*Figure 2.25*). Equipotential lines resemble contour lines on the map of a hill: the closer they are, the greater is the voltage gradient. The change of potential in a given direction is

$$V = -\int E \cdot \mathrm{d}x$$

where E is the electric field strength, or potential gradient, in the direction x; whence $E = -\mathrm{d}V/\mathrm{d}x$.

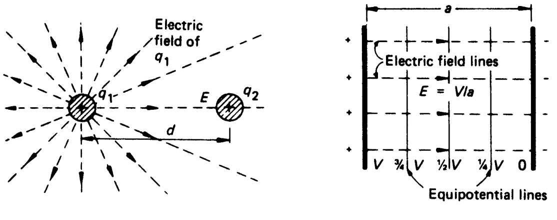


Figure 2.25 Electric fields

2.5.2 Capacitance

The configuration and geometry of the conductor system in *Figure 2.24* depends on the relation between the charge q on each conductor and the p.d., V, between them. Then $q = CV$, where C is the capacitance of the *capacitor* formed by the system. A capacitor has unit capacitance (1 F) if (a) the energy stored in the associated electric field is $\frac{1}{2}$ J for a p.d. of 1 V, or (b) the p.d. is 1 V for a charge of 1 C. Definition (a) follows from the energy storage property: if the p.d. across a capacitor is raised uniformly from zero to V_1, a charge $q_1 = CV_1$ is established at an average p.d. $\frac{1}{2}V_1$, so that the energy input is

$$W = \tfrac{1}{2}V_1 \cdot q_1 = \tfrac{1}{2}V_1 \cdot CV_1 = \tfrac{1}{2}CV_1^2$$

2.5.2.1 Dielectrics

Up to this point it has been assumed that the electric field between the plates of a capacitor has been established through vacuous space. If a material insulator is used—gas, liquid or solid—the electric field will exist therein. It will act on the molecules of the *dielectric* material in accordance with electrostatic principles to 'stretch' or 'rotate' them, and so to orientate the positive and negative molecular charges in opposite directions. This *polarisation* of the dielectric may be imagined to take place as in *Figure 2.26*. Before the p.d. is applied, the molecules of the dielectric material are neutral and unstrained. As the p.d. is raised from zero as in (*a*), the electric field acts to separate the positive and negative elements, the small charge displacement forming a *polarisation* current.

The effect of the application of a p.d. to a capacitor with a material dielectric, then, is to displace a surface charge q_d of polarisation, having a polarity opposite to that of the adjacent capacitor plate q_c. The electric field in the dielectric is due to the resultant or net charge $q = q_c - q_d$. The field strength (and therefore the p.d.) is less than would be expected for the charge q_c on the plates: the relative reduction is found to be approximately constant for a given dielectric. It is called the *relative permittivity*, symbol ϵ_r. The same field strength as for a vacuum will exist in the dielectric for ϵ_r times as much charge on the capacitor plates, so that the capacitor has ϵ_r times the capacitance of a similar capacitor having free space between its plates.

Permittivity effects can thus be taken into account either by assigning to a dielectric a relative permittivity or by considering its polarisation. The latter is of use where internal dielectric forces are concerned, and in dielectric breakdown (*Figure 2.26*(*b*)).

2.5.2.2 Calculation of capacitance

The capacitance of capacitors of simple geometry can be found by assigning, respectively, charges of +1 C and −1 C to the plates or other electrodes, between which the total

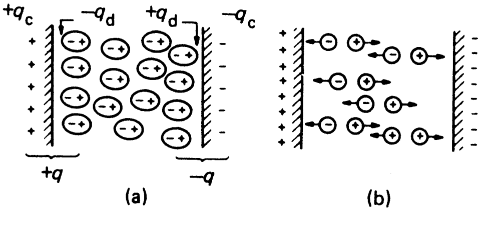


Figure 2.26 Dielectric polarisation and breakdown

electric flux is 1 C. From the field pattern the electric flux density D at any point is found. Then the electric field strength at the point is $E = D/\epsilon$, where $\epsilon = \epsilon_r\epsilon_0$ is the absolute permittivity of the insulating medium in which the electric flux is established. Integration of E over any path from one electrode to the other gives the p.d. V, whence the capacitance is $C = 1/V$.

Parallel plates (*Figure 2.27*(*a*)) The electric flux density is uniform except near the edges. By use of a *guard ring* maintained at the potential of the plate that it surrounds, the capacitance of the inner part is calculable on the reasonable assumption of uniform field conditions. With a charge of 1 C on each plate, and plates of area S spaced a apart, the electric flux density is $D = 1/S$, the electric field intensity is $E = D/\epsilon = 1/S\epsilon$, the potential difference is $V = Ea = a/S\epsilon$, and the capacitance is therefore

$$C = q/V = \epsilon(S/a)\Leftarrow$$

A case of interest is that of a parallel plate arrangement (*Figure 2.27*(*b*)), with two dielectric materials, of thickness a_1 and a_2 and absolute permittivity ϵ_1 and ϵ_2, respectively. The voltage gradient is inversely proportional to the permittivity, so that $E_1\epsilon_1 = E_2\epsilon_2$. The field pattern makes it evident that the difference in polarisation produces an interface charge, but in terms of the charge q_c on the plates themselves the electric flux density is constant throughout. The total voltage between the plates is $V = V_1 + \Leftarrow V_2 = E_1a_1 + E_2a_2$, from which the total capacitance can be obtained.

Concentric cylinders (*Figure 2.27*(*c*)) With a charge of 1 C per metre length, the electric flux density at radius x is $1/2\pi x$, whence $E_x = 1/2\pi x\epsilon$. Integrating for the p.d. gives

$$V = (1/2\pi\epsilon)\ln(R/r)\Leftarrow$$

The capacitance is consequently

$$C = 2\pi\epsilon/\ln(R/r) \text{ farad/metre}$$

The electric field strength (voltage gradient) E is inversely proportional to the radius, over which it is distributed hyperbolically. The maximum gradient occurs at the surface of the inner conductor and amounts to

$$E_m = V/r\ln(R/r)\Leftarrow$$

At any other radius x, $E_x = E_m(r/x)$. For a given p.d. V and gradient E_m there is one value of r to give minimum overall radius R: this is

$$r = V/E_m \text{ and } R = 2.72r$$

For the cylindrical capacitor (*d*) with two dielectrics, of permittivity ϵ_1 between radii r and ρ and ϵ_2 between ρ and R, the maximum gradients are related by $E_{m1}\epsilon_1 r = E_{m2}\epsilon_2\rho$.

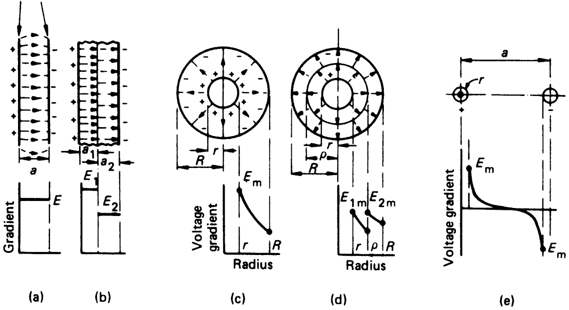


Figure 2.27 Capacitance and voltage gradient

Parallel cylinders (*Figure 2.27*(*e*)) The calculation leads to the value

$$C = \pi\epsilon/\ln(a/r) \text{ farad/metre}$$

for the capacitance between the conductors, provided that $a \gg r$. It can be considered as composed of two series-connected capacitors each of

$$C_0 = 2\pi\epsilon/\ln(a/r) \text{ farad/metre}$$

C_0 being the *line-to-neutral* capacitance. A three-phase line has a line-to-neutral capacitance identical with C_0, the interpretation of the spacing a for transposed asymmetrical lines being the same as for their inductance.

The voltage gradient of a two-wire line is shown in *Figure 2.27*(*e*). If $a \gg r$, the gradient in the immediate vicinity of a wire may be taken as due to the charge thereon, the further wire having little effect: consequently,

$$E_m = V/r\ln(a/r)\Leftarrow$$

is the voltage gradient at a conductor surface.

2.5.2.3 *Connection of capacitors*

If a bank of capacitors of capacitance C_1, C_2, $C_3\ldots$, be connected in *parallel* and raised in combination each to the p.d. V, the total charge is the sum of the individual charges VC_1, VC_2, $VC_3\ldots$, whence the total combined capacitance is

$$C = C_1 + C_2 + C_3 + \cdots\Leftarrow$$

With a *series* connection, the same displacement current occurs in each capacitor and the p.d. V across the series assembly is the sum of the individual p.d.s:

$$\begin{aligned} V &= V_1 + V_2 + V_3 + \cdots\Leftarrow \\ &= q[(1/C_1) + (1/C_2) + (1/C_3) + \cdots]\Leftarrow \\ &= q/C \end{aligned}$$

so that the combined capacitance is obtained from

$$C = 1/[(1/C_1) + (1/C_2) + (1/C_3) + \cdots]\Leftarrow$$

2.5.2.4 *Voltage applied to a capacitor*

The basis for determining the conditions in a circuit containing a capacitor to which a voltage is applied is that the p.d. v across the capacitor is related definitely by its capacitance C to the charge q displaced on its plates: $q = Cv$.

Let a direct voltage V be suddenly applied to a circuit devoid of all characteristic parameters except that of capacitance C. At the instant of its application, the capacitor must accept a charge $q = Cv$, resulting in an infinitely large current flowing for a vanishingly short time. The energy stored is $W = \frac{1}{2}Vq = \frac{1}{2}CV^2$ joules. If the voltage is raised or lowered uniformly, the charge must correspondingly change, by a constant charging or discharging current flowing during the change (*Figure 2.28*(*a*)).

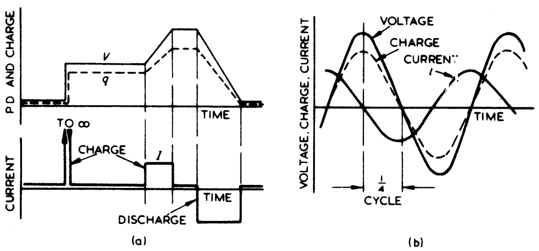


Figure 2.28 Voltage applied to a pure capacitor

If the applied voltage is sinusoidal, as in (*b*), such that $v = v_m \cos 2\pi ft = v_m \cos \omega t$, the same argument leads to the requirement that the charge is $q = q_m \cos \omega t$, where $q_m = Cv_m$. Then the current is $i = dq/dt$, i.e.

$$i = -\omega C v_m \sin \omega t$$

with a peak $i_m = \omega C v_m$ and an r.m.s. value $I = \omega CV = V/X_c$, where $X_c = 1/\omega C$ is the *capacitive reactance*.

2.5.3 Dielectric breakdown

A dielectric material must possess: (a) a high insulation resistivity to avoid leakage conduction, which dissipates the capacitor energy in heat; (b) a permittivity suitable for the purpose—high for capacitors and low for insulation generally; and (c) a high electric strength to withstand large voltage gradients, so that only thin material is required. It is rarely possible to secure optimum properties in one and the same material.

A practical dielectric will break down (i.e. fail to insulate) when the voltage gradient exceeds the value that the material can withstand. The breakdown mechanism is complex.

2.5.3.1 Gases

With gaseous dielectrics (e.g. air and hydrogen), ions are always present, on account of light, heat, sparking, etc. These are set in motion, making additional ionisation, which may be cumulative, causing glow discharge, sparking or arcing unless the field strength is below a critical value. A field strength of the order of 3 MV/m is a limiting value for gases at normal temperature and pressure. The dielectric strength increases with the gas pressure.

The polarisation in gases is small, on account of the comparatively large distances between molecules. Consequently, the relative permittivity is not very different from unity.

2.5.3.2 Liquids

When very pure, liquids may behave like gases. Usually, however, impurities are present. A small proportion of the molecules forms positive or negative ions, and foreign particles in suspension (fibres, dust, water, droplets) are prone to align themselves into semiconducting filaments: heating produces vapour, and gaseous breakdown may be initiated. Water, because of its exceptionally high permittivity, is especially deleterious in liquids such as oil.

2.5.3.3 Solids

Solid dielectrics are rarely homogeneous, and are often hygroscopic. Local space charges may appear, producing absorption effects; filament conducting paths may be present; and local heating (with consequent deterioration) may occur. Breakdown depends on many factors, especially thermal ones, and is a function of the time of application of the p.d.

2.5.3.4 Conduction and absorption

Solid dielectrics in particular, and to some degree liquids also, show conduction and absorption effects. Conduction appears to be mainly ionic in nature. Absorption is an apparent storing of charge *within* the dielectric. When a capacitor is charged, an initial quantity is displaced on its plates due to the *geometric* capacitance. If the p.d. is maintained, the charge gradually grows, owing to *absorptive* capacitance, probably a result of the slow orientation of permanent dipolar molecules. The current finally settles down to a small constant value, owing to conduction.

Absorptive charge leaks out gradually when a capacitor is discharged, a phenomenon observable particularly in cables after a d.c. charge followed by momentary discharge.

2.5.3.5 Grading

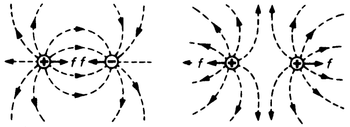
The electric fields set up when high voltages are applied to electrical insulators are accompanied by voltage gradients in various parts thereof. In many cases the gradients are anything but uniform: there is frequently some region where the field is intense, the voltage gradient severe and the dielectric stress high. Such regions may impose a controlling and limiting influence on the insulation design and on the working voltage. The process of securing improved dielectric operating conditions is called *grading*. The chief methods available are:

(1) The avoidance of sharp corners in conductors, near which the gradient is always high.
(2) The application of high-permittivity materials to those parts of the dielectric structure where the stress tends to be high, on the principle that the stress is inversely proportional to the permittivity: it is, of course, necessary to correlate the method with the dielectric strength of the material to be employed.
(3) The use of intersheath conductors maintained at a suitable intermediate potential so as to throw less stress on those parts which would otherwise be subjected to the more intense voltage gradients.

Examples of (1) are commonly observed in high-voltage apparatus working in air, where large rounded conductors are employed and all edges are given a large radius. The application of (2) is restricted by the fact that the choice of materials in any given case is closely circumscribed by the mechanical, chemical and thermal properties necessary. Method (3) is employed in capacitor bushings, in which the intersheaths have potentials adjusted by correlation of their dimensions.

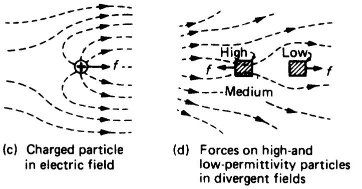
2.5.4 Electromechanical effects

Figure 2.29 summarises the mechanical force effects observable in the electric field. In (*a*), (*b*) and (*c*) are sketched the field patterns for cases already mentioned in connection with the laws of electrostatics. The surface charges developed on high-ε_r materials are instrumental in producing the forces indicated in (*d*). Finally, (*e*) shows the forces on pieces of dielectric material immersed in a gaseous or liquid insulator and subjected to a non-uniform electric field. The force



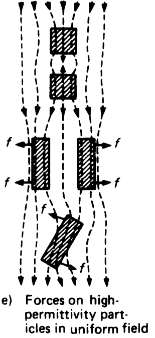
(a) Attraction of opposite charges

(b) Repulsion of like charges



(c) Charged particle in electric field

(d) Forces on high-and low-permittivity particles in divergent fields



(e) Forces on high-permittivity particles in uniform field

Figure 2.29 Electromechanical forces

direction depends upon whether the piece has a higher or lower permittivity than the dielectric medium in which it lies. Thus, pieces of high permittivity are urged towards regions of higher electric field strength.

2.6 Electromagnetic field effects

Electromagnetic field effects occur when electric charges undergo acceleration. The effects may be negligible if the rate of change of velocity is small (e.g. if the operating frequency is low), but other conditions are also significant, and in certain cases effects can be significant even at power frequencies.

2.6.1 Movement of charged particles

Particles of small mass, such as electrons and protons, can be accelerated in vacuum to very high speeds.

Static electric field The force developed on a particle of mass m carrying a positive charge q and lying in an electric field of intensity (or gradient) E is $f = qE$ in the direction of E, i.e. from a high-potential to a low-potential region (*Figure 2.30*(*a*)). (If the charge is negative, the direction of the force is reversed.) The acceleration of the particle is $a = f/m = E(q/m)$; and if it starts from rest its velocity after time t, is $u = at = E(q/m)t$. The kinetic energy $\frac{1}{2}mu^2$ imparted is equal to the change of potential energy Vq, where V is the p.d. between the starting and finishing points in the electric field. Hence, the velocity attained from rest is

$$u = \sqrt{[2V(q/m)]}\Leftarrow$$

For an electron ($q = -1.6 \times 10^{-19}$ C, $m_0 = 0.91 \times 10^{-30}$ kg) falling through a p.d. of 1 V the velocity is 600 km/s and the kinetic energy is $w = Vq = 1.60 \times 10^{-19}$ J, often called an electron-volt, 1 eV.

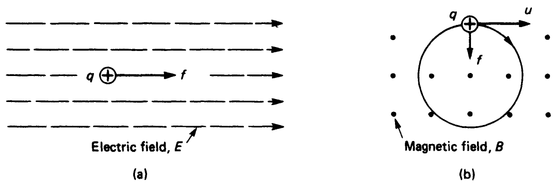
If $V = 2.5$ kV, then $u = 30\,000$ km/s; but the speed cannot be indefinitely raised by increasing V, for as u approaches $c = 300\,000$ km/s, the free-space electromagnetic wave velocity, the effective mass of the particle begins to acquire a rapid relativistic increase to

$$m = m_0/[1 - (u/c)^2]\Leftarrow$$

compared with its 'rest mass' m_0.

Static magnetic field A charge q moving at velocity u is a current $i = qu$, and is therefore subject to a force if it moves across a magnetic field. The force is at right angles to u and to B, the magnetic flux density, and in the simple case of *Figure 2.30*(*b*) we have $f = quB = ma = mu^2/R$, the particle being constrained by the force to move in a circular path of radius $R = (u/B)(m/q)$. For an electron $R = 5.7 \times 10^{-22}$ (u/B).

Combined electric and magnetic fields The two effects described above are superimposed. Thus, if the E and B fields are coaxial, the motion of the particle is helical.



(a) (b)

Figure 2.30 Motion of charged particles

The influence of static (or quasi-static) fields on charged particles is applied in cathode ray oscilloscopes and accelerator machines.

2.6.2 Free space propagation

In Section 1.5.3 the Maxwell equations are applied to propagation of a plane electromagnetic wave in free space. It is shown that basic relations hold between the velocity u of propagation, the electric and magnetic field components E and H, and the electric and magnetic space constants ϵ_0 and μ_0. The relations are:

$$u = 1/\sqrt{(\mu_0\epsilon_0)} \simeq 3 \times 10^8 \text{ metre/second}$$

is the free space propagation velocity. The electric and magnetic properties of space impose a relation between E (in volts per metre) and H (in amps per metre) given by

$$E/H = \sqrt{(\mu_0/\epsilon_0)} = 377\,\Omega.\psi$$

called the *intrinsic impedance* of space. Furthermore, the energy densities of the electric and magnetic components are the same, i.e.

$$\tfrac{1}{2}\epsilon_0 E^2 = \tfrac{1}{2}\mu_0 H^2$$

Propagation in power engineering is not (at present) by space waves but by guided waves, a conducting system being used to direct the electromagnetic energy more effectively in a specified path. The field pattern is modified (although it is still substantially transverse), but the essential physical propagation remains unchanged. Such a guide is called a transmission line, and the fields are normally specified in terms of the inductance and capacitance properties of the line configuration, with an effective impedance $z_0 = \sqrt{(L/C)}$ differing from 377 Ω.

2.6.3 Transmission line propagation (see also Section 36)

If the two wires of a long transmission line, originally dead, are suddenly connected to a supply of p.d. v, an energy wave advances along the line towards the further end at velocity u (*Figure 2.31*). The wave is characterised by the fact that the advance of the voltage charges the line capacitance, for which an advancing current is needed: and the advance of the current establishes a magnetic field against a counter-e.m.f., requiring the voltage for maintaining the advance. Thus, current and voltage are propagated simultaneously. Let losses be neglected, and L and C be the inductance and capacitance per unit length of line. In a brief time interval $\mathrm{d}t$, the waves advance by a distance $u \cdot \mathrm{d}t$. The voltage is established across a capacitance $Cu \cdot \mathrm{d}t$ and the rate of charge, or current, is

$$i = vCu \cdot \mathrm{d}t/\mathrm{d}t = vCu$$

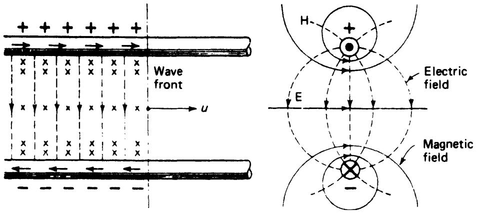


Figure 2.31 Transmission-line field

The current is established in an inductance $Lu \cdot \mathrm{d}t$ producing the magnetic linkages $iLu \cdot \mathrm{d}t$ in time $\mathrm{d}t$, and a corresponding counter-e.m.f. overcome by

$$v = iLu \cdot \mathrm{d}t/\mathrm{d}t = iLu$$

These two expressions, by simple manipulation, yield

Propagation velocity $u = 1/\sqrt{(LC)}$⇐
Surge impedance $z_0 = v/i = \sqrt{(L/C)}$⇐
Energy components $\frac{1}{2}Li^2 = \frac{1}{2}Cv^2$

For a line in air consisting of parallel conductors of radius r spaced a between centres, the inductance (neglecting internal linkage) and capacitance per unit length are

$$L = (\mu_0/\pi)\ln(a/r); \qquad C = \pi\varepsilon_0/\ln(a/r)\Leftarrow$$

whence $u = 1/\sqrt{(\mu_0\epsilon_0)} \simeq 3 \times 10^8$ m/s, exactly as for free space propagation. For lines in which the relative permittivity (and/or, rarely, the relative permeability) of the medium conveying the electromagnetic wave is greater than unity, the speed is reduced by the factor $1/\sqrt{(\mu_r\epsilon_r)}$. Line loss and internal linkage slightly reduce the speed of propagation. With the L and C values quoted, the surge impedance is

$$z_0 = 120\ln(a/r)\Leftarrow$$

which is usually in the range 300–600 Ω. For cables the different geometry and the relative permittivity give a much lower value.

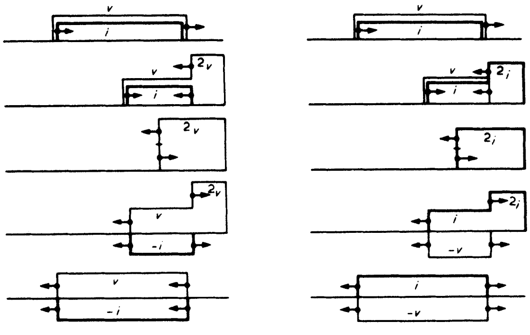
2.6.3.1 Reflection of surges

The relation of p.d. and current direction in a pair of wires forming a long transmission line is determined by the direction in which the energy is travelling (*Figure 2.31*). The current flows in the direction of propagation in the positive conductor, and returns in the negative. Two waves travelling in opposite directions on a line must have *either* the currents *or* the p.d.s in opposite senses. If two such waves meet, either the currents are subtractive and the voltage additive, or the reverse. In each case the natural ratio $v/i = z_0$ for each wave is not apparent: in fact, if the resultant voltage/current ratio is not z_0, the actual distribution of current and voltage must be due to two component waves having opposite directions of propagation. Such conditions arise when a surge is *reflected* at the end of a line.

Consider a steep-fronted surge which reaches the *open-circuited* end of its guiding line. At the point the current must be zero, which requires an equal reflected surge current of opposite sense. The voltages have the same sense and combined to give a doubling effect (*Figure 2.32*(*a*)). Reversal of the energy flow imposed by the line discontinuity, i.e. reflection, is thus accompanied by voltage doubling and an elimination of current. In unit length of a surge, the total energy is $\frac{1}{2}Li^2 + \frac{1}{2}Cv^2$; when two surges (one incident, one reflected) are superimposed, the total energy is electrostatic and of value $\frac{1}{2}C(2v)^2 = 2Cv^2$, which, of course, is equal to the total energy per unit length of the two waves: $2(\frac{1}{2}Cv^2 + \frac{1}{2}Li^2) = Cv^2 + Li^2 = 2Cv^2$.

If the sending-end generator has zero impedance and remains closed on to the line, the returning surge is again reflected at voltage v and the to-and-fro 'oscillation' is maintained indefinitely. In practice, the presence of the losses reduces the reflections and the voltage settles down finally to the steady value v, with a current determined by line admittance.

An incident surge which reaches the *short-circuited* end of a line is presented with a condition of termination across which no p.d. can be maintained. The voltage collapses to zero, initiating a doubling of the current and reflection with reversed voltage (*Figure 2.32*(*b*)).



(a) REFLECTION FROM OPEN END
(b) REFLECTION FROM SHORT-CIRCUITED END

Figure 2.32 Surge reflection

If the sending-end generator has zero impedance and remains closed on to the line, reflections take place repeatedly with an increase of i in the current each time, building up eventually to an infinite value—as would be expected in a lossless line with its end short circuited. Under realisable conditions the current rises with steps of reducing size to a value limited by the series line impedance.

Terminated line If a line is terminated on an impedance z, only partial reflection will take place, some of the incident surge energy being dissipated as heat in the resistive part of z. Let the incident surge be v_1i_1 such that $v_1/i_1 = z_0$; and let the reflected wave be v_2i_2, with $v_2/i_2 = -z_0$. The negative sign is required algebraically to take account of the reversal of either current or voltage by reflection. The voltage v and current i at the termination z during the reflection are such that $v = v_1 + v_2$ and $i = i_1 + i_2$. But $v = iz$, so that, for the reflected current and voltage,

$$v_2 = v_1\frac{z - z_0}{z + z_0} = \alpha v_1 \quad \text{and} \quad i_2 = -i_1\frac{z - z_0}{z + z_0}$$

where α is the *reflection factor*. At the termination itself

$$v = v_1\frac{2z}{z + z_0} = \beta v_1 \quad \text{and} \quad i = i_1\frac{2z_0}{z + z_0}$$

where $\beta = 1 + \alpha$ is the *absorption factor*. The characteristic impedance is resistive, and reflection takes place unless $z = z_0$ and is also resistive. In the latter case v_2 and i_2 vanish, there is no reflection, and the energy in the incident surge v_1i_1 is absorbed in z at the same rate as that at which it arrives.

Line junction or discontinuity A line having an abrupt change of surge impedance from z_0 to z_0'⇐(as, for example, at the junction of a branch line, or at line-cable connection) can be treated as above, substituting z_0'⇐for z. The voltage $v = \beta v_1$ is then characteristic of the *transmitted surge*, and β becomes the transmission factor.

Writing $v = 2v_1 \cdot z/(z + z_0)$, it is seen that v can be considered as the voltage across the load z of a voltage generator of e.m.f. $2v_1$ and internal impedance z_0. Any line termination of reasonable simplicity can now be dealt with. The discontinuity z may be a shunt load, or any combination of loads and extension lines reducible to an equivalent z. If z is resistive, the calculation of the terminal v and i is straightforward; but if z contains a time function (i.e. if it contains inductive and/or capacitive elements), the expressions

$v = \beta v_1$ and $v_2 = \alpha v_1$ become integro-differential equations, to be solved by the methods given in Chapter 5. For z resistive, reflection and transmission (or absorption) of surges takes place without change of shape: for z containing terms in L and C, the shape of the incident surge is modified.

Single reflections Applying the equivalent circuit, a line on open-circuit has $z = \infty$, giving $v = 2v_1$ and $v_2 = v_1$. For a short-circuited line, $z = 0$, $v = 0$ and $v_2 = -v_1$. These cases are shown in *Figure 2.32*.

Shunt inductor termination: Because L offers infinite opposition to infinite rate of current rise, it acts as an open circuit at the instant of surge arrival, degenerating with time constant L/z_0 to a condition of short-circuit. Thus the voltage at the termination is

$$v = 2v_1 \exp(-t \cdot z_0/L) \Leftarrow$$

Shunt capacitor termination: C is initially the equivalent of a short circuit, but charges as the surge continues to arrive, and eventually approaches the charged state, for which it acts as an open circuit. The voltage across it is consequently

$$v = 2v_1[1 - \exp(-t/Cz_0)] \Leftarrow$$

Series inductor termination: If an inductor is inserted into a line of surge impedance z_0, the equivalent circuit is made up of the generator of voltage $2v_1$, loaded with z_0, L and z_0 in series. The voltage transmitted into the line extension is then

$$v' = v_1[1 - \exp(-t \cdot 2z_0/L)] \Leftarrow$$

showing the reduction of wavefront steepness produced. This roughly represents the action of a series surge modifier.

Multiple reflections Surges on power networks will be subject to repeated reflections at terminations, junctions, towers and similar discontinuities. Such cases are handled by developing the appropriate reflection and transmission (or absorption) factors α and β for each discontinuity and each direction. Procedure is facilitated by the use of Bewley's 'lattice diagram': a horizontal scale (of distance along a system in the direction of the initial surge) is combined with a downward vertical scale of time. A lattice of distance–time lines occupies the plane, the lines being sloped to correspond with the velocities of propagation in the system. The lines are marked with their surge voltage values in terms of v_1 and the various transmission and reflection coefficients.

The process is illustrated by the example in *Figure 2.33*. A 500 kV steep-fronted surge reaches the junction P of an overhead transmission line with a cable PQ, 1 km long. At Q the cable is connected to a second overhead line having a short circuit at S, distant 2 km from Q. The overhead lines have a propagation velocity of $u = 0.30$ km/μs and surge impedances z_0 respectively of 500 and 600 Ω; corresponding values for the cable are 0.20 km/μs and 70 Ω. Assuming the surge impedances to be purely resistive and neglecting attenuation, the Bewley lattice diagram is to be drawn for the system and the surge–voltage distribution found for an instant 19 μs after the surge wavefront reaches P.

Five sets of transmission and reflection factors are required, one at S and two each at P and Q for the two directions of propagation. The reflection factor is $\alpha = (z - z_0)/(z + z_0)$ and the transmission (or absorption) factor is $\beta = (1 + \alpha)$. Then

Direction left to right :

At P: $\alpha = (70 - 500)/(70 + 500) = -0.75$; $\beta = +0.25$

At Q: $\alpha = (600 - 70)/(600 + 70) = +0.79$; $\beta = +1.79$

At S: $\alpha = (0 - 600)/(0 + 600) \Leftarrow -1.0$; $\beta = 0$

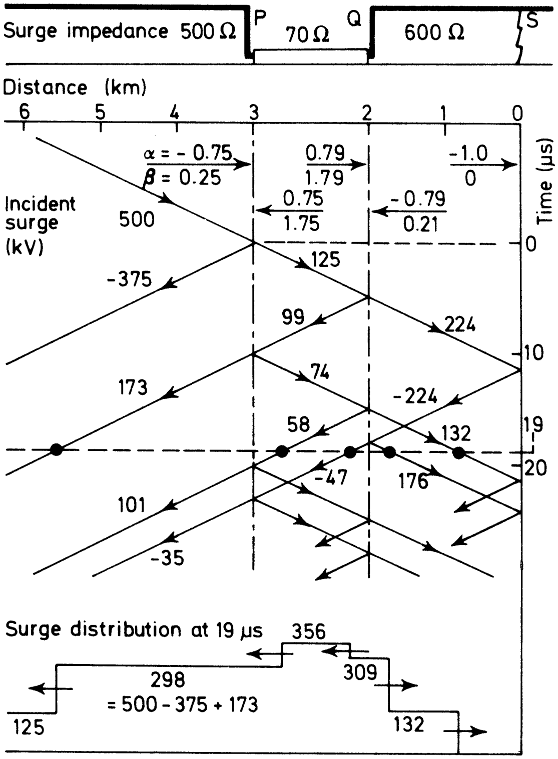


Figure 2.33 Bewley's lattice diagram

Direction right to left :

At P: $\alpha = (500 - 70)/(500 + 70) = +0.75$; $\beta = +1.75$

At Q: $\alpha = (70 - 600)/(70 + 600) = -0.79$; $\beta = +0.21$

The lattice diagram has distance plotted horizontally. The distance scale for the cable section PQ is enlarged by the factor 3/2 to take account of its lower propagation velocity. Time is plotted vertically with time zero at the instant that the surge voltage reaches P. Starting from the upper left-hand side, a sloping straight line is drawn to show the position at any instant of the surge wavefront. Reflection occurs at each junction, so that corresponding lines of the same slope (but running downward to the left) are drawn. The junctions are now marked with the appropriate voltages, using the calculated reflection and transmission factors α and β. Thus the incident surge at P is reflected with $\alpha = -0.75$ to give -375 kV; the transmitted voltage, with $\beta = +0.25$, is 125 kV. The former is turned back at P and is superimposed on the incident surge, while the latter proceeds in the direction PQ, to be split at Q into reflected and transmitted components. At any junction the marked voltages must sum on each side to the same total. The surge–voltage distribution for 19 μs is found by summing the voltages marked on the sloping lines up to this instant.

2.7 Electrical discharges

2.7.1 Introduction

Electric currents may be induced to flow through normally insulating materials by the formation of an electrical discharge. The discharge is maintained by the creation and movement of ions and electrons which in many discharges constitute a plasma (see Section 10.8). Such discharges may be produced between two electrodes which form part of an electrical network and across which a sufficient potential difference exists to ionise the insulation. Alternatively, discharges may be electromagnetically induced, for instance by strong radiofrequency fields.

Electrical discharges may be characterised for electrical network applications in terms of the current and voltage values needed for their occurrence (*Figure 2.34*).

2.7.2 Types of discharge

Discharges have historically been subdivided into two categories namely *self-sustaining* and *non-self-sustaining* discharges. The transition between the two forms (which constitutes the electrical breakdown of the gas) is sudden and occurs through the formation of a *spark*.

Non-self-sustaining discharges occur at relatively low currents ($\sim 10^{-8}$ A) (region 0A, *Figure 2.34*) of which *Townsend discharges* are a particular type. The form of the current–voltage characteristic in this region is governed firstly by a current increase caused by primary electrons ionising the gas by collision to produce secondary electrons, and subsequently by the positive ions formed in this process gaining sufficient energy to produce further ionisation. Such discharges may be induced by irradiating the gas in between two electrodes to produce the initial ionisation. They are non-sustaining because the current flow ceases as soon as the ionising radiation is removed.

When the voltage across the electrodes reaches a critical value V_s (*Figure 2.34*), current level $\sim \cancel{10}^{-5}$ A, the current increases rapidly via a *spark* to form a self-sustaining discharge. The sparking potential V_s for ideal operating conditions (uniform electric field) varies with the product of gas pressure (p) and electrode separation (d) according to Paschen's law (*Figure 2.35*). There is a critical value of pd for which the breakdown voltage V_s is a minimum.

The self-sustaining discharge following breakdown may be either a *glow* or *arc* discharge (regions B–C and D–F, respectively, in *Figure 2.34*) depending on the discharge path and the nature of the connected electric circuit.

The region between V_s and B (*Figure 2.34*) is known as a '*normal*' *glow discharge* and is characterised by the potential difference across the discharge being nearly independent of current, extending to at least 10^{-3} A if not several amperes. For higher currents the voltage increases to form the '*abnormal*' *glow discharge* (region C, *Figure 2.34*).

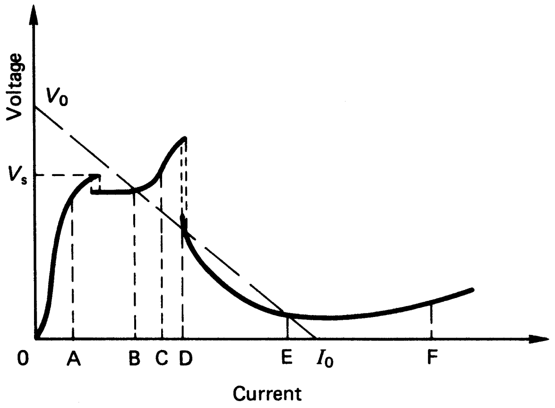


Figure 2.34 Current–voltage characteristic for electrical discharges. 0A, Townsend discharge; B, normal glow; C, abnormal glow; DF, arc; V_s, spark; V_o–I_o, load line

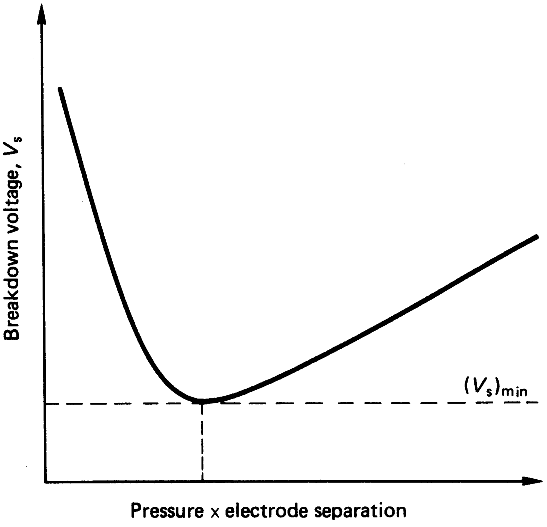


Figure 2.35 Breakdown voltage as a function of the pressure–electrode separation product (Paschen's law)

The glow discharge is manifest as a diffusely luminous plasma extending across the discharge volume but may consist of alternate light and dark regions extending from the cathode in the order: Aston dark space, cathode glow, cathode dark space, negative glow, Faraday dark space, positive glow (which is extensive in volume), the anode glow and the anode dark space (*Figure 2.36*). The glowing regions correspond to ionisation and excitation processes being particularly active and their occurrence and extent depends on the particular operating conditions.

The voltage across the glow discharge consists of two major components: the cathode fall and the positive column (*Figure 2.36a*). Most of the voltage drop occurs across the cathode fall.

For sharply curved surfaces (e.g. wires) and long electrode separations the gas near the surface breaks down at a voltage less than V_s to form a local glow discharge known as a *corona*.

The *electric arc* is a self-sustained discharge requiring only a low voltage for its sustenance and capable of causing currents from typically 10^{-1} A to above 10^5 A to flow.

A major difference between arc and glow discharges is that the current density at the cathode of the arc is greater than that at the glow cathode (*Figure 2.37*). The implication is that the electron emission process for the arc is different from that of the glow and is often thermionic in nature. At higher gas pressures both the anode and cathode of the arc may be at the boiling temperature of the electrode material. Materials having high boiling points (e.g. carbon and tungsten) have lower cathode current densities (ca. $5 \times \cancel{10}^2$ A/cm^2) than materials with lower boiling points (e.g. copper and iron; ca. $5 \times \cancel{10}^3$ A/cm^2).

The arc voltage is the sum of three distinct components (*Figure 2.36*(*b*)): the cathode and anode falls and the positive column. Cathode and anode fall voltages are each typically about 10 V. Short arcs are governed by the electrode fall regions, whereas longer arcs are dominated by the positive column.

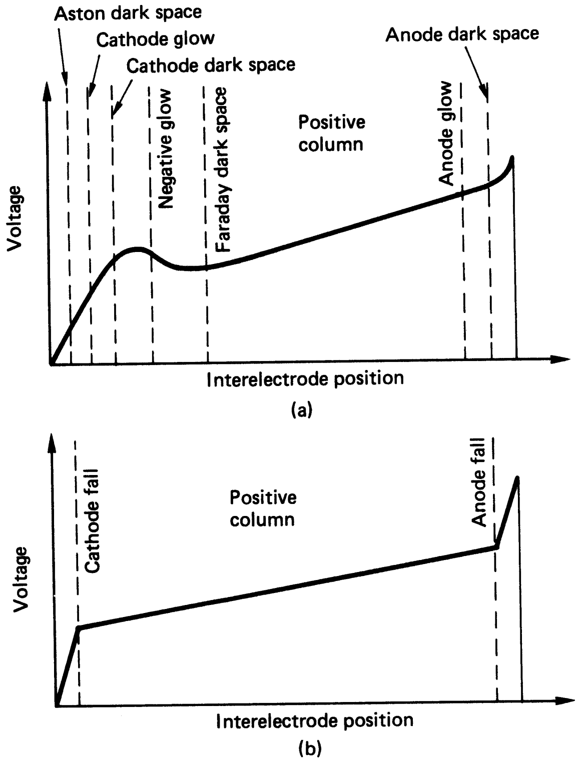


Figure 2.36 Voltage distributions between discharge electrodes: (a) glow; (b) arc

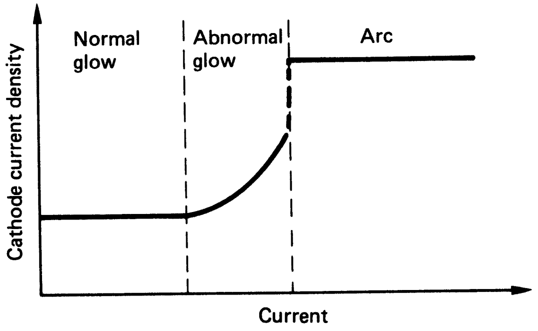


Figure 2.37 Cathode current density distinction between glow and arc discharges

At low pressures the arc may be luminously diffuse and the plasma is not in thermal equilibrium. The temperature of the gas atoms and ions is seldom more than a few hundred degrees Celsius whereas the temperature of the electrons may be as high as 4×10^4 K (*Figure 2.38*).

At atmospheric pressure and above, the arc discharge is manifest as a constricted, highly luminous core surrounded by a more diffusively luminous aureole. The arc plasma column is generally, although not exclusively, in thermal equilibrium.

The arc core is typically at temperatures in the range 5×10^3 to 30×10^3 K so that the gas is completely dissociated and highly ionised. Conversely, the temperature of the aureole spans the range over which dissociation and chemical reactions occur (ca. 2×10^3 to 5×10^3 K).

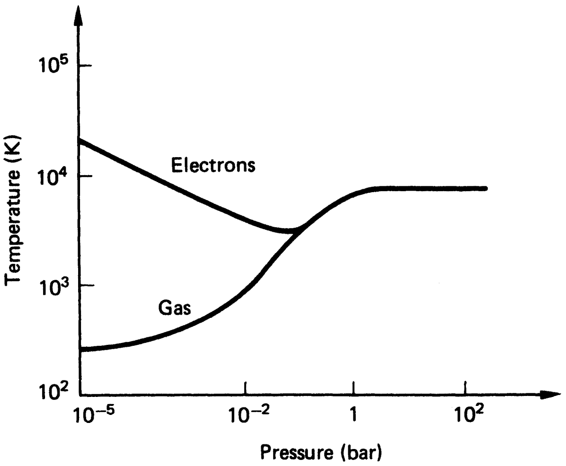


Figure 2.38 Typical electron and gas temperature variations with pressure for arc plasma

The current voltage characteristic of the long electric arc is governed by the electric power (VI) dissipated in the arc column to overcome thermal losses. At lower current levels (10^{-1} to 10^2 A) the are column is governed by thermal conduction losses leading to a negative gradient for the current-voltage characteristic. At higher current levels radiation becomes the dominant loss mechanism yielding a positive gradient characteristic (*Figure 2.34*).

Thermal losses and hence electric power dissipation increase with arc length (e.g. longer electrode separation), gas pressure, convection (e.g. arcs in supersonic flows) and radiation (e.g. entrained metal vapours). As a result the arc voltage at a given current is also increased, causing the discharge characteristic (*Figure 2.34*) to be displaced parallel to the voltage axis.

2.7.3 Discharge–network interaction

For quasi-steady situations the interaction between an electrical discharge and the interconnected network is governed by the intersection of the load line $V_o I_o$ (*Figure 2.34*)

$$V = V_o - iR$$

(where R is the series-network resistance and V_o is the source voltage) and the current–voltage characteristic of the discharge.

The negative gradient of the low current arc branch of the discharge characteristic produces a negative incremental resistance which makes operation at point D (*Figure 2.34*) unstable whereas operation at point E or B is stable.

In practice the operating point is determined by the manner in which the discharge is initiated. Initiation by electrode separation (e.g. circuit breaker contact opening) or by fuse rupture leads to arc operation at E. However, if the discharge is initiated by reducing the series resistance R gradually so that the load line is rotated about V_o, operation as a glow discharge at B may be maintained. If the cathode is heated to provide a large supply of electrons a transition from B to E may occur.

A variation of the source voltage V_o causes the points of intersection of the load line and discharge characteristic to vary so that new points of stability are produced. When the voltage falls below a value, which makes the load line tangential to the negative characteristic, the arc is, in principle, not sustainable. However, in practice, the thermal inertia of the arc plasma may maintain ionisation and so delay eventual arc extinction.

The behaviour of electric arcs in a.c. networks is governed by the competing effects of the thermal inertia of the arc column (due to the thermal capacity of the arc plasma) and the electrical inertia of the network (produced by circuit inductance and manifest as a phase difference between current and voltage). The current–voltage characteristic of the discharge changes from the quasi-steady (d.c.) form of *Figure 2.34* (corresponding to arc inertia being considerably less than the electrical inertia) via an intermediate form (when the thermal and electrical inertias are comparable) to an approximately resistive form (when the thermal inertia is considerably greater than the electrical inertia) (*Figure 2.39*).

2.7.4 Discharge applications

Electrical discharges occur in a number of engineering situations either as limiting or as essential operating features of systems and devices.

Spark discharges are used in applications which utilise their transitional nature. These include spark gaps for protecting equipment against high frequency, high voltage transients and as rapid acting make switches for high power test equipment or pulsed power applications. They are also used for spark erosion in machining materials to high tolerances.

Glow discharges are utilised in a variety of lamps, in gas lasers, in the processing of semiconductor materials and for the surface hardening of materials (e.g. nitriding). Operational problems in all cases involve maintaining the discharge against extinction during the low current part of the driving a.c. at one extreme and preventing transition into an arcing mode (which could lead to destructive thermal overload) at the other extreme.

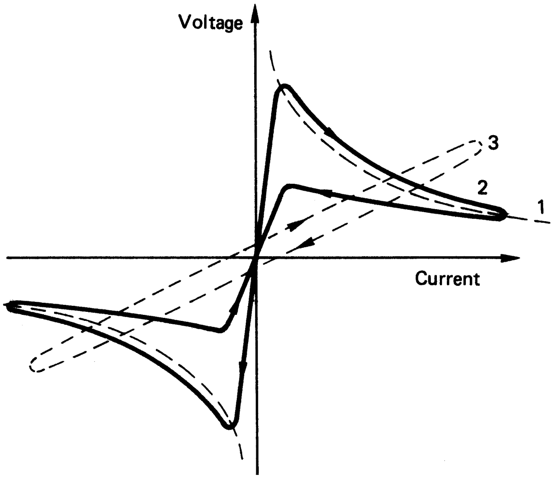


Figure 2.39 Arc current–voltage characteristics for a.c. conditions having different thermal/electrical inertia ratios: (1) thermal ≪ electrical inertia (d.c. case); (2) thermal ≃ electrical inertia; (3) thermal ≫ electrical inertia (e.g. high frequency, resistive case)

Glow discharge lamps either rely on short discharge gaps in which all the light is produced from the negative glow covering the cathode (e.g. neon indicator lamps) or long discharge gaps in which all the light comes from the positive column confined in a long tube (e.g. neon advertising lights).

In materials processing the glow is used to provide the required active ionic species for surface treating the material which forms a cathodic electrode. Both etching of surface layers and deposition of complex layers can be achieved with important applications for the production of integrated circuits for the electronics industry. Metallic surfaces (e.g. titanium steel) may be hardened by nitriding in glow discharges.

Corona on high voltage transmission lines constitute a continuous power loss which for long-distance transmission may be substantial and economically undesirable. Furthermore, such corona can cause a deterioration of insulating materials through the combined action of the discharges (ion bombardment) and the effect of chemical compounds (e.g. ozone and nitrogen oxides) formed in the discharge on the surface.

Arc discharges are used in high pressure lamps, gas lasers, welding, and arc heaters and also occur in current-interruption devices. The distinction between the needs of the two classes of applications is that for lamps and heaters the arc needs to be stably sustained whereas for circuit interruption the arc needs to be extinguished in a controlled manner. The implication is that, when for the former applications an a.c. supply is used, the arc thermal inertia needs to be relatively long compared with the electrical inertia of the network (e.g. to minimise lamp flicker). For circuit-interruption applications the opposite is required in order to accelerate arc extinction and provide efficient current interruption. Such applications require that a number of different current waveforms (*Figure 2.40*) should be interruptable in a controlled manner. High voltage a.c. networks need to be interrupted as the current passes naturally through zero to avoid excessive transient voltages being produced by the inductive nature of such networks. Low voltage, domestic type networks benefit from interruption via the current limiting action of a rapidly lengthening arc. High voltage d.c. network interruption relies upon the arc producing controlled instabilities to force the current artificially to zero.

The arc discharges which are utilised for these applications are configured in a number of different ways. Some basic forms are shown in *Figure 2.41*. These may be divided into two major categories corresponding to the symmetry of the arc. Axisymmetrical arcs include those which are free burning vertically (so that symmetry is maintained by buoyancy forces) wall stabilised arcs, ablation stabilised arcs (which essentially represent arcs in fuses) and axial convection controlled arcs (which are used for both gas heating, welding and high voltage circuit interruption). Non-axisymmetric arcs include the crossflow arc, the linearly driven electromagnetic arc (which has potential for the electromagnetic drive of projectiles or by driving the arc into deionising plates for circuit interruption), the rotary driven electromagnetic arc (which may be configured either between ring electrodes and used for circuit interruption, or helically and used both for gas heating and circuit interruption) and the spiral arc with wall stabilisation.

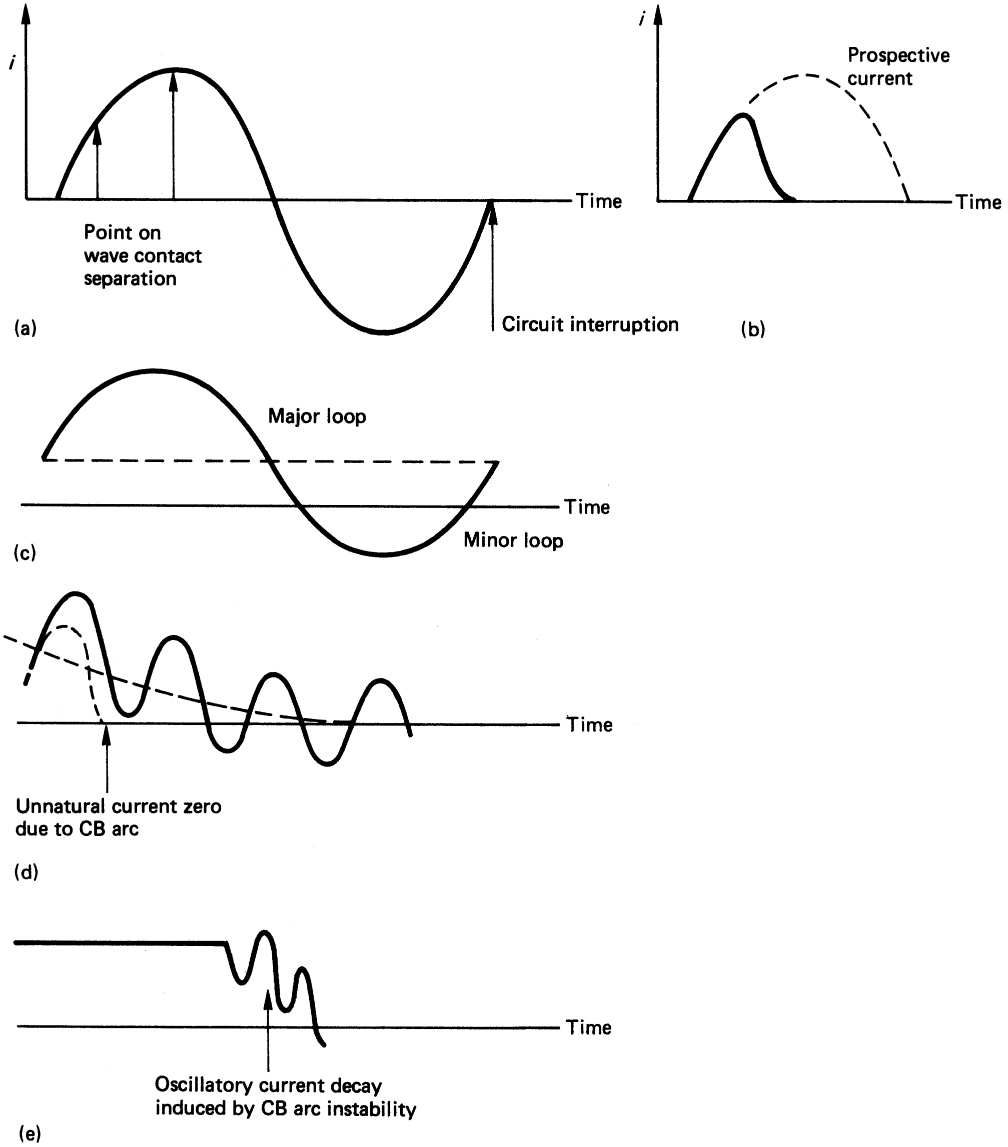

i
Time
Point on
wave contact
separation
(a)
Circuit interruption
i
Prospective
current
Time
(b)
Major loop
Time
Minor loop
(c)
Time
Unnatural current zero
due to CB arc
(d)
Time
Oscillatory current decay
induced by CB arc instability
(e)

Figure 2.40 Circuit breaker (CB) current waveforms: (a) a.c. arc interruption; (b) current limiting effect (domestic CB); (c) asymmetric waveform; (d) generator fault waveform; (e) d.c. arc interruption

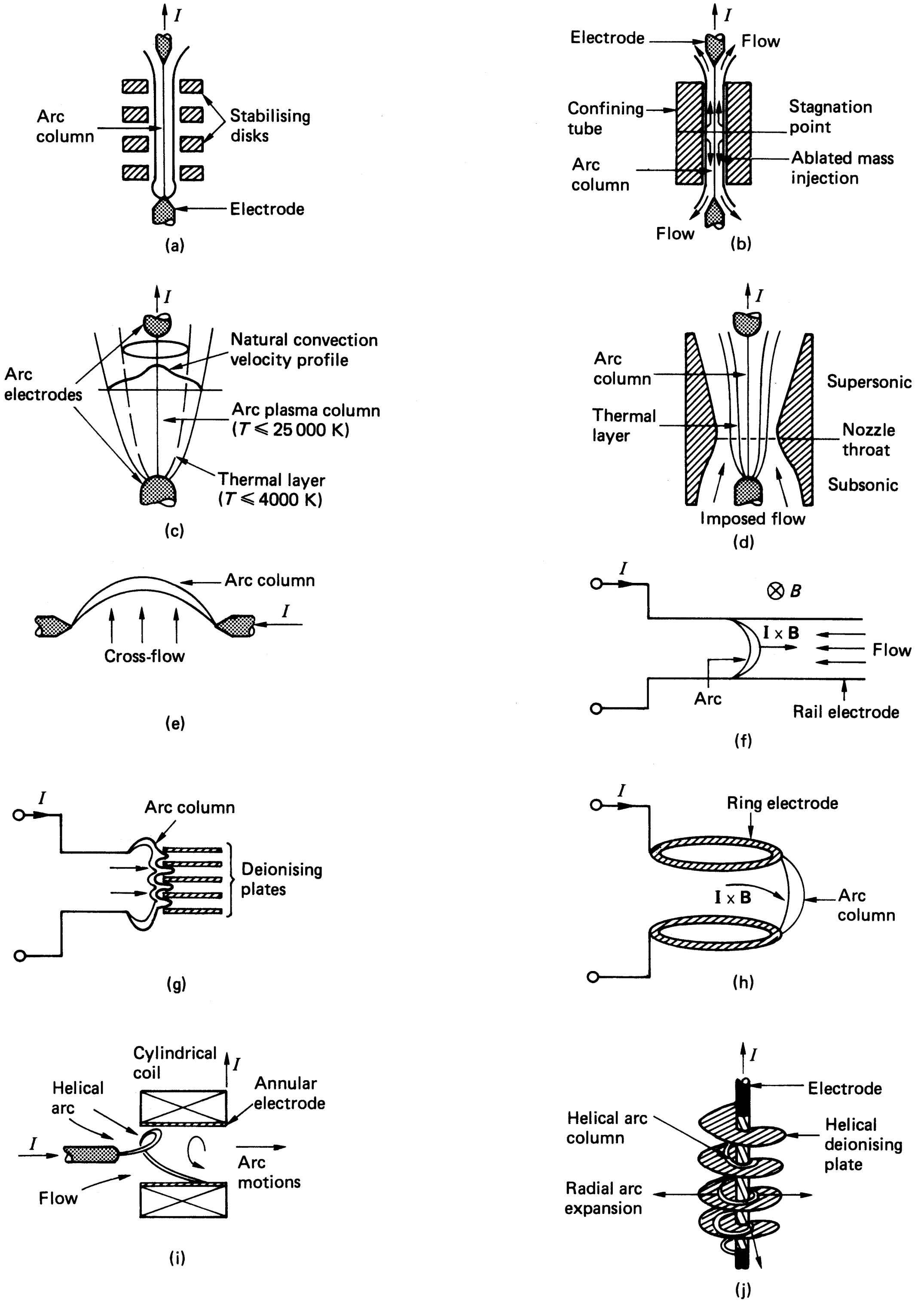

I
Arc column
Stabilising disks
Electrode
(a)
I
Electrode
Flow
Confining tube
Stagnation point
Arc column
Ablated mass injection
Flow
(b)
I
Arc electrodes
Natural convection velocity profile
Arc plasma column ($T \leqslant 25\,000$ K)
Thermal layer ($T \leqslant 4000$ K)
(c)
I
Arc column
Thermal layer
Supersonic
Nozzle throat
Subsonic
Imposed flow
(d)
Arc column
I
Cross-flow
(e)
I
⊗ B
I × B
Flow
Arc
Rail electrode
(f)
I
Arc column
Deionising plates
(g)
I
Ring electrode
I × B
Arc column
(h)
Cylindrical coil
I
Helical arc
Annular electrode
I
Arc motions
Flow
(i)
I
Electrode
Helical arc column
Helical deionising plate
Radial arc expansion
(j)

Figure 2.41 Different arc types: (a) wall stabilised; (b) ablation stabilised; (c) free burning (vertical); (d) axial convection; (e) cross flow; (f) linear electromagnetic acceleration; (g) deionising plates; (h) rotating arc (ring electrodes); (i) helical arc; (j) spiral arc with deionising plate

3

Network Analysis

M G Say PhD, MSc, CEng, FRSE, FIEE, FIERE, ACGI, DIC
Formerly of Heriot-Watt University

M A Laughton BASc, PhD, DSc(Eng), FREng, FIEE
Formerly of Queen Mary & Westfield College,
University of London
(Sections 3.3.1–3.3.5)

Contents

3.1 Introduction 3/3

3.2 Basic network analysis 3/3
  3.2.1 Network elements 3/3
  3.2.2 Network laws 3/4
  3.2.3 Network solution 3/4
  3.2.4 Network theorems 3/5
  3.2.5 Two-ports 3/6
  3.2.6 Network topology 3/7
  3.2.7 Steady-state d.c. networks 3/10
  3.2.8 Steady-state a.c. networks 3/10
  3.2.9 Sinusoidal alternating quantities 3/10
  3.2.10 Non-sinusoidal alternating quantities 3/14
  3.2.11 Three-phase systems 3/15
  3.2.12 Symmetrical components 3/17
  3.2.13 Line transmission 3/18
  3.2.14 Network transients 3/19
  3.2.15 System functions 3/22
  3.2.16 Non-linearity 3/26

3.3 Power-system network analysis 3/28
  3.3.1 Conventions 3/28
  3.3.2 Load-flow analysis 3/29
  3.3.3 Fault-level analysis 3/31
  3.3.4 System-fault analysis 3/31
  3.3.5 Phase co-ordinate analysis 3/34
  3.3.6 Network power limits and stability 3/42

3.1 Introduction

In an electrical network, electrical energy is conveyed from *sources* to an array of interconnected *branches* in which energy is converted, dissipated or stored. Each branch has a characteristic voltage–current relation that defines its *parameters.* The analysis of networks is concerned with the solution of source and branch currents and voltages in a given *network configuration.* Basic and general network concepts are discussed in Section 3.2. Section 3.3 is concerned with the special techniques applied in the analysis of power-system networks.

3.2 Basic network analysis

3.2.1 Network elements

Given the sources (generators, batteries, thermocouples, etc.), the network configuration and its branch parameters, then the network solution proceeds through network equations set up in accordance with the Kirchhoff laws.

3.2.1.1 Sources

In most cases a source can be represented as in *Figure 3.1*(*a*) by an electromotive force (e.m.f.) E_0 acting through an internal series impedance Z_0 and supplying an external 'load' Z with a current I at a terminal voltage V. This is the Helmholtz–Thevenin *equivalent voltage generator*. As regards the load voltage V and current I, the source could equally well be represented by the Helmholtz–Norton *equivalent current generator* in *Figure 3.1*(*b*), comprising a source current I_0 shunted by an internal admittance Y_0 which is effectively in parallel with the load of admittance Y. Comparing the two forms for the same load current I and terminal voltage V in a load of impedance Z or admittance $Y = 1/Z$, we have:

Voltage generator	*Current generator*
$V = E_0 - IZ_0$	$I = I_0 - VY_0$
$I = (E_0 - V)/Z_0$	$V = (I_0 - I)/Y_0$
$= E_0/Z_0 - V/Z_0$	$= I_0/Y_0 - I/Y_0$
$= I_0 - VY_0$	$= E_0 - IZ_0$

These are identical provided that $I_0 = E_0/Z_0$ and $Y_0 = 1/Z_0$. The identity applies only to the *load* terminals, for internally the sources have quite different operating conditions. The two forms are *duals*. Sources with $Z_0 = 0$ and $Y_0 = 0$ (so that $V = E_0$ and $I = I_0$) are termed *ideal* generators.

3.2.1.2 Parameters

When a real *physical* network is set up by interconnecting sources and loads by conducting wires and cables, all parts (including the connections) have associated electric and magnetic fields. A resistor, for example, has resistance as the prime property, but the passage of a current implies a magnetic field, while the potential difference (p.d.) across the resistor implies an electric field, both fields being present in and around the resistor. In the *equivalent* circuit drawn to represent the physical one it is usual to lump together the significant resistances into a limited number of *lumped* resistances. Similarly, electric-field effects are represented by lumped capacitance and magnetic-field effects by lumped inductance. The equivalent circuit then behaves like the physical prototype if it is so constructed as to include all significant effects.

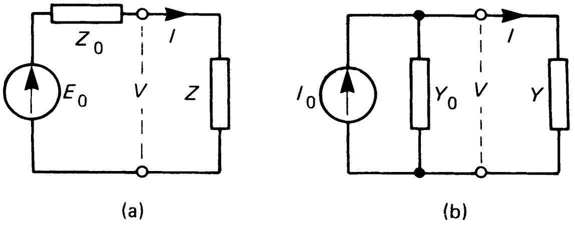


Figure 3.1 (a) Voltage and (b) current sources

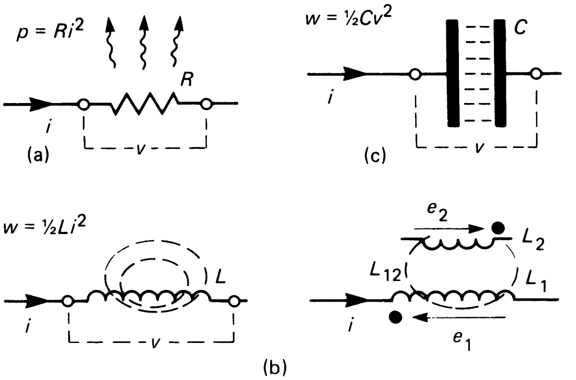


Figure 3.2 Pure parameters

The lumped parameters can now be considered to be free from 'residuals' and *pure* in the sense that simple laws of behaviour apply. These are indicated in *Figure 3.2.*

(*a*) *Resistance* For a pure resistance R carrying an instantaneous current i, the p.d. is $v = Ri$ and the rate of heat production is $p = vi = Ri^2$. Alternatively, if the conductance $G = 1/R$ is used, then $i = Gv$ and $p = vi = Gv^2$. There is a constant relation

$$v = Ri = v/G; \quad i = Gv = v/R; \quad p = Ri^2 = Gv^2$$

(*b*) *Inductance* With a *self*-inductance L, the magnetic linkage is Li, and the source voltage is required only when the linkage changes, i.e. $v = \mathrm{d}(Li)/\mathrm{d}t = L(\mathrm{d}i/\mathrm{d}t)$. An inductor stores in its magnetic field the energy $w = \frac{1}{2}Li^2$. The behaviour equations are

$$v = L(\mathrm{d}i/\mathrm{d}t); \quad i = (1/L)\int v\,\mathrm{d}t; \quad w = \tfrac{1}{2}Li^2$$

Two inductances L_1 and L_2 with a common magnetic field have a *mutual* inductance $L_{12} = L_{21}$ such that an e.m.f. is induced in one when current changes in the other:

$$e_1 = L_{12}(\mathrm{d}i_2/\mathrm{d}t); \quad e_2 = L_{21}(\mathrm{d}i_1/\mathrm{d}t) \Leftarrow$$

The direction of the e.m.f.s depends on the change (increase or decrease) of current and on the 'sense' in which the inductors are wound. The 'dot convention' for establishing the sense is to place a dot at one end of the symbol for L_1, and a dot at that end of L_2 which has the same polarity as the dotted end of L_1 for a given change in the common flux.

(*c*) *Capacitance* The stored charge q is proportional to the p.d. such that $q = Cv$. When v is changed, a charge must enter or leave at the rate $i = \mathrm{d}q/\mathrm{d}t = C(\mathrm{d}v/\mathrm{d}t)$. The electric-field energy in a charged capacitor is $w = \frac{1}{2}Cv^2$. Thus

$$i = C(\mathrm{d}v/\mathrm{d}t); \quad v = (1/C)\int i\,\mathrm{d}t; \quad w = \tfrac{1}{2}Cv^2$$

It can be seen that there is a duality between the inductor and the capacitor. Some typical cases of the behaviour of pure parameters are given in *Figures 2.3*, *2.21* and *2.28*.

A more concise representation of the behaviour of pure parameters uses the differential operator p for d/dt and the inverse 1/p for the integral operator: then

(*a*) Resistance: $v = Ri = v/G$; $i = Gv = (1/R)v$
(*b*) Self-inductance: $v = L\mathrm{p}i$; $i = (1/L\mathrm{p})v$
Mutual inductance: $e_1 = L_{12}\mathrm{p}i_2$; $e_2 = L_{21}\mathrm{p}i_1$
(*c*) Capacitance: $v = (1/C\mathrm{p})i$; $i = C\mathrm{p}v$

For the steady-state direct-current (d.c.) case, p = 0. For steady-state sinusoidal alternating current (a.c.), p = jω, giving for L and C the forms jωL and 1/jωC where ω is the angular frequency. In general, Lp and 1/Cp are the *operational* impedance parameters.

3.2.1.3 Configuration

The assembly of sources and loads forms a network of branches that interconnect nodes (junctions) and form meshes. The seven-branch network shown in *Figure 3.3* has five nodes (a, b, c, d, e) and four meshes (1, 2, 3, 4). Branch ab contains a voltage source; the other branches have (unspecified) impedance parameters. Inspection shows that not all the meshes are independent: mesh 4, for example, contains branches already accounted for by meshes 1, 2 and 3. Further, if one node (say, e) is taken as a *reference* node, the voltages of nodes a, b, c and d can be taken as their p.d.s with respect to node e. The network is then taken as having $b = 7$ branches, $m = 3$ independent meshes and $n = 4$ independent nodes. In general, $m = b - n$.

3.2.2 Network laws

The behaviour of networks (i.e. the branch currents and node voltages for given source conditions) is based on the two Kirchhoff laws (*Figure 3.4*).

(1) *Node law* The total current flowing into a node is zero, $\Sigma i = 0$. The sum of the branch currents flowing into a node must equal the sum of the currents flowing from it; this is a result of the 'particle' nature of conduction current.
(2) *Mesh law* The sum of the voltages around a closed mesh is zero, $\Sigma v = 0$. A rise of potential in sources is absorbed by a fall in potential in the successive branches forming the mesh. This is the result of the nature of a network as an energy system.

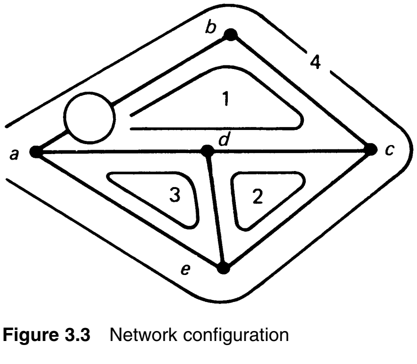


Figure 3.3 Network configuration

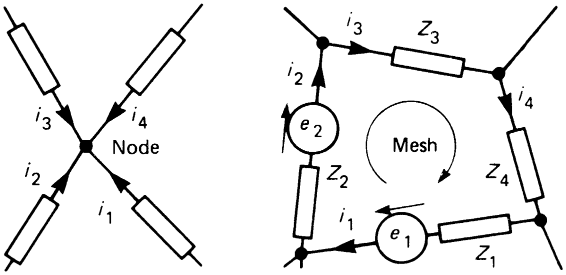


Figure 3.4 The Kirchhoff laws

The Kirchhoff laws apply to all networks. Whether the evaluation of node voltages and mesh currents is tractable or not depends not only on the complexity of the network configuration but also on the branch parameters. These may be active or passive (i.e. containing or not containing sources), linear or non-linear. Non-linearity, in which the parameters are not constant but depend on the voltage and/or current magnitude and polarity, is in fact the normal condition, but where possible the minor non-linearities are ignored in order to permit the use of greatly simplified analysis and the principle of superposition.

3.2.2.1 Superposition

In a strictly linear network, the current in any branch is the sum of the currents due to each source acting *separately*, all other sources being replaced meantime by their internal impedances. The principle applies to voltages and currents, but not to powers, which are current–voltage products.

3.2.3 Network solution

A general solution presents the voltages and currents everywhere in the network; it is initiated by the solution simultaneously of the network equations in terms of voltages, currents and parameters.

The Kirchhoff laws can be applied systematically by use of the *Maxwell circulating-current* process. To each mesh is assigned a circulating current, and the laws are applied with due regard to the fact that certain branches, being common to two adjacent meshes, have net currents given by the superposition of the individual mesh currents postulated. Generalising, the network can be considered as either (i) a set of independent nodes with appropriate node-voltage equations, or (ii) a set of independent meshes with corresponding mesh-current equations.

3.2.3.1 Mesh-current equations

This is a formulation of the Maxwell circulating-current process. If source e.m.f.s are written as E, currents as I and impedances as Z, then for the m independent meshes

$$\begin{aligned}
E_1 &= I_1 Z_{11} + I_2 Z_{12} + \cdots + I_\mathrm{m} Z_{1m}\\
E_2 &= I_1 Z_{21} + I_2 Z_{22} + \cdots + I_\mathrm{m} Z_{2m}\\
&\vdots \qquad\qquad \vdots \qquad\qquad\qquad \vdots\\
E_m &= I_1 Z_{m1} + I_2 Z_{m2} + \cdots + I_m Z_{mm}
\end{aligned}$$

Here Z_{11}, $Z_{22}, \ldots, Z_{mm}$ are the *self-impedances* of meshes 1, 2, …, m, i.e. the total series impedance around each of the chosen meshes; and Z_{12}, $Z_{pq}, \ldots$, are the *mutual impedances* of meshes 1 and 2, p and q, …, i.e. the impedances common to the designated meshes.

The mutual impedance is defined as follows. Z_{pq} is the p.d. per ampere of I_q in the direction of I_p, and Z_{qp} is the p.d. per ampere of I_p in the direction of I_q. The sign of a mutual impedance depends on the current directions chosen for the meshes concerned. If the network is co-planar (i.e. it can be drawn on a diagram with no cross-over) it is usual to select a single consistent direction—say clockwise—for each mesh current. In such a case the mutual impedances are *negative* because the currents are oppositely directed in the common branches.

3.2.3.2 Node-voltage equations

Of the network nodes, one is chosen as a reference node to which all other node voltages are related. The sources are represented by current generators feeding specified currents into their respective nodes and the branches are in terms of admittance Y. Then for the n independent nodes

$$
\begin{aligned}
I_a &= V_a Y_{aa} + V_b Y_{ab} + \cdots + V_n Y_{an} \\
I_b &= V_a Y_{ba} + V_b Y_{bb} + \cdots + V_n Y_{bn} \\
\vdots \quad & \qquad \vdots \qquad\qquad\qquad \vdots \\
I_n &= V_a Y_{na} + V_b Y_{nb} + \cdots + V_n Y_{nn}
\end{aligned}
$$

Here Y_{aa}, $Y_{bb}, \ldots,$ Y_{nn} are the *self-admittances* of nodes a, $b, \ldots,$ n, i.e. the sum of the admittances terminating on nodes $a, b, \ldots, n$; and Y_{ab}, $Y_{pq}, \ldots,$ are the *mutual admittances*, those that link nodes a and b, p and $q, \ldots,$ respectively, and which are usually negative.

The mesh-current and node-voltage methods are general and basic; they are applicable to all network conditions. Simplified and auxiliary techniques are applied in special cases.

3.2.3.3 Techniques

Steady-state conditions Transient phenomena are absent. For d.c. networks the constant current implies absence of inductive effects, and capacitors (having a constant charge) are equivalent to an open circuit. Only resistance is taken into account, using the Ohm law.

For a.c. networks with sinusoidal current and voltage, complexor algebra, phasor diagrams, locus diagrams and symmetrical components are used, while for a.c. networks with periodic but non-sinusoidal waveforms harmonic analysis with superposition of harmonic components is employed.

Transient conditions Operational forms of stimuli and parameters are used and the solutions are found using Laplace transforms.

3.2.4 Network theorems

Network theorems can simplify complicated networks, facilitate the solution of specific network branches and deal with particular network configurations (such as two-ports). They are applicable to linear networks for which *superposition* is valid, and to any form (scalar, complexor, or operational) of voltage, current, impedance and admittance. In the following, the Ohm and Kirchhoff laws, and the reciprocity and compensation theorems, are basic; star–delta transformation and the Millman theorem are applied to network simplification; and the Helmholtz–Thevenin and Helmholtz–Norton theorems deal with specified branches of a network. Two-ports are dealt with in Section 3.2.5.

3.2.4.1 Ohm (Figure 3.2(a))

For a branch of resistance R or conductance G,

$I = V/R = VG; \quad V = I/R = I/G; \quad R = V/I = 1/G$

Summation of resistances R_1, $R_2, \ldots,$ in series or parallel gives

$$
\begin{array}{ll}
\text{Series:} & R = R_1 + R_2 + \cdots \text{ or } G = 1/(1/G_1 + 1/G_2 + \cdots) \\
\text{Parallel:} & R = 1/(1/R_1 + 1/R_2 + \cdots) \text{ or } G = G_1 + G_2 + \cdots
\end{array}
$$

The Ohm law is generalised for a.c. and transient cases by $I = V/Z$ or $I(\mathrm{p}) = V(\mathrm{p})/Z(\mathrm{p})$, where p is the operator d/dt.

3.2.4.2 Kirchhoff (Figure 3.4)

The node and mesh laws are

$$
\begin{array}{ll}
\text{Node:} & i_1 + i_2 + \cdots = \Sigma i = 0 \\
\text{Mesh:} & e_1 + e_2 + \cdots = i_1 Z_1 + i_2 Z_2 + \cdots \text{ or } \quad \Sigma e = \Sigma i Z
\end{array}
$$

3.2.4.3 Reciprocity

If an e.m.f. in branch P of a network produces a current in branch Q, then the same e.m.f. in Q produces the same current in P. The ratio of the e.m.f. to the current is then the *transfer* impedance or admittance.

3.2.4.4 Compensation

For given circuit conditions, any impedance Z in a network that carries a current I can be replaced by a generator of zero internal impedance and of e.m.f. $E = -IZ$. Further, if Z is changed by ΔZ, then the effect on all other branches is that which would be produced by an e.m.f. $-I \Delta Z$ in series with the changed branch. By use of this theorem, if the network currents have been solved for given conditions, the effect of a changed branch impedance can be found without re-solving the entire network.

3.2.4.5 Star–delta (Figure 3.5)

At a given frequency (including zero) a three-branch star impedance network can be replaced by a three-branch delta network, and conversely. For a star Z_a, Z_b, Z_c to be equivalent between terminals AB, BC, CA to a delta Z_1, Z_2, Z_3, it is necessary that

$$
\begin{aligned}
Z_a &= Z_3 Z_1 / Z; \quad & Z_1 &= Z_a + Z_b + Z_a Z_b / Z_c \\
Z_b &= Z_1 Z_2 / Z; \quad & Z_2 &= Z_b + Z_c + Z_b Z_c / Z_a \\
Z_c &= Z_2 Z_3 / Z; \quad & Z_3 &= Z_c + Z_a + Z_c Z_a / Z_b
\end{aligned}
$$

where $Z = Z_1 + Z_2 + Z_3$. The *general* star–mesh conversion concerns the replacement of an n-branch star by a mesh of $\frac{1}{2}n(n-1)$ branches, but *not* conversely; and as the number of mesh branches is greater than the number of star branches when $n > 3$, the conversion is only rarely of use.

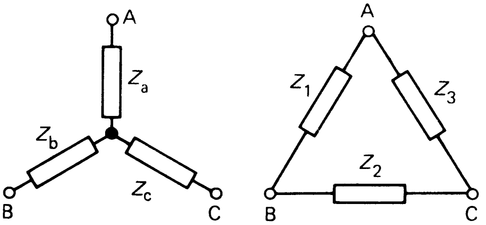


Figure 3.5 Star–delta conversion

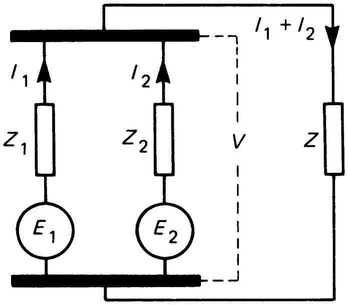


Figure 3.6 The Millman theorem

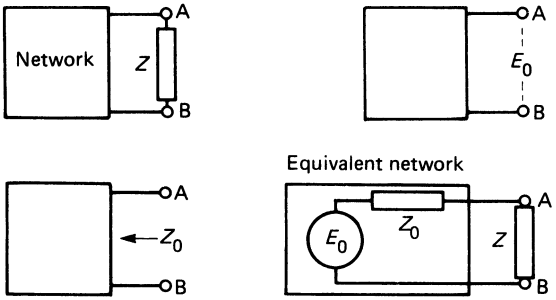


Figure 3.7 The Helmholtz–Thevenin theorem

3.2.4.6 Millman (*Figure 3.6*)

The Millman theorem is also known as the *parallel-generator* theorem. The common terminal voltage of a number of sources connected in parallel to a common load of impedance Z is $V = I_{sc}Z_p$, where I_{sc} is the sum of the short-circuit currents of the individual source branches and Z_p is the effective impedance of all the branches in parallel, including the load Z. If E_1 and E_2 are the e.m.f.s of two sources with internal impedances Z_1 and Z_2 connected in parallel to supply a load Z, and if I_1 and I_2 are the currents contributed by these sources to the load Z, then their common terminal voltage V must be

$$V = (I_1 + I_2)Z = [(E_1 - V)/Z_1 + (E_2 - V)/Z_2]Z$$

whence

$$V(1/Z + 1/Z_1 + 1/Z_2) = E_1/Z_1 + E_2/Z_2$$

The term in parentheses on the left-hand side of the equation is the effective admittance of all the branches in parallel. The right-hand side of the equation is the sum of the individual source short-circuit currents, totalling I_{sc}. Thus $V = I_{sc}Z_p$. The theorem holds for any number of sources.

3.2.4.7 Helmholtz–Thevenin (*Figure 3.7*)

The current in any branch Z of a network is the same as if that branch were connected to a voltage source of e.m.f. E_0 and internal impedance Z_0, where E_0 is the p.d. appearing across the branch terminals when they are open-circuited and Z_0 is the impedance of the network looking into the branch terminals with all sources represented by their internal impedance.

In *Figure 3.7*, the network has a branch AB in which it is required to find the current. The branch impedance Z is removed, and a p.d. E_0 appears across AB. With all sources replaced by their internal impedance, the network presents the impedance Z_0 to AB. The current in Z when it is replaced into the original network is

$$I = E_0/(Z_0 + Z) \Leftarrow$$

The whole network apart from the branch AB has been replaced by an equivalent *voltage* source, resulting in the simplified condition of *Figure 3.1*(*a*).

3.2.4.8 Helmholtz–Norton

The Helmholtz–Norton theorem is the dual of the Helmholtz–Thevenin theorem. The voltage across any branch Y of a network is the same as if that branch were connected to a current source I_0 with internal shunt admittance Y_0, where I_0 is the current between the branch terminals when short circuited and Y_0 is the admittance of the network looking into the branch terminals with all sources represented by their internal admittance. Then across the terminals AB in *Figure 3.7* the voltage is

$$V = I_0/(Y_0 + Y) \Leftarrow$$

Thus the whole network apart from the branch AB has been replaced by an equivalent *current* source, i.e. the system in *Figure 3.1*(*b*).

3.2.5 Two-ports

In power and signal transmission, input voltage and current at one port (i.e. one terminal-pair) yield voltage and current at another port of the interconnecting network. Thus in *Figure 3.8a* voltage source at the input port 1 delivers to the passive network a voltage V_1 and a current I_1. The corresponding values at the output port 2 are V_2 and I_2.

3.2.5.1 Lacour

According to the theorem originated by Lacour, any passive linear network between two ports can be replaced by a two-mesh or T network, and in general no simpler form can be found. Such a result is obtained by iterative star–delta conversion to give the T equivalent; by one more star–delta conversion the Π-equivalent is obtained (*Figure 3.9*). In general, the equivalent networks are asymmetric; in some cases, however, they are symmetric. It can be shown that a passive two-port has the input and output voltages and currents related by

$$V_1 = \boldsymbol{A}V_2 + \boldsymbol{B}I_2 \quad \text{and} \quad I_1 = \boldsymbol{C}V_2 + \boldsymbol{D}I_2$$

where $\boldsymbol{ABCD}$ are the general two-port parameters, constants for a given frequency and with $\boldsymbol{AD} - \boldsymbol{BC} = 1$. The conventions for voltage polarity and current direction are those given in *Figure 3.8*.

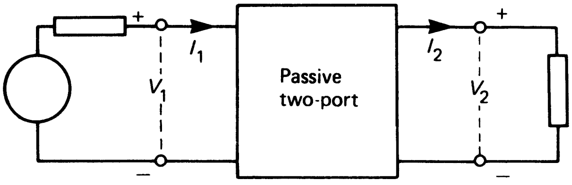


Figure 3.8 Two-port network

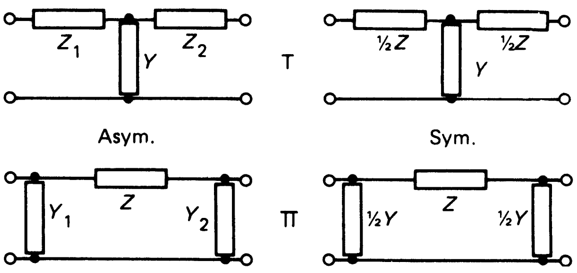


Figure 3.9 T and Π two-ports

3.2.5.2 T *network*

Consider the asymmetric T in *Figure 3.9*. Application of the Kirchhoff laws gives

$$I_1 = I_2 + (V_2 + I_2Z_2)Y = V_2Y + I_2(1 + YZ_2)$$
$$V_1 = V_2(1 + YZ_1) + I_2(Z_1 + Z_2 + Z_1Z_2Y)$$

Hence in terms of the series and parallel branch components

$$\boldsymbol{A} = 1 + YZ_1$$
$$\boldsymbol{B} = Z_1 + Z_2 + Z_1Z_2Y$$
$$\boldsymbol{C} = Y$$
$$\boldsymbol{D} = 1 + YZ_2$$

Multiplication shows that $\boldsymbol{AD} - \boldsymbol{BC} = 1$.

For the symmetric T with $Z_1 = Z_2 = \frac{1}{2}Z$,

$$\boldsymbol{A} = 1 + \tfrac{1}{2}YZ = \boldsymbol{D}; \quad \boldsymbol{B} = Z + \tfrac{1}{4}YZ^2; \quad \boldsymbol{C} = Y$$

3.2.5.3 Π *network*

In a similar way, the general parameters for the asymmetric case are

$$\boldsymbol{A} = 1 + Y_2Z; \quad \boldsymbol{B} = Z; \quad \boldsymbol{C} = Y_1 + Y_2 + Y_1Y_2Z; \quad \boldsymbol{D} = 1 + Y_1Z$$

which reduce with symmetry to

$$\boldsymbol{A} = 1 + \tfrac{1}{2}YZ = \boldsymbol{D}; \quad \boldsymbol{B} = Z; \quad \boldsymbol{C} = Y + \tfrac{1}{4}Y^2Z$$

The values of the ***ABCD*** parameters, in matrix form,

$$\begin{pmatrix} \boldsymbol{A} & \boldsymbol{B} \\ \boldsymbol{C} & \boldsymbol{D} \end{pmatrix}$$

are set out in *Table 3.1* for a number of common cases.

3.2.5.4 *Characteristic impedance*

If the output terminals of a two-port are closed through an impedance $V_2/I_2 = Z_0$, and if the input impedance V_1/I_1 is then also Z_0, the quantity Z_0 is the *characteristic impedance*. Consider a symmetrical two-port ($\boldsymbol{A} = \boldsymbol{D}$) so terminated: if V_1/I_1 is to be Z_0 we have

$$\frac{V_1}{I_1} = \frac{V_2\boldsymbol{A} + I_2\boldsymbol{B}}{V_2\boldsymbol{C} + I_2\boldsymbol{A}} = \frac{V_2(\boldsymbol{A} + \boldsymbol{B}/Z_0)}{I_2(\boldsymbol{A} + \boldsymbol{C}Z_0)} = Z_0\frac{\boldsymbol{A} + \boldsymbol{B}/Z_0}{\boldsymbol{A} + \boldsymbol{C}Z_0}$$

which is Z_0 for $\boldsymbol{B}/Z_0 = \boldsymbol{C}Z_0$. Thus the characteristic impedance is $Z_0 = \sqrt{(\boldsymbol{B}/\boldsymbol{C})}$. The same result is obtainable from the input impedances with the output terminals first open circuited ($I_2 = 0$) giving Z_{oc}, then short circuited ($V_2 = 0$) giving Z_{sc}: thus

$$Z_{oc} = \boldsymbol{A}/\boldsymbol{C}; \quad Z_{sc} = \boldsymbol{B}/\boldsymbol{A}; \quad Z_0 = \sqrt{(Z_{oc}Z_{sc}} = \sqrt{(\boldsymbol{B}/\boldsymbol{C})}$$

3.2.5.5 *Propagation coefficient*

The parameters ***ABCD*** are functions of frequency, and Z_0 is a complex operator. For the Z_0 termination of a symmetrical two-port (for which $\boldsymbol{A}^2 - \boldsymbol{BC} = 1$) the input/output voltage or current ratio is

$$V_1/V_2 = I_1/I_2 = \boldsymbol{A} + \sqrt{(\boldsymbol{BC})} = \boldsymbol{A} + \sqrt{(\boldsymbol{A}^2 - 1)}$$
$$= \exp(\gamma) = \exp(\alpha + j\beta)$$

The magnitude of V_1 exceeds that of V_2 by the factor $\exp(\alpha)$ and leads it by the angle β, where α is the *attenuation* coefficient, β is the *phase* coefficient and the combination $\gamma = \alpha + j\beta$ is the *propagation* coefficient.

3.2.5.6 *Alternative two-port parameters*

There are other ways of expressing two-port relationships. For generality, both terminal voltages are taken as *applied* and both currents are *input* currents. With this convention it is necessary to write $-I_2$ for I_2 in the general parameters so far discussed. The mesh-current and node-voltage methods (Section 3.2.4) give $V_1 = I_1z_{11} + I_2z_{12}$, etc., and $I_1 = V_1y_{11} + V_2y_{12}$, etc., respectively. A further method relates V_1 and I_2 to I_1 and V_2 by hybrid (impedance and admittance) parameters. The four relationships are then obtained as follows:

General

$$\begin{pmatrix} V_1 \\ I_1 \end{pmatrix} = \begin{pmatrix} \boldsymbol{A} & \boldsymbol{B} \\ \boldsymbol{C} & \boldsymbol{D} \end{pmatrix}\begin{pmatrix} V_2 \\ -I_2 \end{pmatrix}$$

Impedance

$$\begin{pmatrix} V_1 \\ V_2 \end{pmatrix} = \begin{pmatrix} z_{11} & z_{12} \\ z_{21} & z_{22} \end{pmatrix}\begin{pmatrix} I_1 \\ I_2 \end{pmatrix}$$

Admittance

$$\begin{pmatrix} I_1 \\ I_2 \end{pmatrix} = \begin{pmatrix} y_{11} & y_{12} \\ y_{21} & y_{22} \end{pmatrix}\begin{pmatrix} V_1 \\ V_2 \end{pmatrix}$$

Hybrid

$$\begin{pmatrix} V_1 \\ I_2 \end{pmatrix} = \begin{pmatrix} h_{11} & h_{12} \\ h_{21} & h_{22} \end{pmatrix}\begin{pmatrix} I_1 \\ V_2 \end{pmatrix}$$

If the networks are passive, then $z_{12} = z_{21}$, $y_{12} = y_{21}$ and $h_{12} = -h_{21}$. If, in addition, the networks are symmetrical, then $\boldsymbol{A} = \boldsymbol{D}$, $z_{11} = z_{22}$ and $y_{11} = y_{22}$. If the networks are active (i.e. they contain sources), then reciprocity does not apply and there is no necessary relation between the terms of the 2 × 2 matrix.

3.2.6 Network topology

In multibranch networks the solution process is aided by representing the network as a graph of nodes and interconnections. The topology is the scheme of interconnections. A network is planar if it can be drawn on a closed spherical (or plane) surface without cross-overs. A non-planar network cannot be so drawn: a single cross-over can be eliminated if the network is drawn on a more complicated surface resembling a doughnut, and more cross-overs require closed surfaces with more holes.

The nomenclature employed in topology is as follows.

Graph A diagram of the network showing all the nodes, with each branch represented by a plain line.

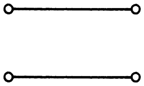
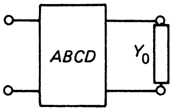
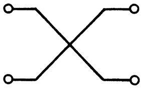
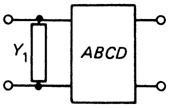
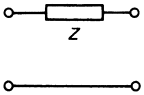
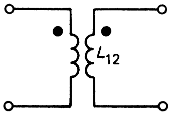
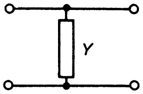
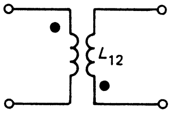
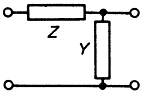
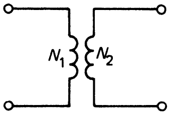
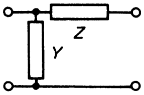
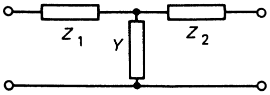
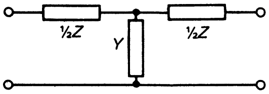
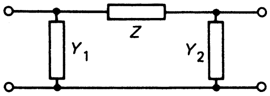
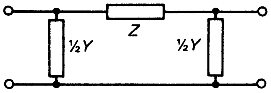
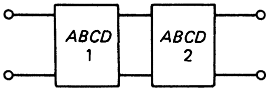
Tree Any arrangement of branches that connects all nodes together without forming loops. A *tree branch* is one branch of such a tree.

Link A branch that, added to a tree, completes a closed loop.

Tie set A loop of branches with one a link and the others tree branches.

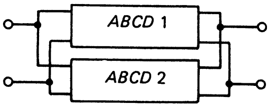
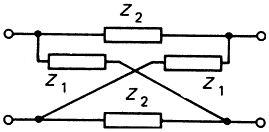
Cut set A set of branches comprising one tree branch, the other branches being tree links. A cut set dissociates two main portions of a network in such a way that replacing any one element destroys the dissociation.

Table 3.1 General ***ABCD*** two-port parameters

Network	*Matrix*	*Network*	*Matrix*
	Direct connection $\begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$		Loaded network $\begin{pmatrix} \boldsymbol{A} + \boldsymbol{B}Y_0 & \boldsymbol{B} \\ \boldsymbol{C} + \boldsymbol{D}Y_0 & \boldsymbol{D} \end{pmatrix}$
	Cross-connection $\begin{pmatrix} -1 & 0 \\ 0 & -1 \end{pmatrix}$		Shunted network $\begin{pmatrix} \boldsymbol{A} & \boldsymbol{B} \\ \boldsymbol{C} + \boldsymbol{A}Y_1 & \boldsymbol{D} + \boldsymbol{B}Y_1 \end{pmatrix}$
	Series impedance $\begin{pmatrix} 1 & Z \\ 0 & 1 \end{pmatrix}$		Mutual inductance $\begin{pmatrix} 0 & -\mathrm{j}\omega L_{12} \\ -1/\mathrm{j}\omega L_{12} & 0 \end{pmatrix}$
	Shunt admittance $\begin{pmatrix} 1 & 0 \\ Y & 1 \end{pmatrix}$		Mutual inductance $\begin{pmatrix} 0 & -\mathrm{j}\omega L_{12} \\ 1/\mathrm{j}\omega L_{12} & 0 \end{pmatrix}$
	L network $\begin{pmatrix} 1 + YZ & Z \\ Y & 1 \end{pmatrix}$		Ideal transformer $\begin{pmatrix} N_1/N_2 & 0 \\ 0 & N_2/N_1 \end{pmatrix}$
	L network $\begin{pmatrix} 1 & Z \\ Y & 1 + YZ \end{pmatrix}$		
	T network $\begin{pmatrix} 1 + YZ_1 & Z_1 + Z_2 + YZ_1Z_2 \\ Y & 1 + YZ_2 \end{pmatrix}$		
	Symmetrical T network $\begin{pmatrix} 1 + YZ/2 & Z(1 + YZ/4) \\ Y & 1 + YZ/2 \end{pmatrix}$		
	Π network $\begin{pmatrix} 1 + Y_2Z & Z \\ Y_1 + Y_2 + Y_1Y_2Z & 1 + Y_1Z \end{pmatrix}$		
	Symmetrical Π network $\begin{pmatrix} 1 + YZ/2 & Z \\ Y(1 + YZ/4) & 1 + YZ/2 \end{pmatrix}$		
	Cascaded networks $\begin{pmatrix} \boldsymbol{A}_1\boldsymbol{A}_2 + \boldsymbol{B}_1\boldsymbol{C}_2 & \boldsymbol{A}_1\boldsymbol{B}_2 + \boldsymbol{B}_1\boldsymbol{D}_2 \\ \boldsymbol{A}_2\boldsymbol{C}_1 + \boldsymbol{C}_2\boldsymbol{D}_1 & \boldsymbol{B}_2\boldsymbol{C}_1 + \boldsymbol{D}_1\boldsymbol{D}_2 \end{pmatrix}$		

cont'd

Table 3.1 *(continued)*

Network	*Matrix*
	Parallel networks $\begin{matrix} (A_1B_2+A_2B_1)/(B_1+B_2) & B_1B_2/(B_1+B_2) \\ C_1+C_2+\dfrac{(A_1-A_1)(D_1-D_2)}{B_1+B_2} & \dfrac{B_1D_2+B_2D_1}{B_1+B_2} \end{matrix}$
	Symmetrical lattice network $\begin{pmatrix} (Z_1+Z_2)(Z_1-Z_2) & 2Z_1Z_2/(Z_1-Z_2) \\ 2/(Z_1-Z_2) & (Z_1+Z_2)/(Z_1-Z_2) \end{pmatrix}$

Before setting up the equations for network solution, some guide is necessary in forming the proper number of independent equations. Given the network (*a*) in *Figure 3.10*, the first step is to draw the graph (*b*). Two of its possible trees are shown in (*c*). The trees are then used to set up the equations.

3.2.6.1 Network equations

Voltage The network diagram for the upper tree in *Figure 3.10*(*c*) is drawn in (*d*). Specifying the tree-branch voltages specifies also the voltages across the links. It is convenient to choose r as a reference node, leaving $n=6$ independent nodes requiring $n=6$ voltage equations.

Current A tree has no closed paths. As the links are added with specified currents, each creates one loop. Then the sum of the links m is the number of currents to be evaluated. For a network of b branches and n independent nodes, the number of independent meshes is $m=b-n$.

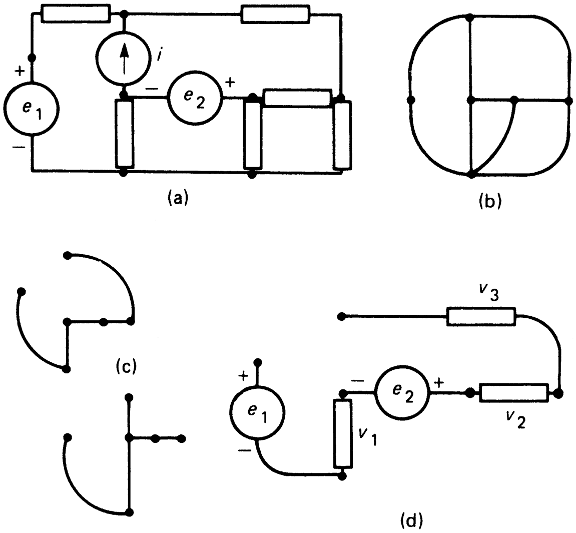


Figure 3.10 Network topology

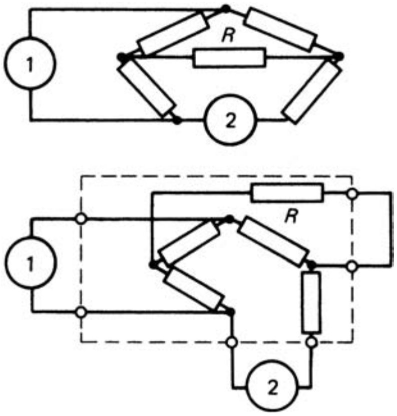


Figure 3.11 Conversion to a passive network

3.2.6.2 Ports

It is often helpful to place the sources outside the network and to regard their connections to the (now passive) remainder as ports. Again, branches of interest can be taken outside and used to terminate ports, as in *Figure 3.11*.

A multiport network (*Figure 3.12*) has the following characteristic definitions:

(a) All ports but one are open-circuited: a voltage V_1 is applied to port 1 and a current I_1 flows into it. Then V_1/I_1 is the *open-circuit* (*o.c.*) *driving-point impedance*

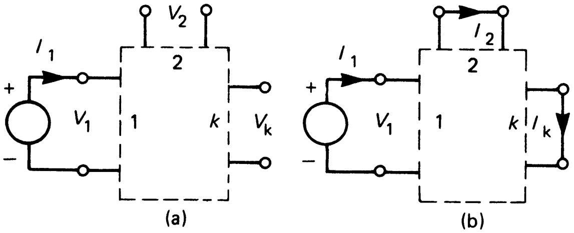


Figure 3.12 A multiport network

at port 1, V_k/I_1 is the *o.c. transfer impedance* from port 1 to port k, and V_k/V_1 is the *o.c. voltage ratio* of port k to port 1.

(b) All ports but one are short circuited: a current I_1 (requiring a voltage V_1) is fed in at port 1. Then I_1/V_1 is the *short-circuit* (*s.c.*) *driving-point admittance* at port 1, I_k/V_1 is the *s.c. transfer admittance* from port 1 to port k, and I_k/I_1 is the *s.c. current ratio* of port k to port 1.

3.2.7 Steady-state d.c. networks

The steady state implies that energy storage in electric and magnetic fields does not change, and only the resistance is significant. In *Figure 3.13* a source of constant e.m.f. E and internal resistance r provides a current I at terminal voltage V to a network represented by an equivalent resistance R. On open circuit ($R=\infty$), $I=0$ and $V=E$. As R is reduced the source provides a current $I=E/(R+r)=(E-V)/r$. The greatest output power $P=VI$ occurs for the condition $R=r$; further reduction of R reduces the network power, down to a short-circuit condition for $R=0$ and $V=0$ when the source power is dissipated entirely in r. The maximum-power condition is utilised only with sources whose power capability is very small.

3.2.8 Steady-state a.c. networks

An a.c. flows alternately in the specified *positive* direction and then in the *negative* direction in a circuit, repeating this cycle continuously. A graph of current or voltage to a time base shows the *waveform* as a succession of *instantaneous values*. In general there will be a maximum or *peak* value in both positive and negative half-periods where the current or voltage is greatest. The time for one complete *cycle* is the *period T*. The number of periods per second is the *frequency* $f=1/T$.

An a.c. is produced by an alternating voltage. Two such quantities may have a difference of *phase*, to which a precise meaning can be given only when the quantities are both sinusoidal functions of time.

3.2.8.1 Root-mean-square (r.m.s.) value

The numerical value assigned to an a.c. or voltage is normally defined in terms of mean power in a pure resistor. An *a.c.* of 1 A is that which produces heat energy at the same mean rate as a direct current of 1 A in the same non-reactive resistor. If i is the instantaneous value of an a.c. in a pure resistance R, the heat developed in a time element dt is $dw=i^2R\,dt$. The mean rate (i.e. the mean power) over a complete period T is

$$P=(1/T)\int_0^T dw=(1/T)\int_0^T i^2R\,dt=I^2R$$

and I is the r.m.s. value of the current. An *alternating voltage* is defined in a similar way; the instantaneous power is v^2/R, and the mean is V^2/R where V is the square root of the mean v^2.

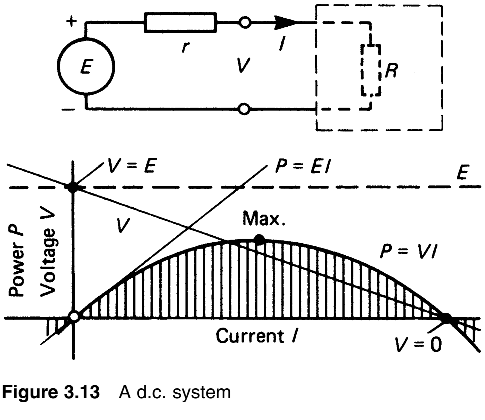


Figure 3.13 A d.c. system

Table 3.2 Values of alternating quantities (peak = a)

Waveform	*r.m.s.*	*Mean*	K_f	K_p
Sinusoidal	$a/\sqrt{2}$	$a(2/\pi)$	1.11	1.41
Half-wave rectified sine	$a/2$	$a/\pi\psi$	1.57	2.0
Full-wave rectified sine	$a/\sqrt{2}$	$a(2/\pi)$	1.11	1.41
Rectangular	a	a	1.0	1.0
Triangular	$a/\sqrt{3}$	$a/2$	1.16	1.73

In some cases the *peak* or the *mean* value of the current or voltage waveform is more significant, particularly with asymmetric, pulse or rectified waveforms. The value to be understood by the term 'mean' is then obvious. In the case of a symmetrical wave, the *half-period mean* value is intended, as the mean over a complete period is zero. *Table 3.2* gives the mean and r.m.s. values for a number of typical waveforms, together with the values of

$$\text{Form factor } K_f = \text{r.m.s./mean}$$
$$\text{Peak factor } K_p = \text{peak/r.m.s.}\psi$$

The techniques developed for the solution of steady-state a.c. networks depend on the waveform. One technique applies to purely sinusoidal quantities, another to periodic but non-sinusoidal waveforms. In each case the network is assumed to be linear so that the principle of superposition is valid.

3.2.9 Sinusoidal alternating quantities

For pure sinusoidal waveforms, a current can be expressed as a function of time, $i=i_m\sin(2\pi ft)=i_m\sin(\omega t)$, completing f cycles in 1 s with a period $T=1/f$. The quantity $2\pi f$ is contracted to ω, the *angular frequency*. The sine-wave shape has the advantages that (i) it is mathematically simple and its integral and differential are both cosinusoidal, (ii) it is a waveform desirable for power generation, transmission and utilisation, and (iii) it lends itself to phasor and complexor representation.

The graph of a sinusoidal current or voltage of frequency f can be plotted to a time-angle base ωt by use of trigonometric tables. Alternatively it can be represented by the projection of a line of length equal to the peak value and rotating counter-clockwise at angular speed $\omega\psi$about one end O. A *stationary* line can represent the sine wave, particularly in relation to other sine waves of the same frequency but 'out of step'. Two such waves, say v and i with peak values v_m and i_m, respectively, can be written

$$v=v_m\sin\omega t \quad \text{and} \quad i=i_m\sin(\omega t-\phi)\Leftarrow$$

and drawn as in *Figure 3.14*, the phase difference or phase angle between them being ϕ rad. Then two lines, OA and OB, having an angular displacement ϕ, can represent the two waves in peak magnitude and relative time phase.

Although developed from rotating lines of peak-value length, it is more convenient to change the scale and treat the lengths as r.m.s. values. The processes of addition and

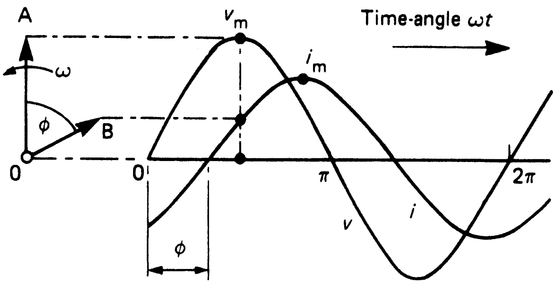


Figure 3.14 Phasors

subtraction of r.m.s. values are performed as if the lines were co-planar vector forces in mechanics. Physically, however, the lines are not vectors: they substitute for scalar quantities, alternating sinusoidally with time. They are termed *phasors*. Certain associated quantities, such as impedance, admittance and apparent power, can also be represented by directed lines, but as they are not sinusoids they are termed *complexors* or *complex operators*. Both phasors and complexors can be dealt with by application of the theory of complex numbers. The definitions concerned are listed below.

Complexor A generic term for a non-vector quantity expressed as a complex number.
Phasor A complexor (e.g. voltage or current) derived from a time-varying sinusoidal quantity and expressed as a complex number.
Complex operator A complexor derived for the ratio of two phasors (e.g. impedance and admittance); or a complexor which, operating on a phasor, gives another phasor (e.g. $V = IZ$, in which V and I are phasors, but Z is a complex operator).

3.2.9.1 Complexor algebra

The four arithmetic processes for complexors are applications of the theory of complex numbers. Complexor $\boldsymbol{a}$ in *Figure 3.15* can be expressed by its magnitude a and its angle θ with respect to an arbitrary 'datum' direction (here taken as horizontal) as the simple *polar* form $\boldsymbol{a} = a \angle \theta$. Alternatively it can be written as $\boldsymbol{a} = p + \mathrm{j}q$, the *rectangular* form, in terms of its projection p on the datum and q on a quadrature axis at right angles thereto: q (as a scalar magnitude along the datum) is rotated counter-clockwise by angle $\frac{1}{2}\pi$ rad (90°) by the operator j. Two successive operations by j (written as j^2) give a rotation of π rad (180°), making the original $+q$ into $-q$, in effect a multiplication by -1. Three operations (j^3) give $-\mathrm{j}q$ and four give $+q$. Thus any complexor can be located in the complex datum–quadrature plane. Further obvious forms are the *trigonometric*, $\boldsymbol{a} = a(\cos\theta + \mathrm{j}\sin\theta)$, and the *exponential*, $\boldsymbol{a} = a\exp(\mathrm{j}\theta)$. Summarising, the four descriptions are:

Polar: $\boldsymbol{a} = a \angle \theta$
Rectangular: $\boldsymbol{a} = p + \mathrm{j}q$

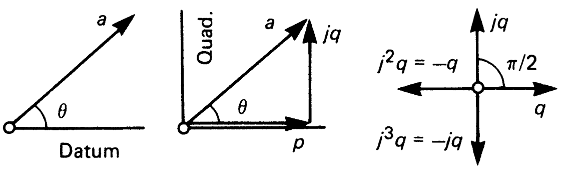


Figure 3.15 Complexors

Exponential: $\boldsymbol{a} = a\exp(\mathrm{j}\theta)$
Trigonometric: $\boldsymbol{a} = a(\cos\theta + \mathrm{j}\sin\theta)$

where $a = \sqrt{(p^2 + q^2)}$ and $\theta = \arctan(q/p)$.

Consider complexors $\boldsymbol{a} = p + \mathrm{j}q = a \angle \alpha$ and $\boldsymbol{b} = r + \mathrm{j}s = b \angle \beta$. The basic manipulations are:

Addition $\boldsymbol{a} + \boldsymbol{b} = (p + r) + \mathrm{j}(q + s)$

Subtraction $\boldsymbol{a} - \boldsymbol{b} = (p - r) + \mathrm{j}(q - s)$

Multiplication The exponential and polar forms are more direct than the rectangular or trigonometric:

$$\boldsymbol{ab} = (pr - qs) + \mathrm{j}(qr + ps)$$
$$= ab\exp[\mathrm{j}(\alpha + \beta)] = ab \angle (\alpha + \beta)$$

Division Here also the angular forms are preferred:

$$\boldsymbol{a}/\boldsymbol{b} = [(pr + qs) + \mathrm{j}(qr - ps)]/(r^2 + s^2)$$
$$= (a/b)\exp[\mathrm{j}(\alpha - \beta)] = (a/b) \angle (\alpha - \beta)$$

Conjugate The conjugate of a complexor $a = p + \mathrm{j}q = a \angle \alpha$ is $\boldsymbol{a}^* = p - \mathrm{j}q = a \angle (-\alpha)$, the quadrature component (and therefore the angle) being reversed. Then

$$\boldsymbol{ab}^* = ab \angle (\alpha - \beta)$$
$$\boldsymbol{a}^*\boldsymbol{b} = ab \angle (\beta - \alpha)$$
$$\boldsymbol{a}^*\boldsymbol{a} = \boldsymbol{aa}^* = a^2 = p^2 + q^2$$

The last expression is used to 'rationalise' the denominator in the complexor division process.

3.2.9.2 Impedance and admittance operators

Sinusoidal voltages and currents can be represented by phasors in the expressions $V = IZ = I/Y$ and $I = VY = V/Z$. Current and voltage phasors are related by multiplication or division with the complex operators Z and Y. Series resistance R and reactance $\mathrm{j}X$ can be arranged as a right-angled triangle of hypotenuse $Z = \sqrt{(R^2 + X^2)}$ and the angle between Z and R is $\theta = \arctan(X/R)$. The relation between Z and Y for the same series network elements with $Z = R + \mathrm{j}X$ is

$$Y = \frac{1}{Z} = \frac{1}{R + \mathrm{j}X} = \frac{R - \mathrm{j}X}{(R + \mathrm{j}X)(R - \mathrm{j}X)} = \frac{R - \mathrm{j}X}{R^2 + X^2}$$
$$= R/Z^2 - \mathrm{j}(X/Z^2) = G - \mathrm{j}B$$

where G and B are defined in terms of R, X and Z. The *series* components R and X become *parallel* branches in Y, one a pure conductance, the other a pure susceptance. Further, a positive-angled impedance has, as inverse equivalent, a negative-angled admittance (*Figure 3.16*).

The impedance and phase angle of a number of circuit combinations are given in *Table 3.3*.

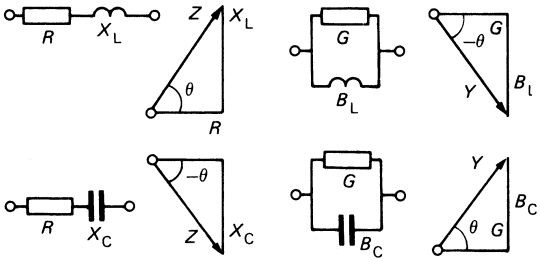
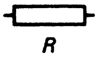
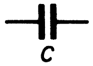
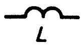
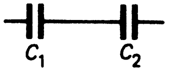
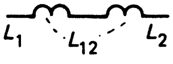
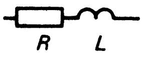
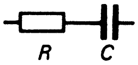
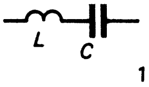
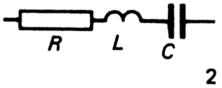
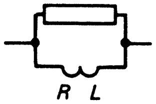
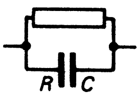
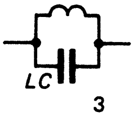


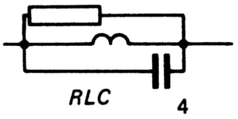
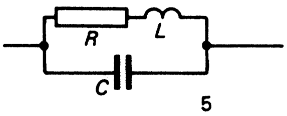
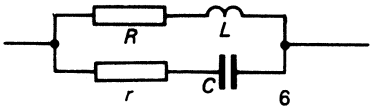
Figure 3.16 Impedance and admittance triangles

Table 3.3 Impedance of network elements at angular frequency ω (rad/s)

Impedance:	$Z = R + jX = \|Z\| \angle \theta$	$\|Z\| = \sqrt{(R^2 + X^2)}$	$\theta = \arctan(X/R)$
Admittance:	$Y = 1/Z = \|Y\| \angle (-\theta)$	$\|Y\| = \sqrt{[(R/Z^2)^2 + (X/Z^2)^2]}$	$\theta = -\arctan(X/R)$

						
Z: R	$1/j\omega C$	$j\omega L$	$(C_1 + C_2)/j\omega C_1C_2$	$j\omega(L_1 + 2L_{12})$	$R + j\omega L$	$R + 1/j\omega C$
θ: 0	$-\pi/2$	$+\pi/2$	$-\pi/2$	$+\pi/2$	$\arctan(\omega L/R)$	$-\arctan(j\omega CR)$

 1	 2			 3
Z: $j(\omega L - 1/\omega C)$	$R + j(\omega L - 1/\omega C)$	$\omega LR\dfrac{\omega L + jR}{R^2 + \omega^2L^2}$	$R\dfrac{1 - j\omega CR}{1 + \omega^2C^2R^2}$	$\dfrac{j\omega L}{1 - \omega^2LC}$
θ: $\pm\pi/2$	$\arctan[(\omega L - 1/\omega C)/R]$	$\arctan(R/\omega L)$	$-\arctan(\omega CR)$	$\pm\pi/2$

 4	 5	 6
$Z: \dfrac{1/R - j(\omega C - 1/\omega L)}{(1/R)^2 + (\omega C - 1/\omega L)^2}$	$\dfrac{R + j\omega[L(1 - \omega^2LC) - CR^2]}{(1 - \omega^2LC)^2 + \omega^2C^2R^2}$	$\dfrac{A + jB}{(R + r)^2 + (\omega L - 1/\omega C)^2}$ $A = Rr(R + r) + \omega^2L^2r + R/\omega^2C^2$ $B = \omega Lr^2 - R^2/\omega C - (L/C)(\omega L - 1/\omega C)$
θ: $-\arctan[R(\omega C - 1/\omega L)]$	$\arctan\{\omega[L(1 - \omega^2 LC) - CR^2]/R\}$	$\arctan(B/A)$

Resonance conditions for LC networks numbered 1–6 above, for $\omega = \omega_0 = 1/\sqrt{(LC)}$:
(1) $|Z| = 0$, $\theta = 0$; (2) $|Z| = R$, $\theta = 0$; (3) $|Z| = \infty$, $\theta = 0$; (4) $|Z| = R$, $\omega = 0$;
(5) $|Z| = L/CR$, $\theta = -\arctan(\omega CR)$ for $R \ll \omega L$; (6) $|Z| = R$ (const.) for $R = r = \sqrt{(L/C)}$

Impedance and admittance loci If the characteristics of a device or a circuit can be expressed in terms of an equivalent circuit in which the impedances and/or admittances vary according to some law, then the current taken for a given applied voltage (or the voltage for a given current) can be obtained graphically by use of an admittance or impedance locus diagram.

In *Figure 3.17*(*a*), let OP represent an impedance $Z = R + jX$ and OQ the corresponding admittance $Y = G - jB$. Point Q is obtained from P by finding first the *geometric* inverse point Q′ such that $OQ' = 1/OP$ to scale, and then reflecting OQ′ across the datum line to give OQ and thus a reversed angle $-\theta$, a process termed *complexor* inversion. If Z has successive values $Z_1, Z_2, \ldots$, on the impedance locus, the corresponding admittances $Y_1, Y_2, \ldots$, lie on the admittance locus. The inversion process may be point-by-point, but in many cases certain propositions can reduce the labour:

(1) *Inverse of a straight line*—the geometric inverse of a straight line AB about a point O not on the line is a circle passing through O with its centre M on the perpendicular OC from O to AB (*Figure 3.17*(*b*)). Then A′ is the geometric inverse of A, B′ of B, etc.; also, A is the inverse of A′, B of B′, etc.

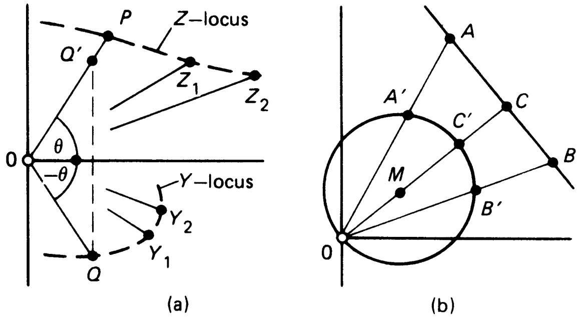


Figure 3.17 Inversion

(2) *Inverse of a circle*—from the foregoing, the geometric inverse of a circle about a point O on its circumference is a straight line. If, however, O is not on the circumference, the inverse is a second circle between the same tangents; but the distances OM and OM′ from the origin O to the centres M and M′ of the circles are not inverses of each other.

The choice of scales arises in the inversion process: for example, the inverse of an impedance $Z = 50 \angle 70° \Leftarrow \Omega$ is $Y = 0.02 \angle (-70°)$ S. It is usually possible to decide on a scale by taking a salient feature (such as a circle diameter) as a basis.

3.2.9.3 Power

The instantaneous power delivered to a load is the product of the instantaneous voltage v and current i. Let $v = v_m \sin \omega t$ and $i = i_m \sin(\omega t - \phi)$ as in *Figure 3.18(a)*; then the instantaneous power is

$$p = \tfrac{1}{2} v_m i_m [\cos\phi - \cos(2\omega t - \phi)]$$

This is a quantity fluctuating at angular frequency $2\omega \Leftarrow$ with, in general, excursions into negative power (i.e. that returned by the load to the source). Over an integral number of periods the mean power is

$$P = \tfrac{1}{2} v_m i_m \cos\phi \Leftarrow= VI\cos\phi$$

where V and I are r.m.s. values.

Now resolve i into the active and reactive components

$$i_p = (i_m \cos\phi)\sin\omega t \text{ and } i_q = (i_m \sin\phi)\sin(\omega t - \tfrac{1}{2}\pi)$$

as in *Figure 3.18(b)*; then the instantaneous power can be written

$$p = (v_m(i_m\cos\phi)\sin^2\omega t - v_m(i_m\sin\phi)\sin\omega t\cos\omega t$$

Over a whole number of periods the average of the first term is

$$P = \tfrac{1}{2} v_m i_m \cos\phi \Leftarrow= VI\cos\phi$$

giving the *average* rate of energy transfer from source to load. The second term is a double-frequency sinusoid of average value zero, the energy flow changing direction rhythmically between source and load at a *peak* rate

$$Q = \tfrac{1}{2} v_m i_m \sin\phi \Leftarrow= VI\sin\phi$$

The power conditions thus summarise to the following:

Active power P The *mean* of the instantaneous power over an integral number of periods giving the mean rate of energy transfer from source to load in watts (W).

Reactive power Q The *maximum* rate of energy interchange between source and load in reactive volt-amperes (var).

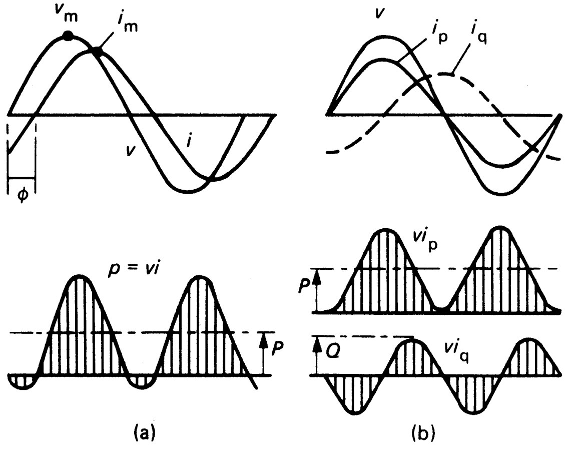


Figure 3.18 Active and reactive power

Apparent power S The product of the r.m.s. voltage and current in volt-amperes (V-A).

Both P and Q represent real power. The apparent power S is not a power at all, but is an arbitrary product VI. Nevertheless, because of the way in which P and Q are defined, we can write

$$P^2 + Q^2 = (VI)^2(\cos^2\phi \Leftarrow + \sin^2\phi) = (VI)^2$$

whence $S = \sqrt{(P^2 + Q^2)}$, a convenient combination of mean active power with peak power circulation.

Complex power The active and reactive powers can be determined for voltage and current phasors by

$$S = P \pm jQ = VI^* \quad \text{or} \quad S = V^*I$$

using the conjugate of either I or V.

Power factor The ratio of active to apparent power, $P/S = \cos\phi \Leftarrow$ for sinusoidal conditions.

3.2.9.4 Resonance

A condition of resonance occurs when the load contains two forms of energy-storing element (L and C) such that, at the frequency of operation, the two energies are equal. The reactive power requirements are then satisfied internally, as the inductor releases energy at the rate that the capacitor requires it. The source supplies only the active power demand of the energy-dissipating load components, the load externally appearing to be purely resistive.

Acceptor resonance The series *RLC* circuit in *Figure 3.19(a)* has, at angular frequency ω, the impedance $Z = R + jX$, where X is $\omega L - 1/\omega C$, which for $\omega = \omega_0 = 1/\sqrt{(LC)}$ is zero. The impedance is then $Z = R$ and the input current has a maximum $I_0 = V/R$, conditions of *acceptor* resonance. Internally, large voltages appear across the reactive components, *viz.*

$$V_L = I_0\omega L = V\omega_0(L/R) \text{ and } V_C = I_0(1/\omega_0 C) = V/\omega_0(CR) \Leftarrow$$

The terms L/R and $1/CR$ are the time constants of the reactive elements, and $\omega_0 L/R$ is the Q value of a practical inductor of

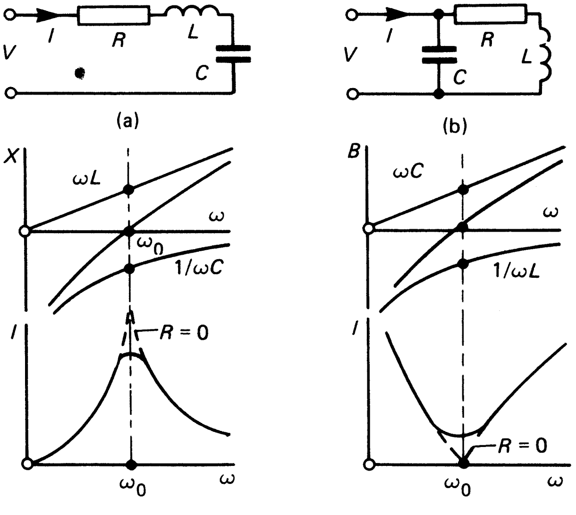


Figure 3.19 Resonance

inductance L and loss-resistance R. The Q value may be large (e.g. 100) for resonance at a high frequency.

Rejector resonance This occurs in a parallel combination of L and C, the energies circulating around the closed LC loop. If in *Figure 3.19(b)* the resistance R is zero, the terminal input admittance vanishes at angular frequency $\omega_0 = 1/\sqrt{(LC)}$, with $\omega_0 C = 1/\omega_0 L$ and an input susceptance $B = 0$. Where the circuit contains resistance R the resonance conditions are less definite. Three possible criteria are: (i) $\omega_0 = 1/\sqrt{(LC)}$, (ii) the input admittance is a minimum, and (iii) the input admittance is purely conductive. All three criteria are satisfied simultaneously in the simple acceptor circuit, but differ in rejector conditions; however, where resonance is an intended property of the circuit, the differences are small.

Some expressions for resonance are given for six circuit arrangements in *Table 3.3*.

3.2.10 Non-sinusoidal alternating quantities

Periodic but non-sinusoidal currents occur: (i) with non-sinusoidal e.m.f. sources, (ii) with sinusoidal sources applied to non-linear loads, and (iii) with any combination of (i) and (ii).

3.2.10.1 Fourier series

Any univalued periodic waveform can be represented as a summation of sine waves comprising a *fundamental*, where frequency is that of the periodic occurrence, and a series of *harmonic* waves of frequency 2, 3, ..., n times that of the fundamental. The Fourier series for a periodic function $y = f(x)$ takes either of the following equivalent forms:

$$(1)\ y = c_0 + c_1 \sin(x + \alpha_1) + c_2 \sin(2x + \alpha_2) + \cdots \Leftarrow$$

$$(2)\ y = c_0 + a_1 \cos x + a_2 \cos 2x + \cdots + a_n \cos nx + b_1 \sin x + b_2 \sin 2x + \cdots + b_n \sin nx$$

where c_0 is a constant, $c_n = \sqrt{(a_n{}^2 + b_n{}^2)}$ and $\alpha_n = \Leftarrow$ arctan (a_n/b_n). The coefficients of the terms are given by

$$c_0 = (1/2\pi)\int_0^{2\pi\psi} f(x)\,\mathrm{d}x = \text{mean of the wave over one period}$$

$$a_n = (1/\pi)\int_0^{2\pi\psi} f(x) \cos nx\,\mathrm{d}x$$

$$b_n = (1/\pi)\int_0^{2\pi} f(x) \sin nx\,\mathrm{d}x$$

These can be evaluated mathematically for simple cases. The work may sometimes be reduced by inspection: thus $c_0 = 0$ for a wave symmetrical about the baseline; or only odd-order harmonics may be present.

3.2.10.2 Analysis

The series for a range of mathematically tractable waveforms are given in *Table 1.10*. For experimentally derived waveforms there are several methods, but none yields the amplitude of higher order harmonics without considerable labour, unless a computer program is available.

A particular harmonic, say the nth, may be found by superimposing n copies of the wave, displaced relatively by $2\pi/n$, $4\pi/n, \ldots$, and adding the corresponding ordinates. The result is a wave of frequency n times that of the harmonic sought (with the addition, however, of harmonics of orders kn where k is an integer). The method gives also the phase angle α_n.

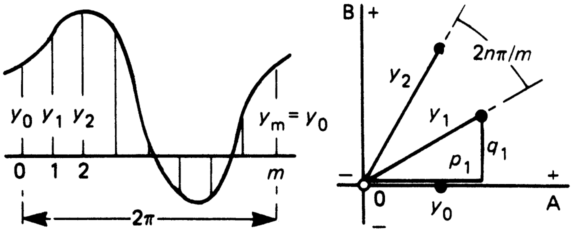


Figure 3.20 Graphical harmonic analysis

A semi-graphical method is shown in *Figure 3.20*. The base of a complete period $2\pi\psi$is divided into m parts, the corresponding ordinates being $y_0, y_1, y_2, \ldots, y_m$. Construct axes OA and OB; set out the radii y_0 to y_m $(=y_0)$ at angles 0, $2n\pi/m$, $4n\pi/m, \ldots$, from the axis OA. Then project the extremities horizontally (p) and vertically (q), and take the sum of the two sets of projections with due regard to their sign. Then for the nth harmonic

$$a_n = \Leftarrow \frac{2}{\pi\psi}\sum_0^{m-1} p \quad \text{and} \quad b_n = \Leftarrow \frac{2}{\pi\psi}\sum_0^{m-1} q$$

The labour is reduced if $2\pi/m$ is a simple fraction of 2π, for then some groups of radii are coincident.

3.2.10.3 Power

The r.m.s. value of a current

$$i = I_0 + i_1 \sin(\omega t + \alpha_1) + i_2 \sin(2\omega t + \alpha_2) + \cdots \Leftarrow$$

is obtained from the square root of the average squared value, resulting in

$$I = \sqrt{(I_0^2 + \tfrac{1}{2}i_1^2 + \tfrac{1}{2}i_2^2 + \cdots)} = \sqrt{(I_0^2 + I_1^2 + I_2^2 + \cdots)} \Leftarrow$$

where $I_1 = i_1/\sqrt{2}$, $I_2 = i_2/\sqrt{2}$, etc., are the r.m.s. values of the individual harmonic components. The r.m.s. voltage is obtained in a similar way.

Power The instantaneous power p in a circuit with an applied voltage

$$v = v_1 \sin(\omega t + \alpha_1) + v_2 \sin(2\omega t + \alpha_2) + \cdots \Leftarrow$$

producing a current

$$i = i_1\ \sin(\omega t + \alpha_1 - \phi_1) + i_2\ \sin(2\omega t + \alpha_2 - \phi_2) + \cdots \Leftarrow$$

is the product vi: this includes (i) a series of the form

$$v_n i_n \sin(n\omega t + \alpha_n) \sin(n\omega t + \alpha_n - \phi_n) \Leftarrow$$

all terms of which have a fundamental-period average $\frac{1}{2} v_n i_n \cos \phi_n$; and (ii) a series of the form

$$v_p i_q \sin(p\omega t + \alpha_p) \sin(q\omega t + \alpha_q - \phi_q) \Leftarrow$$

which, over a fundamental period, averages zero. Power is circulated by a voltage and a current of different frequencies, but the circulation averages zero. The mean (active) power is therefore

$$P = \tfrac{1}{2} v_1 i_1 \cos \phi_1 + \tfrac{1}{2} v_2 i_2 \cos \phi_2 + \cdots + \tfrac{1}{2} v_n i_n \cos \phi_n + \cdots \Leftarrow$$
$$= V_1 I_1 \cos \phi_1 + V_2 I_2 \cos \phi_2 + \cdots + V_n I_n \cos \phi_n + \cdots \Leftarrow$$

where the capital letters denote component r.m.s. values. Thus the harmonics contribute power separately.

Power factor The ratio of the active power P to the apparent power S is

$$P/S = (V_1 I_1 \cos\phi_1 + \cdots + V_n I_n \cos\phi_n + \cdots)/VI$$

This may be less than unity even with all phase angles zero if the ratio V_n/I_n is not the same for each component. Where the applied voltage is a *pure sinusoid* of fundamental frequency there can be no harmonic powers; the active power is $P = V_1 I_1 \cos\phi_1$. Then

$$P/S = (V_1 I_1 \cos\phi_1)/V_1 I = (I_1/I)\cos\phi_1$$

where $I_1/I = I_1/\sqrt{(I_1^2 + I_2^2 + \cdots + I_n^2)} = \delta$, the *distortion factor*. The overall power factor is consequently $\delta\cos\phi_1$. This is typical of circuits containing non-linear elements.

3.2.11 Three-phase systems

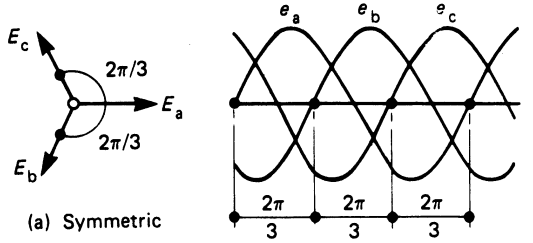
A symmetrical m-phase system has m source e.m.f.s, all of the same waveform and frequency, and displaced $2\pi/m$ rad or $1/m$ period in time; m is most commonly 3, but is occasionally 6, 12 or 24.

Symmetric three-phase system In *Figure 3.21*(*a*) the symmetric sinusoidal three-phase system has source e.m.f.s in phases A, B and C given by

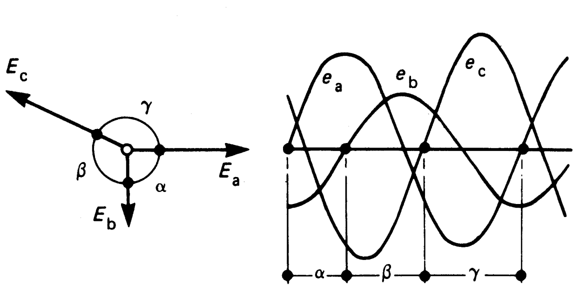
$$e_a = e_m \sin\omega t; \quad e_b = e_m \sin(\omega t - 2\pi/3); \quad e_c = e_m \sin(\omega t - 4\pi/3) \Leftarrow$$

The instantaneous sum of the phase e.m.f.s (and also the phasor sum of the corresponding r.m.s. phasors E_a, E_b and E_c) is zero.

Asymmetric three-phase system The asymmetric system in *Figure 3.21*(*b*) has, in general, unequal phase voltages and phase displacements. Such asymmetry may occur in machines with unbalanced phase windings and in power supply systems when faults occur; the usual method of dealing with asymmetry is described in Section 3.2.12. Attention here is confined to the basic symmetric cases.



(a) Symmetric



(b) Asymmetric

Figure 3.21 Three-phase systems

3.2.11.1 *Phase interlinkage*

While individual phase sources can be used separately, they are generated in the same machine and are normally interlinked. Using r.m.s. phasors, let the e.m.f. in a generator winding XY be such as to drive positive current out at X: then X is positive to Y and its e.m.f. E_{XY} is represented by an arrow with its point at X. The e.m.f. E_{YX} between terminals Y and X is therefore $-E_{XY}$. Further, when two windings XN and YN have a common terminal N, the e.m.f.s are

X to Y: $E_{XY} = E_{XN} - E_{YN}$
Y to X: $E_{YX} = E_{YN} - E_{XN} = -E_{XY}$

Common phase interconnections are shown in *Figure 3.22*.

3.2.11.2 *Star*

Let the phase e.m.f.s be E_{an}, E_{bn} and E_{cn} with an arbitrary positive direction outward from the star-point N. Then the line e.m.f.s are

$$E_{ab} = E_{an} - E_{bn}; \quad E_{bc} = E_{bn} - E_{cn}; \quad E_{ca} = E_{cn} - E_{an}$$

These are of magnitude $\sqrt{3}$ times that of a phase e.m.f., and provide a symmetric three-phase system of line e.m.f.s, with E_{ab} leading E_{an} by 30°. Thus $E_l = \sqrt{3}\,E_{ph}$ and $I_t = I_{ph}$, the subscripts l and ph referring to line and phase quantities respectively.

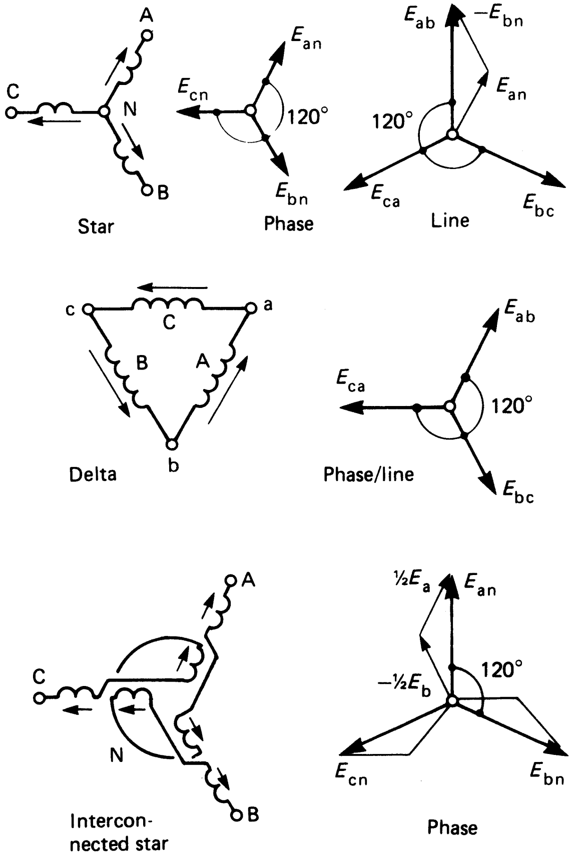


Figure 3.22 Three-phase interconnections

3.2.11.3 Delta

The line-to-line e.m.f. is that of the phase across which the lines are connected. The line current is the difference of the currents in the phases forming the line junction, so that the relations for symmetric loading are $E_l = E_{ph}$ and $I_l = \sqrt{3}\, I_{ph}$.

3.2.11.4 Interconnected star

A line-to-neutral e.m.f. comprises contributions from successive half-phases and sums to $\frac{1}{2}\sqrt{3}$ of a complete phase e.m.f. The line-to-line e.m.f. is $1\frac{1}{2}$ times the magnitude of a complete phase e.m.f. and the line current is numerically equal to the phase current.

3.2.11.5 Power

The total power delivered to or absorbed by a polyphase system, be it symmetric and balanced or not, is the algebraic sum of the individual phase powers. Consider an m-phase system with instantaneous line currents $i_1, i_2, \ldots, i_m$, the algebraic sum of which is zero by the Kirchhoff node law. Let the voltages of the input (or output) terminals, with reference to a common point X, be $v_1 - v_x$, $v_2 - v_X, \ldots, v_m - v_x$; then the instantaneous powers will be $(v_1 - v_x)i_1$, $(v_2 - v_x)i_2, \ldots$, $(v_m - v_x)i_m$, which together sum to the total instantaneous power p. There is no restriction on the choice of X; it can be any of the terminals, say M. In this case $v_m - v_x = v_m - v_m = 0$, and the power summation has only $m-1$ terms. The average power over a full period T is, therefore,

$$P = (1/T)\int_0^T [(v_1 - v_m)i_1 + \cdots + (v_{m-1} - v_m)i_{m-1}]\, dt$$

The first term of the sum in brackets represents the indication of a wattmeter with i_1 in its current circuit and $v_1 - v_m$ across its volt circuit, i.e. connected between terminals 1 and M. It follows that three wattmeters can measure the power in a three-phase four-wire system, and two in a three-phase three-wire system. Some of the common cases are listed below.

(1) *Three-phase, four-wire, load unbalanced*—The connections are shown in *Figure 3.23*(*a*). Wattmeters W_1, W_2 and W_3 measure the phase powers separately. The total power is the sum of the indications:

$$P = P_1 + P_2 + P_3$$

(2) *Three-phase, four-wire, load balanced*—with the connections shown in *Figure 3.23*(*a*), all the meters read the same. Two of the wattmeters can be omitted and the reading of the remaining instrument multiplied by 3.

(3) *Three-phase, three-wire, load unbalanced*—two wattmeters are connected with their current circuits in any pair of lines, as in *Figure 3.23*(*b*). The total power is the algebraic sum of the readings, regardless of waveform. A two-element wattmeter summates the power automatically; with separate instruments, one will tend to read reversed under certain conditions, given below.

(4) *Three-phase, three-wire, load balanced*—with sinusoidal voltage and current the conditions in *Figure 3.23*(*c*) obtain. Wattmeters W_1 and W_2 indicate powers P_1 and P_2 where

$$P_1 = V_{ab}I_a \cos(30° + \phi) = V_l I_l \cos(30° + \phi)$$
$$P_2 = V_{cb}I_c \cos(30° - \phi) = V_l I_l \cos(30° - \phi)$$

The total active power $P = P_1 + P_2$ is therefore

$$P = V_l I_l[\cos(30° + \phi) + \cos(30° - \phi)] = \sqrt{3} V_l I_l \cos\phi$$

where $\cos\phi$ is the *phase* power factor. The algebraic difference is $P_1 - P_2 = V_l I_l \sin\phi$, whence the reactive power is given by

$$Q = \sqrt{3} V_l I_l \sin\phi = \sqrt{3}(P_1 - P_2)$$

and the phase angle can be obtained from $\phi = \arctan(Q/P)$. For $\phi = 0$ (unity power factor) both wattmeters read alike; for $\phi = 60°$ (power factor 0.5 lag) W_1 reads zero; and for lower lagging power factors W_1 tends to read backwards.

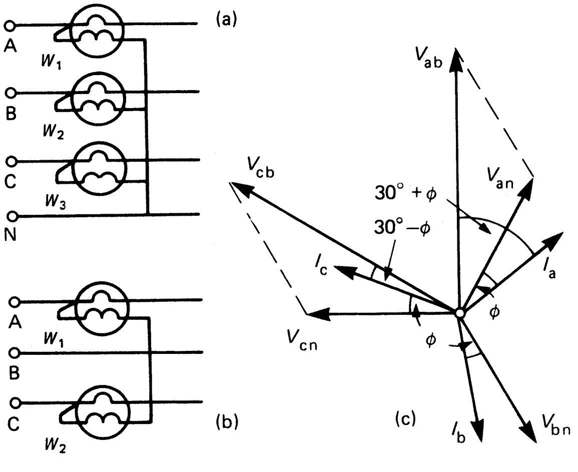


Figure 3.23 Three-phase power measurement

The active power of a single phase has a double-frequency pulsation (*Figure 3.18*). For the asymmetric two-phase system under balanced conditions and a phase displacement of 90°, and for all symmetric systems with $m = 3$ or more, the total power is constant.

3.2.11.6 Harmonics

Considering a symmetrical balanced system of three-phase non-sinusoidal voltages, and omitting phase displacements (which are in the context not significant), let the voltage of phase A be

$$v_a = v_1 \sin\, \omega t + v_2 \sin\, 2\omega t + v_3 \sin\, 3\omega t + \cdots \Leftarrow$$

Writing $\omega t - \frac{2}{3}\pi$ and $\omega t - \frac{4}{3}\pi$, respectively, for phases B and C, and simplifying, we obtain

$$v_a = v_1 \sin\, \omega t + v_2 \sin\, 2\omega t + v_3 \sin\, 3\omega t + \cdots \Leftarrow$$
$$v_b = v_1 \sin\, (\omega t - \tfrac{2}{3}\pi) + v_2 \sin\, 2(\omega t - \tfrac{4}{3}\pi) + v_3 \sin\, 3\omega t + \cdots \Leftarrow$$
$$v_c = v_1 \sin\, (\omega t - \tfrac{4}{3}\pi) + v_2 \sin\, 2(\omega t - \tfrac{2}{3}\pi) + v_3 \sin\, 3\omega t + \cdots \Leftarrow$$

The fundamentals have a normal $2\pi/3$ rad (120°) phase relation in the sequence ABC, as also do the 4th, 7th, 10th, ..., harmonics. The 2nd (and 5th, 8th, 11th, ...) harmonics have the $2\pi/3$ rad phase relation but of reversed sequence ACB. The *triplen* harmonics (those of the order of a multiple of 3) are, however, *co-phasal* and form a zero-sequence set.

The relation $V_l = \sqrt{3}\, V_{ph}$ in a three-phase star-connected system is applicable only for sine waveforms. If harmonics are present, the line- and phase-voltage waveforms differ because of the effective phase angle and sequence of the harmonic components. The nth harmonic voltages to neutral in two successive phases AB are $v_n \sin n\omega t$ and $v_n \sin n(\omega t - \frac{2}{3}\pi)$, and between the corresponding line terminals the nth harmonic voltage is $2v_n \sin n(\frac{1}{3}\pi)$. For triplen harmonics this is zero; hence no triplens are present in balanced line voltages because, in the associated phases, their components are equal and in opposition. In a balanced delta connection, again no triplens are present between lines: the delta forms a closed circuit to triplen circulating currents, the impedance drop of which absorbs the harmonic e.m.f.s.

3.2.12 Symmetrical components

Figure 3.21(a) shows the sine waves and phasors of a balanced symmetric three-phase system of e.m.f.s of sequence ABC. The magnitudes are equal and the phase displacements are $2\pi/3$ rad. In *Figure 3.21(b)*, the asymmetric sine waveforms have also the sequence ABC, but they are of different magnitudes and have the phase displacements α, β and γ. Problems of asymmetry occur in the unbalanced loading of a.c. machines and in fault conditions on power networks. While a solution is possible by the Kirchhoff laws, the method of *symmetrical components* greatly simplifies analysis.

Any set of asymmetric three-phase e.m.f.s or currents can be resolved into a summation of three sets of symmetrical components, respectively of positive phase-sequence (p.p.s.) ABC, negative phase-sequence (n.p.s.) ACB, and zero phase-sequence (z.p.s.). Use is made of the operator α, resembling the 90° operator j (Section 3.2.9.1) but implying a counter-clockwise rotation of $2\pi/3$ rad (120°). Thus

$$\alpha = 1 \angle 120^\circ = \tfrac{1}{2}(-1 + \mathrm{j}\sqrt{3})$$

$$\alpha^2 = 1 \angle 240^\circ = \tfrac{1}{2}(-1 - \mathrm{j}\sqrt{3})$$

$$\alpha^3 = 1 \angle 360^\circ = 1 + \mathrm{j}0$$

$$1 + \alpha + \alpha^2 = 0$$

A symmetric three-phase system has only p.p.s. components

$$E_\mathrm{a} = E_{\mathrm{a}+}; \quad E_\mathrm{b} = \alpha^2 E_{\mathrm{a}+}; \quad E_\mathrm{c} = \alpha E_{\mathrm{a}+}$$

whereas an asymmetric system (*Figure 3.24*) comprises the three sets

$$\text{z.p.s.: } E_{\mathrm{a}0}; \quad E_{\mathrm{b}0} = E_{\mathrm{a}0}; \quad E_{\mathrm{c}0} = E_{\mathrm{a}0}$$

$$\text{p.p.s.: } E_{\mathrm{a}+}; \quad E_{\mathrm{b}+} = \alpha^2 E_{\mathrm{a}+}; \quad E_{\mathrm{c}+} = \alpha E_{\mathrm{a}+}$$

$$\text{n.p.s.: } E_{\mathrm{a}-}; \quad E_{\mathrm{b}-} = \alpha E_{\mathrm{a}-}; \quad E_{\mathrm{c}-} = \alpha^2 E_{\mathrm{a}-}$$

where the subscripts 0, + and − designate the z.p.s., p.p.s. and n.p.s. components, respectively. The p.p.s. and the n.p.s. components sum individually to zero. Therefore, if the originating phasors E_a, E_b, E_c also sum to zero there are no z.p.s. components; if they do not, their residual is the sum of the three z.p.s. components.

The asymmetrical phasors have now been reduced to the sum of three sets of symmetrical components:

$$E_\mathrm{a} = E_{\mathrm{a}0} + E_{\mathrm{a}+} + E_{\mathrm{a}-}$$

$$E_\mathrm{b} = E_{\mathrm{b}0} + E_{\mathrm{b}+} + E_{\mathrm{b}-}$$

$$E_\mathrm{c} = E_{\mathrm{c}0} + E_{\mathrm{c}+} + E_{\mathrm{c}-}$$

The components are evaluated from the arbitrary identities

$$E_\mathrm{a} = Z + P + N$$

$$E_\mathrm{b} = Z + \alpha^2 P + \alpha N$$

$$E_\mathrm{c} = Z + \alpha P + \alpha^2 N$$

where

$$Z = (E_\mathrm{a} + E_\mathrm{b} + E_\mathrm{c})/3$$

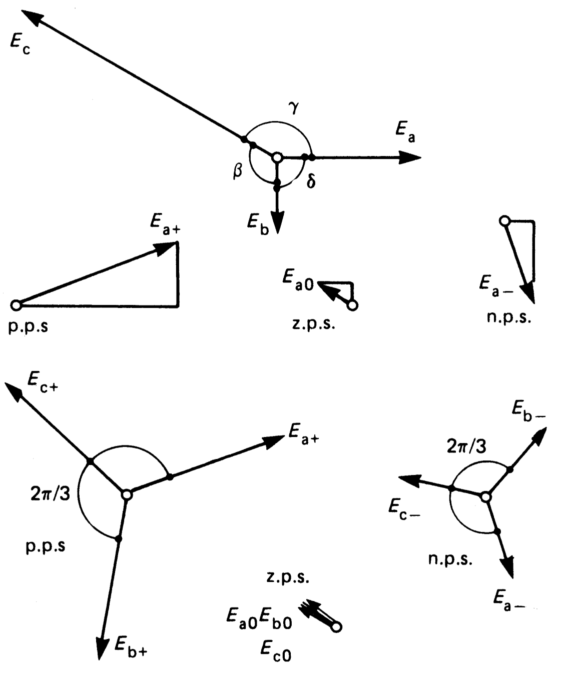


Figure 3.24 Symmetrical components

$$P = (E_\mathrm{a} + \alpha E_\mathrm{b} + \alpha^2 E_\mathrm{c})/3$$

$$N = (E_\mathrm{a} + \alpha^2 E_\mathrm{b} + \alpha E_\mathrm{c})/3$$

Figure 3.24 is drawn for an asymmetric system with voltages $E_\mathrm{a} = 200$, $E_\mathrm{b} = 100$ and $E_\mathrm{c} = 400\,\mathrm{V}$, and phase-displacement angles $\delta = 90^\circ$, $\beta = 120^\circ$ and $\gamma = 150^\circ$. In phasor terms,

$$E_\mathrm{a} = 200 \angle 0^\circ = 200 + \mathrm{j}0\,\mathrm{V}$$

$$E_\mathrm{b} = 100 \angle (-90^\circ) = 0 - \mathrm{j}100\,\mathrm{V}$$

$$E_\mathrm{c} = 400 \angle 150^\circ = -346 + \mathrm{j}200\,\mathrm{V}$$

Then

$$Z = (200 - \mathrm{j}100 - 346 + \mathrm{j}200)/3 = -49 + \mathrm{j}33 = E_{\mathrm{a}0}$$

$$P = (200 + 87 + \mathrm{j}50 + 347 + \mathrm{j}200)/3 = 211 + \mathrm{j}83 = E_{\mathrm{a}+}$$

$$N = (200 - 87 + \mathrm{j}50 - \mathrm{j}400)/3 = 38 - \mathrm{j}117 = E_{\mathrm{a}-}$$

The summation $E_{\mathrm{a}0} + E_{\mathrm{a}+} + E_{\mathrm{a}-} = 200 + \mathrm{j}0 = E_\mathrm{a}$. The p.p.s and n.p.s. components of E_b and E_c are readily obtained.

3.2.12.1 Power

In linear networks there is no interference between currents of different sequences. Thus p.p.s. voltages produce only p.p.s. currents, etc. The total power is therefore

$$P = P_\mathrm{a} + P_\mathrm{b} + P_\mathrm{c}$$
$$= 3(V_0 I_0 \cos\phi_0 + V_+ I_+ \cos\phi_+ + V_- I_- \cos\phi_-)$$

This is equivalent to the more obvious summation of phase powers

$$P = V_\mathrm{a} I_\mathrm{a} \cos\phi_\mathrm{a} + V_\mathrm{b} I_\mathrm{b} \cos\phi_\mathrm{b} + V_\mathrm{c} I_\mathrm{c} \cos\phi_\mathrm{c}$$

Symmetrical-component techniques are useful in the analysis of power-system networks with faults or unbalanced loads: an example is given in Section 3.3.4. Machine performance is also affected when the machine is supplied from an asymmetric voltage system: thus in a three-phase induction motor the n.p.s. components set up a torque in opposition to that of the (normal) p.p.s. voltages.

3.2.13 Line transmission

Networks of small physical dimensions and operated at low frequency are usually considered to have a zero propagation time; a current started in a closed circuit appears at every point in the circuit simultaneously. In extended circuits, such as long transmission lines, the propagation time is significant and cannot properly be ignored.

The basics of energy propagation on an ideal loss-free line are discussed in another section. Propagation takes place as a voltage wave v accompanied by a current wave i such that $v/i = z_0$ (the surge impedance) travelling at speed u. Both z_0 and u are functions of the line configuration, the electric and magnetic space constants ϵ_0 and μ_0, and the relative permittivity and permeability of the medium surrounding the line conductors. At the receiving end of a line of finite length, an abrupt change of the electromagnetic-field pattern (and therefore of the ratio v/i) is imposed by the discontinuity unless the receiving-end load is z_0, a termination called the *natural load* in a power line and a *matching impedance* in a telecommunication line. For a non-matching termination, wave reflection takes place with an electromagnetic wave running back towards the sending end. After many successive reflections of rapidly diminishing amplitude, the system settles down to a steady state determined by the sending-end voltage, the receiving-end load impedance and the line parameters.

3.2.13.1 A.c. power transmission

The steady-state condition considered is the transfer of a constant balanced apparent power per phase from a generator at the sending end (s) to a load at the receiving end (r) by a sinusoidal voltage and current at a frequency $f = \omega/2\pi$. The line has uniformly distributed parameters: a conductor resistance r and a loop inductance L effectively in series, and an insulation conductance g and capacitance C in shunt, all per phase and per unit length. The series impedance, shunt admittance and propagation coefficient per unit length are $z = r + j\omega L$, $y = g + j\omega C$ and $\gamma = \sqrt{(yz)}$, respectively. For a line of length l the overall parameters are $zl = Z$, $yl = Y$ and $l\sqrt{(yz)} = \sqrt{(YZ)} = \gamma l$. The solution for the receiving-end terminal conditions is in terms of $\sqrt{(YZ)}$ and its hyperbolic functions as a two-port:

$$V_s = V_r\boldsymbol{A} + I_r\boldsymbol{B} = V_r\cosh(\sqrt{YZ}) + I_r z_0 \sinh(\sqrt{YZ}) \Leftarrow$$

$$I_s = V_r\boldsymbol{C} + I_r\boldsymbol{D} = V_r(1/z_0)\sinh(\sqrt{YZ}) + I_r\cosh(\sqrt{YZ}) \Leftarrow$$

Using the hyperbolic series (Section 1.2.2) and writing $z_0 = \sqrt{(Z/Y)}$, we obtain for a symmetrical line

$$\boldsymbol{A} = 1 + YZ/2 + (YZ)^2/24 + \cdots = \boldsymbol{D}$$

$$\boldsymbol{B} = Z[1 + YZ/6 + (YZ)^2/120 + \cdots] \Leftarrow$$

$$\boldsymbol{C} = Y[1 + YZ/6 + (YZ)^2/120 + \cdots] \Leftarrow$$

The significance of the higher powers of YZ depends on: (i) the line configuration, (ii) the properties of the ambient medium, and (iii) the physical length of the line in terms of the wavelength $\lambda = u/f$. For air-insulated overhead lines the inductance is large and the capacitance small: the propagation velocity approximates to $u = 3\times10^5$ km/s (corresponding to a wavelength $\lambda = 6000$ km at 50 Hz), with a natural load z_0 of the order of 400–500 Ω. Cable lines have a low inductance and a large capacitance: the permittivity of the dielectric material and the presence of armouring and sheathing result in a propagation velocity around 200 km/s, a surge impedance below 100 Ω, and the possibility (in high-voltage cables) that the charging current may be comparable with the load current if the cable length exceeds 25–30 km.

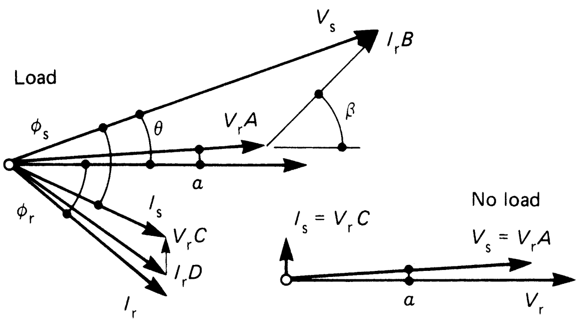


Figure 3.25 Transmission-line phasor diagram

For balanced three-phase power transmission, the general equations are applied for the line-to-neutral voltage, line current and phase power factor. Phasor diagrams for the load and no-load ($I_r = 0$) receiving-end conditions for an overhead-line transmission are shown in *Figure 3.25*, with V_r as datum. On no load, $V_s = V_r\boldsymbol{A}$, and as $\boldsymbol{A}$ has a magnitude less than unity and a small positive angle α, the phasor $V_r\boldsymbol{A}$ is smaller than V_r and leads it by angle α: thus $V_r > V_s$, the *Ferranti effect*. For the loaded condition, $I_r\boldsymbol{B}$ is added to $V_r\boldsymbol{A}$ to give V_s. Similarly $V_r\boldsymbol{C}$ is added to $I_r\boldsymbol{D}$ to obtain I_s.

The product $V_r I_r = I_r(V_s - V_r\mathrm{A})$ is the receiving-end complex apparent power S_r. Let V_s lead V_r by angle θ; then the receiving-end load has the active and reactive powers P_r and Q_r given by

$$P_r = (V_s V_r/\boldsymbol{B})\cos(\theta - \beta) - (V_r^2\boldsymbol{A}/\boldsymbol{B})\cos(\beta - \alpha) \Leftarrow$$

$$Q_r(V_s V_r/\boldsymbol{B})\sin(\theta - \beta) + (V_r^2\boldsymbol{A}/\boldsymbol{B})\sin(\beta - \alpha) \Leftarrow$$

where α and β are the angles in the complexors $\boldsymbol{A}$ and $\boldsymbol{B}$. The importance of $\boldsymbol{B}$ (roughly the overall series impedance) is clear.

Line chart Operating charts for a transmission circuit can be drawn to relate graphically V_s, V_r, P_r and Q_r, using the appropriate overall ***ABCD*** parameters (e.g. with terminal transformers included).

Receiving-end chart A receiving-end chart gives active and reactive power at the receiving end for V_r constant (*Figure* 3.26(*a*)). The co-ordinates (x, y) and the radius (r) of the constant-voltage circles are

$$x = -V_r^2(\boldsymbol{A}/\boldsymbol{B})\cos(\beta - \alpha) \Leftarrow$$

$$y = -V_r^2(\boldsymbol{A}/\boldsymbol{B})\sin(\beta - \alpha) \Leftarrow$$

$$r = V_s V_r/\boldsymbol{B}$$

where $\boldsymbol{A}$ and $\boldsymbol{B}$ are scalar magnitudes, and α and β the angles in $\boldsymbol{A}$ and $\boldsymbol{B}$. For a given V_r the chart comprises a family of concentric circles, each corresponding to a particular V_s. If a given receiving-end load is located by its active and reactive power components, V_s is obtained from the corresponding V_s circle.

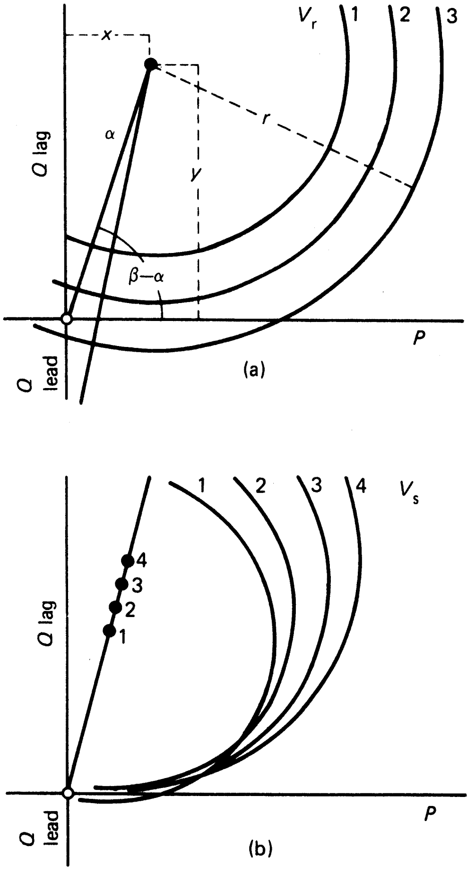


Figure 3.26 Line charts

Sending-end chart For a given V_s the sending-end chart comprises a family of circles as shown in *Figure 3.26(b)*, each circle corresponding to a particular V_r. Load points outside the envelope of these circles cannot be supplied at the V_s for which the chart is drawn.

3.2.13.2 Short line

For an overhead interconnector line the capacitive shunt admittance is neglected, reducing the general parameters to $\boldsymbol{A}=\boldsymbol{D}=1$, $\boldsymbol{B}=Z=\boldsymbol{R}+\mathrm{j}X$ and $\boldsymbol{C}=0$. The operating conditions are those in *Figure 3.27*, with a receiving-end voltage V_r (taken as reference phasor), a sending-end voltage V_s and a load current I at a lagging phase angle $\phi$with respect to V_r and having active and reactive components respectively I_p and I_q. Then

$$\begin{aligned} V_s &= V_r + (I_p - \mathrm{j}I_q)(R+\mathrm{j}X) \Leftarrow \\ &= V_r + (I_pR + I_qX) + \mathrm{j}(I_pX - I_qR) = V_r + v + \mathrm{j}u \end{aligned}$$

To a close approximation, v is the difference of the voltages V_s and V_r, while u determines their phase difference (or transmission angle). The regulation and angle are therefore v/V_s p.u. and $\theta = \arctan(u/V_s)$ rad.

Suppose that $V_r = V_s$; then $v=0$ giving $I_q = -I_p(R/X)$, and $u = I_pX[1+(R/X)^2]$ giving $\theta = \arctan(I_p/V_s)[1+(R/X)^2]$. The consequences are that (i) for a receiving-end active power P the load must be able to absorb a leading reactive power $Q = P(R/X)$, and (ii) the transmission angle is determined largely by X. If $R/X = 0.5$, typical of an overhead line, then $Q = 0.5P$ and $\theta = \arctan[1.25\ X(I_p/V_s)]$. With interconnector cables the R/X ratio is usually greater than unity and shunt capacitance current is no longer negligible.

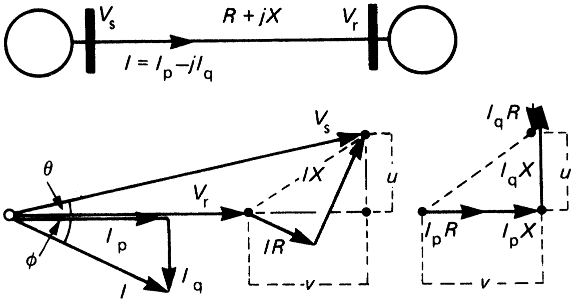


Figure 3.27 Operating conditions for a short transmission line

Independent adjustment of V_s and V_r is not feasible, and effective load control requires adjustment of v (e.g. by transformer taps) and of u (e.g. by quadrature boosting).

3.2.14 Network transients

Energy cannot be instantaneously converted from one form to another, although the time needed for conversion can be very short and the conversion rate (i.e. the power) high. Change between states occurs in a period of *transience* during which the system energies are redistributed in accordance with the energy-conservation principle (Section 1.3.1). For example, in a simple series circuit of resistance R, inductance L and capacitance C connected to a source of instantaneous voltage v, the corresponding rates of energy input, dissipation (in R) and storage (in L and C) are related by

$$p = vi = Rii + [L(\mathrm{d}i/\mathrm{d}t)i + (q/C)i] \Leftarrow$$

Dividing by the common current i and writing the capacitor charge q as the time-integral of the current gives the voltage equation

$$v = Ri + L(\mathrm{d}i/\mathrm{d}t) + (1/C)\int i\ \mathrm{d}t$$

and any changes in the parameters or in the applied voltage demand changes in the distribution of the circuit energy. The integro-differential equation can be solved to yield both steady-state and transient conditions.

In practical circuits the system may be too complex for such a direct solution; the following methods may then be attempted:

(1) formal mathematics for simple cases with linear parameters;
(2) simplification, e.g. by linearising parameters or by neglecting second-order terms;
(3) writing a possible solution based on the known physical behaviour of the system, with a check by differentiation;
(4) setting up a model system on an analogue computer; or
(5) programming a digital computer to give a solution by iteration.

3.2.14.1 Classification

Where the system has only one energy-storage component, *single-energy* transients occur. Where two (or more) different storages are concerned, the transient has a *double-* (or *multiple-*) *energy* form. Transients may occur in the following circumstances.

1. *Initiation*—a system, initially dead, is energised.
2. *Subsidence*—an initially energised system is reduced to a zero-energy condition.
3. *Transition*—a change from one state to another, where both states are energetic.
4. *Complex*—the superposition of more than one disturbance.
5. *Relaxation*—transition between states that, when reached, are themselves unstable.

Further distinctions can be made, e.g. between linear and non-linear parameters, neglect or otherwise of propagation time within the system, etc. Attention here is mainly confined to simple electric networks with constant parameters and, by analogy (Section 1.3.1), to corresponding mechanical systems.

3.2.14.2 Transient forms

During transience, the current i for an impressed voltage stimulus $v(t)$ is considered to be the superposition of a transient component i_t and a final steady-state current i_s, so that at any instant $i = i_s + i_t$. Alternatively, the voltage v for an impressed current stimulus $i(t)$ is the summation $v = v_s + v_t$. The quantities i_s or v_s are readily derived by applying the appropriate steady-state technique. The *form* of i_t or v_t is characteristic of the system itself, is independent of the stimulus and comprises exponential terms $k\exp(\lambda t)$ where k depends on the boundary conditions. This is the case because of the fixed proportionality between the stored energy $\frac{1}{2}Li^2$ and the rate of energy dissipation Ri^2 in an RL circuit; and similarly for $\frac{1}{2}Cv^2$ and v^2/R in an RC circuit. Hence the transient form can be obtained from a case in which the final steady state is of zero energy, i.e. a subsidence transient.

The subsidence transient in a *single-energy* (first-order) system having the general equation $dy/dt + ay = 0$ can be found by substituting λ for d/dt to give $\lambda y + ay = 0$, whence $\lambda = -a$. Then the solution is

$$y = k\exp(\lambda t) = k\exp(-at)$$

a simple exponential decay as in *Figure 1.2* of Section 1.2.2. For a *double-energy* (second-order) system the basic equation is

$$d^2y/dt^2 + a(dy/dt) + by = 0 \quad \text{or} \quad \lambda^2 + a\lambda + b = 0$$

The quadratic in λ has two roots, λ_1 and λ_2, and the solution has a pair of exponential terms that depend on the relation between a and b. For a *multiple-energy* (nth-order) system there will be n roots. From Section 1.2.2 it will be seen that exponential terms can represent oscillatory as well as decay forms of response.

Single-energy system Consider the RL circuit shown in *Figure 3.28*, subsequent to closure of the switch at $t = 0$. The transient current form is obtained from $L(di/dt) + Ri = 0$, or $L\lambda i + Ri = 0$, giving $\lambda = -R/L$. Then

$$i_t = k\exp[-t(R/L)] = k\exp(-t/T)$$

where $T = L/R$ is the *time-constant*. The final steady-state current depends on the source voltage v. In *Figure 3.28(a)*, with $v = V$, a constant direct voltage, $i_s = V/R$. Immediately after switching, with $t = 0+$, the current i is still zero because the inductance prevents any instantaneous rise. Hence

$$i = i_s + i_t = V/R + k\exp(-0) = V/R + k$$

so that $k = -(V/R)$. From $t = 0$ the current is, therefore,

$$i = i_s + i_t = (V/R)[1 - \exp(-t/T)]$$

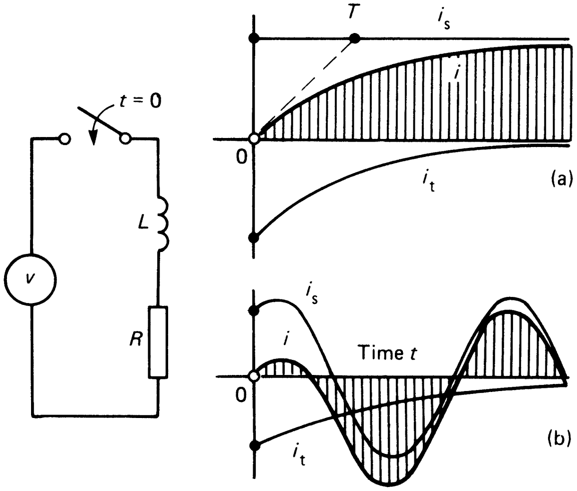


Figure 3.28 Transients in an inductive–resistive circuit

The two terms and their summation are shown in *Figure 3.28(a)*.

If, as in *Figure 3.28(b)*, the source voltage is sinusoidal expressed by $v = v_m\sin(\omega t - \alpha)$ and again switching occurs at $t = 0$, the form of the transient current is unchanged, but the final steady-state current is

$$i_s = (v_m/Z)\sin(\omega t - \alpha - \phi)$$

where $Z = \sqrt{(R^2 + \omega^2 L^2)}$ and $\phi = \arctan(\omega L/R)$. At $t = 0+$

$$i = i_s + i_t = (v_m/Z)\sin(-\alpha - \phi) + k = 0$$

which gives $k = -(v_m/A)\sin(-\alpha - \phi)$. The final steady-state and transient current components are shown in *Figure 3.28(b)* with their resultant. Initially the current is asymmetric, but subsequently the decay of i_t allows the current to approach the steady-state condition.

If $\omega L \gg R$, then approximately $\phi = \frac{1}{2}\pi$. Let the switch be closed at $v = 0$ for which $\alpha = 0$. Then the current is

$$i = (v_m/\omega L)[\sin(\omega t - \tfrac{1}{2}\pi) + 1]$$

which raises i to twice the normal steady-state peak when t reaches a half-period: this is the *doubling effect*. However, if the switch is closed at a source-voltage maximum, the current assumes its steady-state value immediately, with no transient component.

Summary for an RL circuit The transient current has a decaying exponential form, with a value of k such that, when it is added to the final steady-state current, the initial current flowing in the circuit at $t = 0$ is obtained. (In both of the cases in *Figure 3.28* the initial current is zero.) Thus if the initial circuit current is 10 A and the final current is 25 A, the value of k is −15 A.

For the CR circuit in *Figure 3.29*, the form of the transient is found from $Ri + q/C = 0$; differentiating, we obtain

$$R(di/dt) + (1/C)i = 0 \quad \text{or} \quad R\lambda + 1/C = 0$$

from which $\lambda = -1/CR = -1/T$, where $T = CR$ is the *time-constant*. Thus $i_t = k\exp(-t/T)$. With the capacitor initially uncharged and a source direct voltage V switched on at $t = 0$,

$$i = i_s + i_t = 0 + k\exp(-t/T)$$

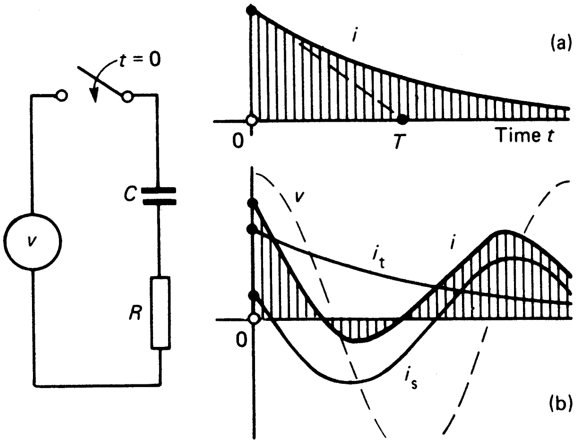


Figure 3.29 Transients in a capacitive–resistive circuit

As this must be V/R at $t=0+$, then $k=V/R$, as shown in *Figure 3.29(a)*. In *Figure 3.29(b)* the initiation of a CR circuit with a sine voltage is shown.

Summary for an RC circuit The transient current is a decaying exponential $k\ \exp(-t/T)$. The initial current is determined by the voltage difference between the voltage applied by the source and that of the capacitor. (In *Figure 3.29* the capacitor is in each case uncharged.) If this p.d. is V_0, then the initial current is V_0/R.

Double-energy system A typical case is that of a series RLC circuit. The transient form is obtained from $L(\mathrm{d}i/\mathrm{d}t) + Ri + q/C = 0$, differentiated to

$$\mathrm{d}^2i/\mathrm{d}t^2 + (R/L)(\mathrm{d}i/\mathrm{d}t) + (1/LC)i = 0$$

Thus $\lambda^2 + (R/L)\lambda + 1/LC = 0$ is the required equation, with the roots

$$\lambda_1, \lambda_2 = -\frac{R}{2L} \pm \left(\frac{R^2}{4L^2} - \frac{1}{LC}\right)^{1/2}$$

The resulting transient depends on the sign of the quantity in parentheses, i.e. on whether $R/2L$ is greater or less than $1/\sqrt{(LC)}$. Four waveforms are shown in *Figure 3.30*.

(1) *Roots real*: $R > 2\sqrt{(L/C)}$. The transient current is unidirectional and results from two simple exponential curves with different rates of decay.

(2) *Roots equal*: $R = 2\sqrt{(L/C)}$. This has more mathematical than physical interest, but it marks the boundary between unidirectional and oscillatory transient current.

(3) *Roots complex*: $R < 2\sqrt{(L/C)}$. The roots take the form $-\alpha \pm \mathrm{j}\omega_\mathrm{n}$, and the transient current oscillates with the interchange of magnetic and electric energies respectively in L and C; but the oscillation amplitude decays by reason of dissipation in R. With $R=0$ the oscillation persists without decay at the undamped natural frequency $\omega_\mathrm{n} = 1/\sqrt{(LC)}$.

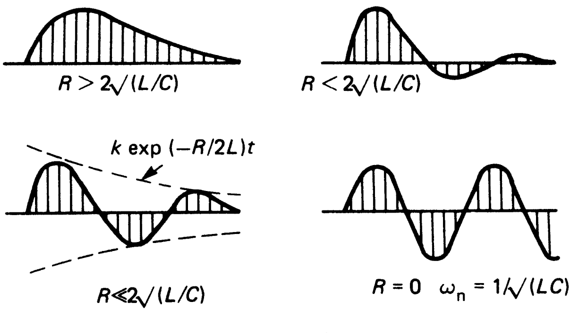


Figure 3.30 Double-energy transient forms

Pulse drive The response of networks to single pulses (or to trains of such pulses) is an important aspect of data transmission. An ideal pulse has a rectangular waveform of duration ('width') t_p. It can be considered as the resultant of two opposing step functions displaced in time by t_p as in *Figure 3.31(a)*.

In practice a pulse cannot rise and fall instantaneously, and often the amplitude is not constant (*Figure 3.31(b)*). Ambiguity in the precise position of the peak value V_p makes it necessary to define the *rise time* as the interval between the levels $0.1\ V_\mathrm{p}$ and $0.9\ V_\mathrm{p}$. The *tilt* is the difference between V_p and the value at the start of the trailing edge, expressed as a fraction of V_p.

The response of the output network to a voltage pulse depends on the network characteristics (in particular its time-constant T) and the pulse width t_p. Consider an ideal input

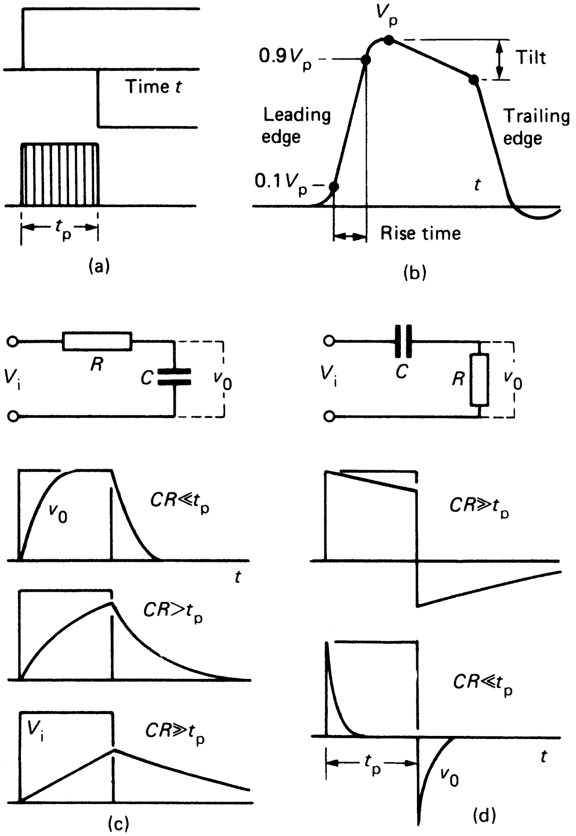


Figure 3.31 Pulse drive

voltage V_i of rectangular waveform applied to an ideal low-pass series network (*Figure 3.31*(*c*)), the output being the voltage v_0 across the capacitor C. Writing p for d/dt, then

$$\frac{v_0}{V_i}=\frac{1/\mathrm{p}C}{R+1/\mathrm{p}C}=\frac{1}{1+\mathrm{p}CR}=\frac{1}{1+\mathrm{p}T}$$

where $T=CR$ is the network time-constant. This represents an exponential growth $v_0=V_i[1-\exp(-t/T)]$ over the interval t_p. The trailing edge is an exponential decay, with t reckoned from the start of the trailing edge. Three typical responses are shown. For $CR \ll t_p$ the output voltage reaches V_i; for $CR > t_p$ the rise is slow and does not reach V_i; for $CR \gg t_p$ the rise is almost linear, the final value is small and the response is a measure of the time-integral of V_i.

With C and R interchanged as in *Figure 3.31*(*d*) to give a high-pass network, the whole of V_i appears across R at the leading edge, falling as C charges. Following the input pulse there is a reversed v_0 during the discharge of the capacitor. The output/input voltage relation is given by

$$\frac{v_0}{V_i}=\frac{R}{R+1/\mathrm{p}C}=\frac{\mathrm{p}CR}{1+\mathrm{p}CR}=\frac{\mathrm{p}T}{1+\mathrm{p}T}$$

For $CR \gg t_p$ the response shows a tilt; for $CR \ll t_p$ the capacitor charges rapidly and the output v_0 comprises positive- and negative-going spikes that give a measure of the time-differential of V_i.

3.2.14.3 Laplace transform method

Application of the Laplace transforms is the most usual method of solving transient problems. The basic features of the Laplace transform are set out in Section 1.2.7 and *Table 3.4*, which gives transform pairs. The advantages of the method are that: (1) any stimulus, including discontinuous and pulse forms, can be handled, (2) the solution is complete with both steady-state and transient components, (3) the initial conditions are introduced at the start, and (4) formal mathematical processes are avoided.

Consider the system in *Figure 3.28*(*a*). The applied direct voltage V has the Laplace transform $V(s)=V/s$; the operational impedance of the circuit is $Z(s)=R+Ls$. Then the Laplace transform of the current is

$$I(s)=\frac{V(s)}{Z(s)}=\frac{V}{s}\,\frac{1}{R+Ls}=\frac{V}{L}\,\frac{1}{s(s+R/L)}$$

The term V/L is a constant unaffected by transformation. The term in s is almost of the form $a/s(s+a)$. So, if we write

$$I(s)=\frac{V}{aL}\,\frac{a}{s(s+a)}$$

where $a=R/L=1/T$, the inverse Laplace transform gives

$$i(t)=(V/aL)[1-\exp(-at)]=(V/R)[1-\exp(-t/T)]$$

which is the complete solution. More complex problems require the development of partial fractions to derive recognisable transforms which are then individually inverse-transformed to give the terms in the solution of $i(t)$.

3.2.15 System functions

It is characteristic of linear constant-coefficient systems that their operational solution involves three parts: (i) the excitation or stimulus, (ii) the output or response and (iii) the system function. Thus in the relation $I(s)=V(s)/Z(s)$ for the current in Z resulting from the application of V, $1/Z(s)$ is the *system function* relating voltage to current. For the simple electrical system shown in *Figure 3.32*(*a*) the system function $Y(s)$ relating $V(s)$ to $I(s)$ in $I(s)=V(s)Y(s)$ is $Y(s)=1/(R+Ls+1/Cs)$. Different functions could relate the capacitor charge or the magnetic linkage in the inductor to the transform $V(s)$ of the stimulus $v(t)$.

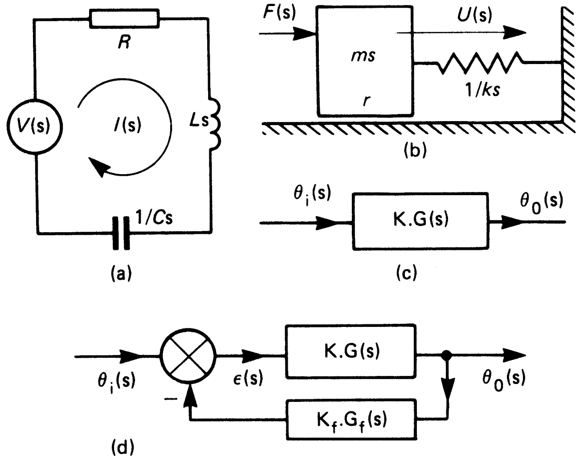


Figure 3.32 System functions

The mechanical analogue (*Figure 3.32*(*b*)) of this electrical system, as indicated in Section 1.3.1, has a system transfer function to relate force $f(t)$ to velocity $u(t)$ of the mass m and one end of the spring of compliance k in the presence of viscous friction of coefficient r. Then $F(s)$ and $U(s)$ are the transforms of $f(t)$ and $u(t)$, and the operational 'mechanical impedance' has the terms ms, $1/ks$ and r. In general, an input $\theta_i(s)$ and an output $\theta_o(s)$ are related by a system transfer function $KG(s)$ (*Figure 3.32*(*c*)), where K is a numerical or a dimensional quantity to include amplification or the value of some physical quantity (such as admittance). The transform of the integro-differential equation of variation with time is expressed by the term $G(s)$. The system is then represented by the block diagram in *Figure 3.32*(*c*); i.e. $\theta_o(s)/\theta_i(s)=KG(s)$.

A number of typical system transfer functions for relatively simple systems are given in *Table 3.4*.

The output of one system may be used as the input to another. Provided that the two do not interact (i.e. the individual transfer functions are not modified by the connection) the overall system function is the product $[K_1G_1(s)]\times[K_2G_2(s)]$ of the individual functions. If the systems are paralleled and their outputs are additively combined, the overall function is their sum.

3.2.15.1 Closed-loop systems

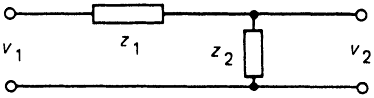
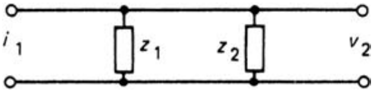
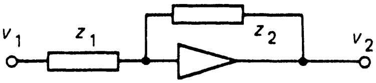
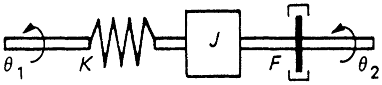
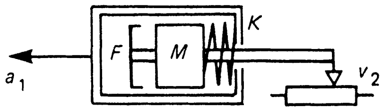
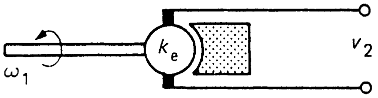
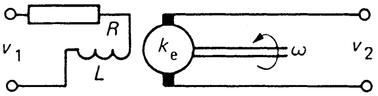
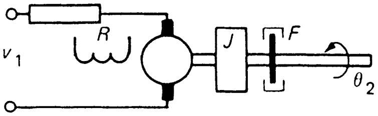
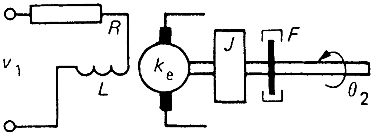
In *Figure 3.32*, parts (*a*), (*b*) and (*c*) are *open-loop* systems. However, the output can be made to modify the input by feedback through a network $K_fG_f(s)$ as in (*d*). The signal

$$\theta_f(s)=[K_fG_f(s)]\theta_o(s)$$

is combined with $\theta_i(s)$ to give the modified input.

For *positive feedback*, the resultant input is $\sigma(s)=\theta_i(s)+\theta_f(s)$, and the effect is usually to produce instability and oscillation.

Table 3.4 System transfer functions [the relation $f_2(t)/f_1(t)$ of output to input quantity in terms of the Laplace transform $F_2(s)/F_1(s)$]

System		*Transfer function*	
1 Electrical network		$\frac{V_2(s)}{V_1(s)} = \frac{Z_2(s)}{Z_1(s) + Z_2(s)}$	
2 Electrical network		$\frac{V_2(s)}{I_1(s)} = \frac{Z_1(s)Z_2(s)}{Z_1(s) + Z_2(s)}$	
3 Feedback amplifier		$\frac{V_2(s)}{V_1(s)} = \frac{Z_2(s)}{Z_1(s)}$	
4 Second-order system		$\frac{\theta_2(s)}{\theta_1(s)} = \frac{1}{1 + 2csT + s^2T^2}$	$T = \sqrt{(J/K)}$ $c = F/2\sqrt{(JK)}$
5 Accelerometer		$\frac{V_2(s)}{A_1(s)} = \frac{k_a}{1 + 2csT + s^2T^2}$	$T = \sqrt{(M/K)}$ $c = F/2\sqrt{(MK)}$
6 Permanent-magnet generator		$\frac{V_2(s)}{\omega_1(s)} = k_e$	
7 Separately excited generator		$\frac{V_2(s)}{V_1(s)} = \frac{k_e\omega}{R(1 + sT)}$	$T = L/R$
8 Motor: armature control		$\frac{\theta_2(s)}{V_1(s)} = \frac{K_e}{s(1 + sT)}$	$K_e = k_e/(FR + k_e^2)$ $T = JR/(FR + k_e^2)$
9 Motor: field control		$\frac{\theta_2(s)}{V_1(s)} = \frac{k_e}{s(1 + sT_1)(1 + sT_2)}$	$T_1 = J/F$, $T_2 = L/R$

v	voltage	L	inductance	θ	angular displacement	M	mass
i	current	k_e	e.m.f. coefficient	ω	angular velocity	J	inertia
Z	impedance	c	damping coefficient	a	acceleration	F	viscous friction coefficient
R, r	resistance	T	time-constant	k_a	acceleration coefficient	K	stiffness

For *negative feedback*, the resultant input is the difference $\epsilon(s) = \theta_i(s) - \theta_f(s)$, an 'error' signal. With the main system $KG(s)$ now relating ϵ and θ_o, the output/input relation is

$$\frac{\theta_o(s)}{\theta_i(s)} = \frac{KG(s)}{1 + [KG(s)][K_fG_f(s)]}$$

Suppose that there is unity feedback $K_fG_f(s) = 1$; then if $KG(s)$ is large

$$\theta_o(s)/\theta_i(s) = KG(s)/[1 + KG(s)] \simeq 1$$

and the output closely follows the input in magnitude and wave shape, a condition sought in servo-mechanisms and feedback controls.

3.2.15.2 System performance

In general, a system function takes the form numerator/denominator, each a polynomial in s, relating response to input stimulus. Two forms are

$$KG(s) = \frac{b_m s^m + b_{m-1}s^{m-1} + \cdots + b_0}{a_n s^n + a_{n-1}s^{n-1} + \cdots + a_0} \qquad (3.1)$$

$$= \frac{b_m(s - z_1)(s - z_2)\cdots(s - z_m)}{a_n(s - p_1)(s - p_2)\cdots(s - p_n)} \qquad (3.2)$$

The response depends both on the system and on the stimulus. Performance can be studied if simple formalised stimuli (e.g. step, ramp or sinusoidal) are assumed; an exponential stimulus is even more direct because (in a linear system) the transient and steady-state responses are then both exponential. With the system function expressed in terms of the complex frequency $s = \sigma + j\omega$ it is necessary to express the stimulus in similar terms and to evaluate the response as a function of time by inverse Laplace transformation. The response in the *frequency domain* (i.e. the output/input relation for sustained sinusoidal stimuli over a frequency range) is obtained by taking $s = j\omega$ and solving the complexor $KG(j\omega)$. Another alternative is to derive the poles (p) and zeroes (z) in equation (3.2) above.

Thus there are several techniques for evaluating system functions. Some are graphical and give a concise representation of the response to specified stimuli.

3.2.15.3 Poles and zeros

In equation (3.2), the numbers z are the values of s for which $KG(s) = 0$; for, if s is set equal to z_1 or $z_2, \ldots$, the numerator has a zero term as a factor. Similarly, if s is set equal to p_1 or $p_2, \ldots$, there is a zero factor in the denominator and $KG(s)$ is infinite. Then the z terms are the *zeros* and the p terms are the *poles* of the system function. Except for the term b_m/a_n, the system function is completely specified by its poles and zeros.

Consider the network of *Figure 3.33*, the system function required being the output voltage v_o in terms of the input voltage v_i. This is the ratio of the paralleled branches R_2L_2C to the whole impedance across the input terminals. Algebra gives

$$KG(s) = 8(s + 1)/(s^3 + 3s^2 + 14s + 16)$$

$$= 8\frac{s + 1}{(s + 1.36)[s + (0.82 + j3.33)][s + (0.82 - j3.33)]}$$

by factorising numerator and denominator. Thus there is one zero for $s = -1$. There are three poles, with $s = -1.36$, and $-0.82 \pm j3.33$. These are plotted on the complex s-plane in *Figure 3.33*. Poles on the real axis σ correspond to simple exponential variations with time, decaying for negative and increasing indefinitely for positive values. Poles in conjugate pairs on the $j\omega$ axis correspond to sustained sinusoidal oscillations. If the poles occur displaced from the origin and not on either axis, they refer to sinusoids with a decay or a growth factor, depending on whether the term σ is negative or positive.

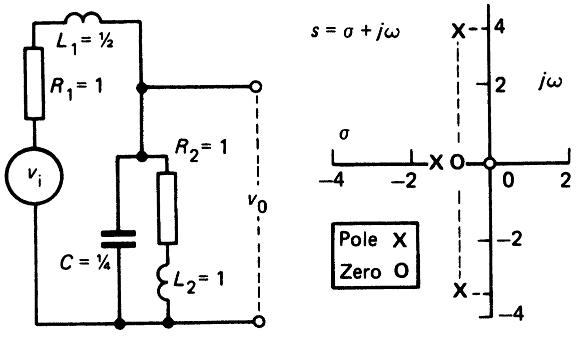


Figure 3.33 Poles and zeros

3.2.15.4 Harmonic response

This is the steady-state response to a sinusoidal input at angular frequency ω. When a sine signal input is applied to a linear system, the steady-state response is also sinusoidal and is related to the input by a relative magnitude M and a phase angle α. The system function is $KG(j\omega)$.

Consider again the network of *Figure 3.33*. Writing $s = j\omega$ and simplifying gives the phasor expression for V_o/V_i as

$$KG(j\omega) = \frac{8(j\omega + 1)}{(16 - 3\omega^2) + j\omega(14 - \omega^2)} = |M| \angle \alpha$$

Plots of $|M|$ and $\angle \alpha$ are shown in *Figure 3.34(a)*. For $\omega = 0$ the network is a simple voltage divider with $V_o/V_i = 0.5$ and a phase angle $\alpha = 0$. For $\omega = \infty$, the terminal capacitor effectively

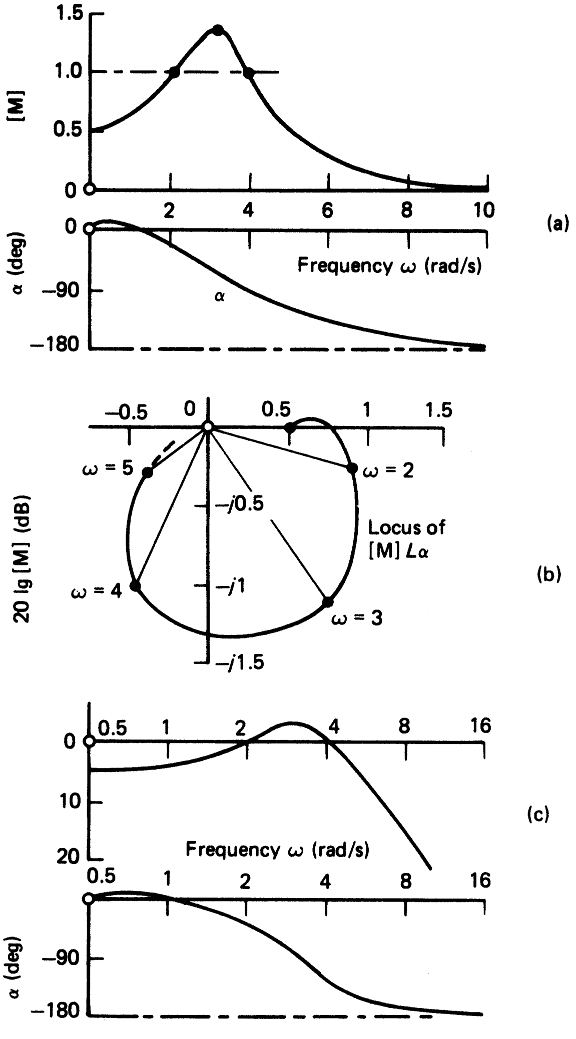


Figure 3.34 Harmonic response

short circuits the output terminals so that $V_o/V_i = 0$. At intermediate frequencies the gain $|M|$ rises to a peak at $\omega = 3.3$ rad/s and thereafter falls toward zero. The phase angle α is small and positive below $\omega = 1$, being always negative thereafter, to become $-180°$ at infinite frequency.

Nyquist diagram The Nyquist diagram is a polar plot of $|M| \angle \alpha$ over the frequency range (*Figure 3.34(b)*), for an input $V_i = 1 + j0$. The plot is particularly useful for feedback systems. If the open-loop transfer function is plotted, and in the direction of increasing ω it encloses the point $(-1 + j0)$, then when the loop is closed the system will be unstable as the output is more than enough to supply a feedback input even when $V_i = 0$. The Nyquist criterion for stability is therefore that the point $(-1 + j0)$ shall not be enclosed by the plot.

Bode diagram The Bode diagram for the system shown in *Figure 3.33* is *Figure 3.34(a)* redrawn with logarithmic ordinates of $|M|$ and a logarithmic scale of ω. Normally the ordinates are expressed as a *gain* $20 \log |M|$ in decibels. For the example being considered, $M = 0.5$ for very low frequencies, so that $20 \log |M| = -6$ dB; for $\omega = 3.3$ the amplitude of M is 1.4 and the corresponding gain is $+2.9$ dB; and at the two frequencies when the output and input magnitudes are the same, $M = 1$ and $20 \log (1) = 0$ dB. All these are shown in the Bode diagram (*Figure 3.34(c)*). On the logarithmic frequency scale, equal ratios of ω are separated by equal distances along the horizontal axis. If successive values 0.5, 1, 2, 4, ..., are marked in equidistantly, their successive ratios 1/0.5, 2/1, ..., are all equal to 2, so that each interval is a *frequency octave*. Correspondingly the equispaced frequencies 0.1, 1, 10, ..., express a *frequency decade*.

The phase-angle plot is drawn in degrees to the same logarithmic scale of frequency.

An advantage of the Bode plot is the ease with which system functions can be built up term by term. The product of complex operators is reduced to the addition of the logarithms of their moduli and phase angles; similarly the quotient is reduced to subtraction. If the system function can adequately be expressed in *simple* terms, the Bode diagram can be rapidly assembled. Such terms are listed below.

(1) $j\omega$: represented by a line through $\omega = 1$ and rising with frequency at 6 dB per octave or 20 dB per decade, and with a constant phase angle $\alpha = 90°$
(2) $1/j\omega$: as for $j\omega$, but falling with frequency, and with $\alpha = -90°$
(3) $1 + j\omega T$: a straight line of zero gain for frequencies up to that for which $\omega T = 1$, and thereafter a second straight line rising at 6 dB per octave; the change of direction occurs at the *break point* (*Figure 3.35(a)*).
(4) $1/(1 + j\omega T)$: as for $1 + j\omega T$, except that after the break point the gain drops with frequency at 6 dB per octave.

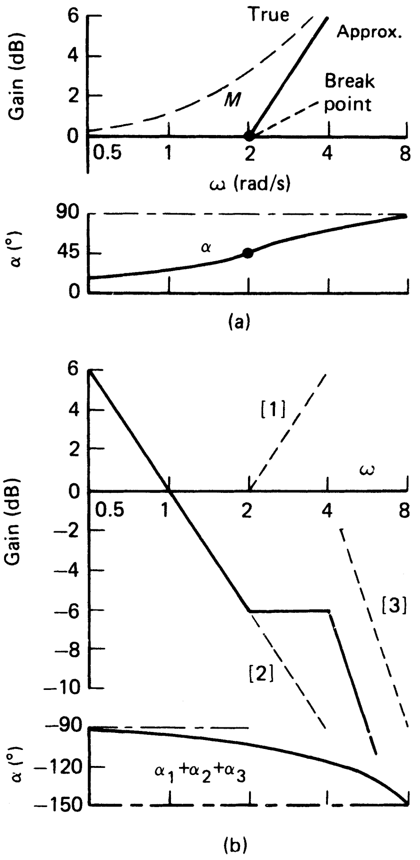


Figure 3.35 Bode diagrams

In *Figure 3.35(a)* the true gain shown by the broken curve is approximated by the two straight lines meeting at the break point. The approximate and true gains, and the phase angles, are given in *Table 3.5* for the term $1 + j\omega T$. The error in the gain is 3 dB at the break point, and 1 dB at one-half and twice the break-point frequency, making correction very simple.

The uncorrected Bode plot for the system function

$$KG(j\omega) = K\frac{(1 + j\omega 0.5)}{j\omega(1 + j\omega 0.25)^2} = K\frac{[1]}{[2][3]}$$

is shown in *Figure 3.35(b)*. Term [1] is the same as in *Figure 3.35(a)*. Term [2] is a straight line running downward

Table 3.5 Gain and phase angle for $1+j\omega T$

ωT	*Gain (dB)*		*Angle (°)*	ωT	*Gain (dB)*		*Angle (°)*
	Approx.	*True*			*Approx.*	*True*	
0	0	0.0	0	2	6	7.0	63.5
0.01	0	0.00	0.5	4	12	12.3	76
0.1	0	0.04	5.7	8	18	18.1	83
0.25	0	0.26	14	10	20	20.0	84
0.5	0	1.0	26.5	16	24	24.0	86.5
1.0	0	3.0	45	100	40	40.0	89.5

through $\omega = 1$ with a slope of 6 dB per octave. Term [3] has a break point at $\omega = 4$, but as it is a squared term its slope for $\omega > 4$ is 12 dB per octave. The full-time plot of gain is obtained by direct superposition. The effect of the constant K is to lift the whole plot upward by 20 log (K). The summed phase angles approach $-90°$ at zero frequency and $-180°$ at infinite frequency.

Nichols diagram The Nichols diagram resembles the Nyquist diagram in construction, but instead of phasor values the magnitudes are the log moduli. The point $(-1 + j0)$ of the Nyquist diagram becomes the point (0 dB, $\angle -180°$). The Nichols diagram is used for determining the closed-loop response of systems.

3.2.16 Non-linearity

A truly linear system, in which *effect* is in all circumstances precisely proportional to *cause*, is a rarity in nature. Yet engineering analyses are most usually based on a linear assumption because it is mathematically much simplified, permits of superposition and can sometimes yield results near enough to reality to be useful. If, however, the non-linearity is a significant property (such as magnetic saturation) or is introduced deliberately for a required effect (as in rectification), a non-linear analysis is essential. Such analyses are mathematically cumbersome. No general method exists, so that *ad hoc* techniques have been applied to deal with specific forms of non-linearity. The treatment depends on whether a steady-state or a transient condition is to be evaluated.

3.2.16.1 Techniques

Some of the techniques used are: (i) step-by-step solution, graphical or by computation; (ii) linearising over finite intervals; (iii) fitting an explicit mathematical function to the non-linear characteristic; and (iv) expressing the non-linear characteristic as a power series.

Step-by-step solution Consider, as an example, the growth of the flux in a ferromagnetic-cored inductor in which the inductance L is a function of the current i in its N turns. Given the flux magnetomotive-force (m.m.f.) characteristic, and the (constant) resistance r, the conditions for the sudden application of a constant voltage V are given by

$$V = ri + \mathrm{d}(Li)/\mathrm{d}t \simeq ri + N(\Delta\Phi/\Delta\tau)$$

which is solved in suitable steps of Δt, successive currents i being evaluated for use with the magnetic characteristic to start the next time-interval.

Linearising A non-linear characteristic may be approximated by a succession of straight lines, so that a piecemeal set of linear equations can be applied, 'matching' the conditions at each discontinuity.

For 'small-signal' perturbations about a fixed quiescent condition, the mean slope of the non-linear characteristic around the point is taken and the corresponding parameters derived therefrom. Oscillation about the quiescent point can then be handled as for a linear system.

Explicit function For the resistance material in a surge diverter the voltage–current relationship $v = ki^x$ has been employed, with x taking a value typically between 0.2 and 0.3. The resistance–temperature relationship of a thermistor, in terms of the resistance values R_1 and R_2 at corresponding absolute temperatures T_1 and T_2 takes the form

$$R_2 = R_1 \exp(k/T_2 - k/T_1)$$

Several functions, such as $y = a \sinh(bx)$, have been used as approximations to magnetic saturation excluding hysteresis. A static-friction effect, of interest at zero speed in a control system, has been expressed as $y = k(\mathrm{sgn}\ x)$, i.e. a constant that acts against the driving torque.

Series A typical form is $y = a_0 + a_1 x + a_2x^2 + \cdots + a_nx^n$, where the coefficients a are independent of x. Such a series may have a restricted range, and the powers limited to even orders if the required characteristic has the same shape for both negative and positive y. A second-degree series $y = a_0 + a_1x + a_2x^2$ can be fitted through any three points on a given function of y, and a third-degree expression through any four points. However, the prototype characteristic must not have any discontinuities.

Rational-fraction expressions have also been developed. The open-circuit voltage of a small synchronous machine in terms of the field current might take the form $v = (27 + 0.006i)/(1 + 0.03i)$. Similarly, the magnetisation curve of an electrical sheet steel might have the B–H relationship

$$H = B(426 - 760B + 440B^2)/(1 - 0.80B + 0.17B^2)$$

with hysteresis neglected. An exponential series

$$B = a[1 - \exp(-bH)] + c[1 - \exp(-dH)] + \cdots + \mu_0 H$$

has been suggested to represent the magnetisation characteristic of a machine, the final term being related to the air gap line.

Non-linear characteristics *Figure 3.36* shows some of the typical relations $y = f(x)$ that may occur in non-linear systems. Not all are analytic, and some may require step-by-step methods.

The *simple* relations shown are: (a) response depending on direction, as in rectification; (b) skew symmetry, showing the effect of saturation; and (c) negative-slope region, but with y univalued.

The *complex* relations are: (d) negative-slope region, with y multivalued; (e) build-up of system with hysteresis, unsaturated; (f) toggle characteristic, typical of idealised saturated hysteresis; and (g) backlash, with y taking any value between the characteristic limit-lines.

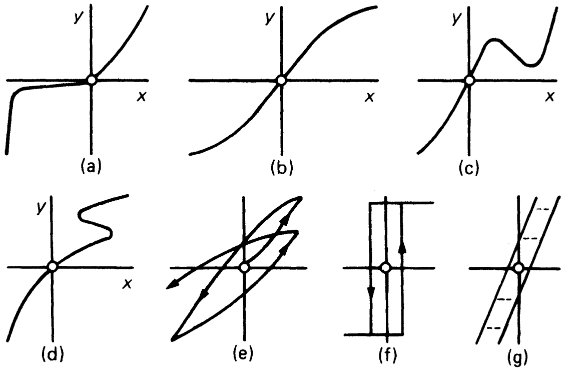


Figure 3.36 Typical non-linear characteristics

3.2.16.2 Examples

A few examples of non-linear parameters and techniques are given here to illustrate their very wide range of interest.

Resistors Thermally sensitive resistors (thermistors) may have positive or negative resistance-temperature coefficients. The latter have a relation between resistance R and absolute temperature T given by $R_2 = R_1 \exp[b(1/T_2 - 1/T_1)]$. They are made from oxides of the iron group of metals with the addition of small amounts of ions of different valency, and are applied to temperature measurement and control. Thermistors with a positive resistance-temperature coefficient made from monocrystalline barium titanate have a resistance that, for example, increases 100-fold over the range 50–100°C; they are used in the protection of machine windings against excessive temperature rise.

Voltage-sensitive resistors, made in disc form from silicon carbide, have a voltage–current relationship approximating to $v = ki^{\beta}$, where β ranges from 0.15 to 0.25. For $\beta = 0.2$ the power dissipated is proportional to v^6, the current doubling for a 12% rise in voltage.

Inductors The current in a load fed from a constant sinusoidal voltage supply can be varied over a wide range economically by use of a series inductor carrying an additional d.c.-excited winding to vary the saturation level and hence the effective inductance. The core material should have a flux-m.m.f. relationship like that in *Figure 3.36(f)*. Grain-oriented nickel and silicon irons are suitable for the inductor core. A related phenomenon accounts for the in-rush current in transformers.

Describing function In a non-linear system a sinusoidal drive does not produce a sinusoidal response. The describing-function technique is devised to obtain the *fundamental-frequency* effect of non-linearity under steady-state (but not transient) conditions.

Consider a stimulus $x = h\cos\omega t$ to give a response $y = f(t)$. As non-linearity inevitably introduces harmonic distortion, y can be expanded as a Fourier series (Section 1.2.5) to give

$$y = a_0 + a_1\cos\omega t + a_2\cos 2\omega t + \cdots$$
$$+ b_1\sin\omega t + b_2\sin 2\omega t + \cdots$$

The components $a_1\cos\omega t$ and $b_1\sin\omega t$ are regarded as the 'true' response, with a gain factor $(a_1 + jb_1)/h$, the other terms being the distortion. The gain factor is the describing function. Let $y = ax + bx^2$ with $x = h\cos\omega t$; applying the expansion gives the fundamental-frequency term $y = (a + \frac{3}{4}bh^2)h\cos\omega t$. The describing function is therefore $a + \frac{3}{4}bh^2$, which is clearly dependent on the magnitude h of the input. Thus the technique consists in evaluating the Fourier series for the output waveform for a sinusoidal input, and finding therefrom the magnitude and phase angle of the fundamental-frequency response.

Ferroresonance The individual r.m.s. current–voltage characteristics of a pure capacitor C and a ferromagnetic-cored (but loss-free) inductor L for a constant-frequency sinusoidal r.m.s. voltage V are shown in *Figure 3.37*. With L and C in series and carrying a common r.m.s. current I, the applied voltage is $V = V_L - V_C$. At low voltage V_L predominates, and the $I - V$ relationship is the line 0P, with I lagging V by 90°. At P, with $V = V_0$ and $I = I_0$, the system is at a limit of stability, for an increase in V results in a reduction in $V_L - V_C$. At a current level Q the difference is zero. The current therefore 'jumps' from I_0 to a higher level I_1 (point R), still with $V = V_0$. During the rapid rise there is an interchange of stored energy, and for $V > V_0$ the circuit is capacitive. When V is reduced from above to below V_0, a sudden current jump from I_1 to I_0 occurs. A comparable jump phenomenon takes place for a parallel connection of C and L.

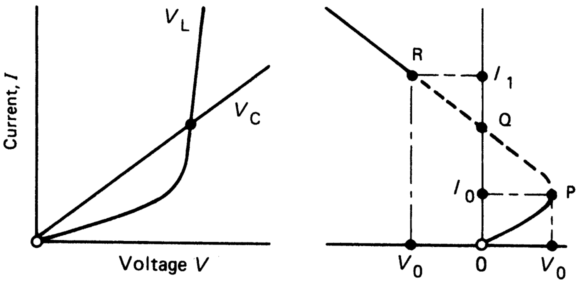


Figure 3.37 Ferroresonance

Phase-plane technique 'Phase' here means 'state' (as in the solid, liquid and vapour 'phases' of water). The phase-plane technique can be used to elucidate non-linear system behaviour graphically. *Figure 3.38* shows a circuit of series R, L and C with a drive having the voltage–current relationship $v = -ri + ai^3$. Then, with constant circuit parameters,

$$L(\mathrm{d}i/\mathrm{d}t) + (R - r + ai^2)i + q/C = 0$$

where q is the time-integral of i. The presence of L and C indicates the possibility of oscillation. The middle ('damping') term can be negative for small currents (increasing the oscillation amplitude) but positive for larger currents (reducing the amplitude). Hence the system seeks a constant amplitude irrespective of the starting condition. The $q - i$ phase-plane loci show the stable condition as related to the degree of drive non-linearity (indicated by the broken curves). With suitable scales the locus for minor non-linearity is circular, indicating near-sinusoidal oscillation; for major drive non-linearity, however, the locus shows abrupt changes and an approach to a 'relaxation' type of waveform.

Isoclines A non-linear second-order system described by

$$\mathrm{d}^2y/\mathrm{d}t^2 + \mathrm{f}(y, \mathrm{d}y/\mathrm{d}t) + \mathrm{g}(y, \mathrm{d}y/\mathrm{d}t)y = 0$$

can be represented at any point in the phase plane having the co-ordinates y and $\mathrm{d}y/\mathrm{d}t$, representing, for example, position

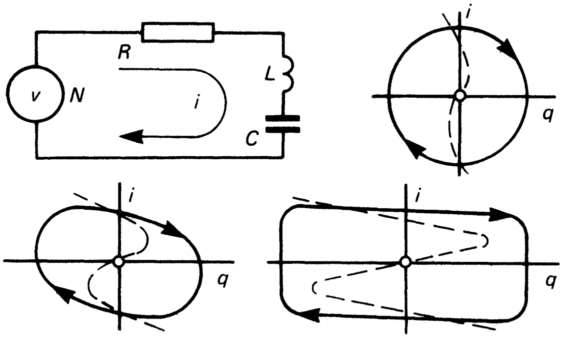


Figure 3.38 Phase-plane trajectories for an oscillatory circuit

and velocity, or charge and current. By writing $\mathrm{d}y/\mathrm{d}t = z$ and eliminating time by division, we obtain a first-order equation relating z and y:

$$\mathrm{d}z/\mathrm{d}y = -[\mathrm{f}(y, z)z + \mathrm{g}(y, z)y]/z$$

Integration gives phase-plane trajectories that everywhere satisfy this equation, starting from any initial condition. If $\mathrm{d}z/\mathrm{d}y$ cannot be directly integrated, it is possible to draw the trajectories with the aid of *isoclines*, i.e. lines along which the *slope* of the trajectory is constant. Make $\mathrm{d}z/\mathrm{d}y = m$, a constant; then $-mz = \mathrm{f}(y, z)z + \mathrm{g}(y, z)y$. Since for $z = 0$ the slope m is infinite (i.e. at right angles to the y axis) the trajectories intersect the horizontal y axis normally, except at singular points.

Consider a linear system with an undamped natural frequency $\omega_n = 1$ and a damping coefficient $c = 0.5$. For zero drive

$$\mathrm{d}^2y/\mathrm{d}t^2 + \mathrm{d}y/\mathrm{d}t + y = 0 \quad \text{or} \quad \mathrm{d}z/\mathrm{d}t = -(z + y)\Leftarrow$$

if z is written for $\mathrm{d}y/\mathrm{d}t$. Dividing the second equation by z and equating it to a constant m gives

$$\mathrm{m} = \mathrm{d}z/\mathrm{d}y = -(z + y)/z \text{ or } z/y = -1/(1 + m)\Leftarrow$$

representing a family of straight lines with the associated values

m	-4	-2	-1	0	1	2	4	$\infty\Leftarrow$
z/y	$\frac{1}{3}$	1	∞	-1	$-\frac{1}{2}$	$-\frac{1}{3}$	$-\frac{1}{5}$	0

Draw the z/y axes on the phase plane (*Figure 3.39*) marked with short lines of the appropriate slope m. Then, starting at any arbitrary point, a trajectory is drawn to cross each axis at the indicated slope. With no drive, all trajectories approach, and finally reach, the origin after oscillations in a counter-clockwise direction; for a steady drive V, the only difference is to shift the vortex to V on the y-axis. The approach to O or V represents the decaying oscillation of the system and its final steady state. Because $\mathrm{d}t = \mathrm{d}y/z$, the finite difference $\Delta t = \Delta y/z$ gives the time interval between successive points on a trajectory.

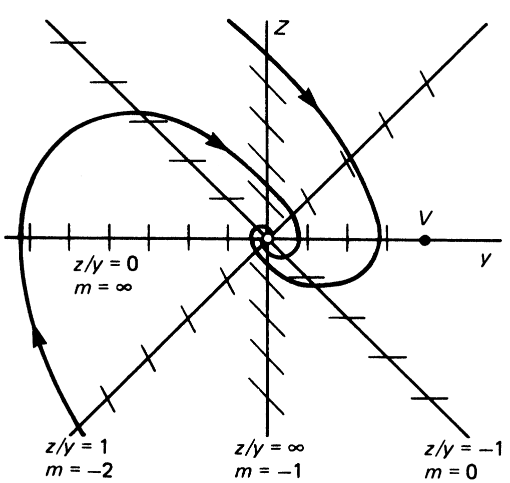


Figure 3.39 Phase-plane trajectories

3.3 Power-system network analysis

3.3.1 Conventions

Modern power-system analyses are based mainly on nodal equations scaled to a per-unit basis, with a particular convention for the sign of reactive power.

3.3.1.1 Per-unit basis

The total apparent power in a three-phase circuit ABC with phase voltages V_a, V_b, V_c, currents I_a, I_b, I_c and phase angles between the associated voltage and current phasors of θ_a, θ_b, θ_c is

$$S = (P_a + P_b + P_c) + \mathrm{j}(Q_a + Q_b + Q_c) = P + \mathrm{j}Q$$

where for phase A the active power is $P_a = V_aI_a \cos\theta_a$ and the reactive power is $Q_a = V_aI_a \sin\theta_a$. Corresponding expressions apply for phases B and C.

If the phases are balanced, all three have the same scalar voltage V and current I, and all phase angles are θ. Then $S = 3(VI \cos\theta + \mathrm{j}VI \sin\theta)$. This can be written in the form

$$\frac{S}{k} = \left(\frac{3}{a}\frac{V}{b}\frac{I}{c}\frac{\cos\theta\psi}{d}\right) + \mathrm{j}\left(\frac{3}{a}\frac{V}{b}\frac{I}{c}\frac{\sin\theta\psi}{d}\right)\Big]$$

where k, a, b, c and d are scaling factors. It is customary to choose $a = 3$ and $d = 1$, leaving c and b, one of which is assigned an independent value while the other takes a value depending on the overall scaling relation $k = abcd$.

In normal operating conditions the scalar voltage V approximates to the rated phase voltage V_R; hence b is taken as V_R so that V/b is the voltage in per unit of V_R. Defining the scaled variables as S_{pu}, V_{pu} and I_{pu}, the *total* apparent power is

$$S_{pu} = (V_{pu}I_{pu}\cos\theta) + \mathrm{j}(V_{pu}I_{pu}\sin\theta)\Leftarrow$$

which is an equation in one-phase form, justifying the use of single-line schematic diagrams to represent three-phase power circuits.

The scaling factors are termed *base* values; i.e. k is the base apparent power, b is the base phase voltage and c is the base current. These definitions imply further base values for impedance and admittance, namely $Z_{base} = b/c$ and $Y_{base} = c/b$.

If the line-to-line voltage V_l is used, then in the foregoing scaling equation $3/a$ and V/b becomes $\sqrt{3}/a'\Leftarrow$and $V_l/b'\Leftarrow$. Choosing $a' = \sqrt{3}$ and $b'\Leftarrow$as rated line voltage leaves S/k unchanged. Note, however that: (i) $\theta\psi$remains as the angle between *phase* voltage and current; and (ii) in the per-unit equation $S = VI^*$ the voltage V is the per-unit *phase* voltage, not the line-to-line voltage (although numerically both have the same per-unit value).

3.3.1.2 Reactive power convention

Reactive power may be lagging or leading. The common convention is to consider lagging reactive power flow to be positive, as calculated from the product of the voltage and current-conjugate phasors: thus $S = VI^* = P + \mathrm{j}Q$, with Q a positive number for a lagging power factor condition. For a leading power factor, Q has the same flow direction but is numerically negative. As a consequence, an inductor *absorbs*, but a capacitor *generates*, lagging reactive power, as shown in *Figure 3.40*.

Note: system engineers refer, for brevity, to the flow of 'power' and 'vars', meaning active power and reactive power, respectively.

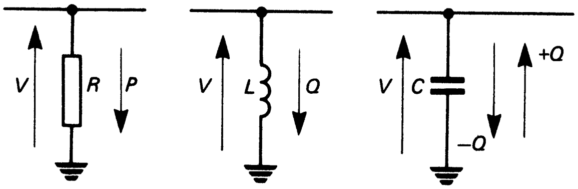


Figure 3.40 Power taken by resistive, inductive and capacitive loads

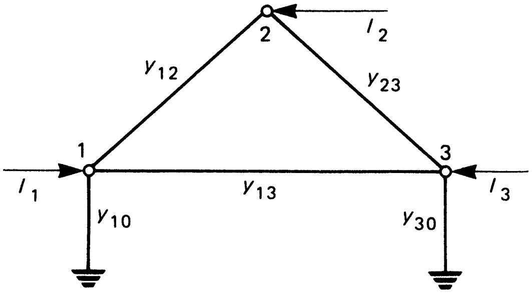


Figure 3.41 Sample network

3.3.1.3 Nodal-admittance equations

The nodal-admittance equations derive from the node-voltage method of analysis in Section 3.2.3. For the network in *Figure 3.41*, currents I_1, I_2 and I_3 are injected respectively into nodes 1, 2 and 3. The loads are linked by branches of admittance y_{12}, y_{23} and y_{31}, and to the earth or external reference node r by branches of admittance y_{10} and y_{30}. Assuming that the node voltages V_1, V_2 and V_3 are expressed with reference to r and that $V_r = 0$, then writing the Kirchhoff current equations and simplifying gives

$$(y_{10} + y_{12} + y_{13})V_1 - y_{12}V_2 - y_{13}V_3 = I_1$$
$$-y_{21}V_1 + (y_{21} + y_{23})V_2 - y_{23}V_3 = I_2$$
$$-y_{31}V_1 - y_{32}V_2 + (y_{30} + y_{31} + y_{32})V_3 = I_3$$

Cast into matrix form, these equations become

$$\left[\begin{pmatrix} y_{10} + y_{12} + y_{13} & -y_{12} & -y_{13} \\ -y_{21} & y_{21} + y_{23} & -y_{23} \\ -y_{31} & -y_{32} & y_{30} + y_{31} + y_{32} \end{pmatrix}\right] \left[\begin{pmatrix} V_1 \\ V_2 \\ V_3 \end{pmatrix}\right] = \left[\begin{pmatrix} I_1 \\ I_2 \\ I_3 \end{pmatrix}\right]$$

This can be abbreviated to the nodal-admittance matrix equation $\boldsymbol{YV} = \boldsymbol{I}$, matrix $\boldsymbol{Y}$ being known as the nodal-admittance matrix. The relationship between the branch elements and the corresponding matrix elements can be seen by inspection.

For some purposes the alternative impedance matrix equations are used: *viz.* $\boldsymbol{ZI} = \boldsymbol{V}$, where $\boldsymbol{Z} = \boldsymbol{Y}^{-1}$ is the nodal-impedance matrix. However, evaluating the inverse of $\boldsymbol{Y}$ is more complicated than finding $\boldsymbol{Y}$, and the admittance form may be considered as the primary (or given) form.

3.3.2 Load-flow analysis

Load-flow analysis is the solution of the nodal equations, subject to various constraints, to establish the node voltages. At the same time generator power outputs, transformer tap settings, branch power flows and powers taken by voltage-sensitive loads (including reactive power compensators) are determined.

3.3.2.1 Problem description

As indicated in *Figure 3.41*, the elements of a power system can be represented either as equivalent branches with appropriate admittance incorporated into the matrix $\boldsymbol{Y}$, or as equivalent current sources added to the matrix $\boldsymbol{I}$.

Transmission lines and cables These are represented by their series admittance and shunt (charging) susceptance. These parameters are actually distributed quantities, but are taken into account by Π (or, occasionally, T) equivalent networks (Section 3.2.5).

Transformers These are modelled by equivalent circuits with an ideal transformer in series with a leakage admittance (*Figure 3.42*(*a*)). With two terminals connected to a common reference point (earth) the circuit reduces to that in *Figure 3.42*(*b*): the ideal transformer with a turns ratio $(1+t)/1$ is replaced by an equivalent Π. The tap setting t represents the per-unit of nominal turns ratio: e.g. $t = \pm 0.05$ for ±5% taps.

Loads These can be represented either as equivalent admittances Y_{k0} connected between bus-bars and earth, or as current sources. If the load demand at bus-bar k is $S_k = P_k + jQ_k$, then the equivalent admittance at voltage V_k is found from

$$S_k = V_k I_k^* = |V_k|^2 Y_{k0}^* \quad \text{or} \quad Y_{k0} = S_k^*/|V_k|^2$$

where $|V_k|$ is assumed to be 1.0 p.u. (i.e. rated voltage). The equivalent admittance is incorporated into the corresponding diagonal element of the nodal-admittance matrix. In the current-source representation, the current I_k is substituted into the current column-matrix $\boldsymbol{I}$ by a fixed power requirement S_k and the (unknown) voltage V_k, where $I_k = S_k^*/V_k^*$.

Generator units or stations can likewise be represented by current sources, but usually the bus-bar to which a generator is connected has a controlled voltage. At a bus-bar m the requirement would be for specified values of active power P_m and voltage $|V_m|$, with the reactive power Q_m to be determined.

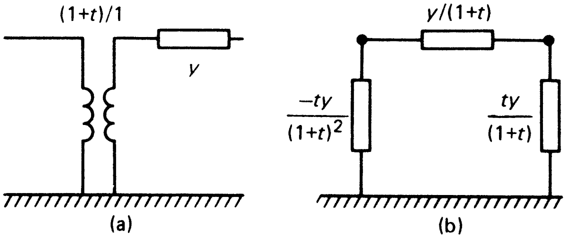


Figure 3.42 Transformer equivalent circuit

Slack bus-bar In a load-flow study, the total active power supplied cannot be specified in advance because the loss in the supply network will not be known. Further, in an *n*-bus-bar network, there are *n* complexor equations involving $2n$ real-number equations. However, there are $2n+2$ unknowns. To reduce this number to $2n$ it is the practice to specify the voltage of one bus-bar in both magnitude and phase angle. This is termed the *slack bus-bar*, to which the chosen *slack generator* is connected. The slack-bus-bar equation can now be removed from the solution process and, when all other voltages have been determined, the slack-bus-bar generation can be found. For a slack bus-bar k, for example, the generation S_k is found from

$$S_k = V_k \Sigma_m (Y_{km}{}^* V_m{}^*) \Leftarrow$$

3.3.2.2 *Network solution process*

Solution of the matrix equation is now sought. As it embodies several simultaneous equations, the solution has to be iterative. The two main procedures are the Gauss–Seidel and the Newton–Raphson methods. An important consideration in the computation process is the rate of convergence.

3.3.2.3 *Gauss–Seidel procedure*

This early (and still effective) technique resembles the over-relaxation method used in linear algebra. Consider a four bus-bar network described by the equations

$$Y_{11}V_1 + Y_{12}V_2 + Y_{13}V_3 + Y_{14}V_4 = S_1{}^*/V_1{}^*$$
$$Y_{21}V_1 + Y_{22}V_2 + Y_{23}V_3 + Y_{24}V_4 = S_2{}^*/V_2{}^*$$
$$Y_{31}V_1 + Y_{32}V_2 + Y_{33}V_3 + Y_{34}V_4 = S_3{}^*/V_3{}^*$$
$$Y_{41}V_1 + Y_{42}V_2 + Y_{43}V_3 + Y_{44}V_4 = S_4{}^*/V_4{}^*$$

where $Y_{12} = -y_{12}$, etc. Let bus-bar 1 be chosen as the slack bus-bar, and let V_1 be $1+\mathrm{j}0$. Then the remaining three equations are to be solved for V_2, V_3 and V_4. The method adopted is one of successive estimation.

First, the equations are rearranged by extracting the diagonal terms $Y_{11}V_1, Y_{22}V_2, \ldots,$ and transferring all other terms to the right-hand side. Each equation is then divided by the diagonal admittance element (Y_{11}, $Y_{22}, \ldots$).

If $V_k^{(p)}$ denotes the *p*th estimate of V_k and the equations are solved in the sequence 2–3–4–2–3– . . . , then an iterative process is the following:

$$\begin{aligned} V_2^{(p+1)\Leftarrow} = & -(Y_{21}/Y_{22})V_1 + 0 - (Y_{23}/Y_{22})V_3^{(p)\Leftarrow} \\ & -(Y_{24}/Y_{22})V_4^{(p)} + S_2{}^*/Y_{22}V_2{}^*(p) \end{aligned}$$

$$\begin{aligned} V_3^{(p+1)\Leftarrow} = & -(Y_{31}/Y_{33})V_1 - (Y_{32}/Y_{33})V_2^{(p+1)} + 0 \\ & -(Y_{34}/Y_{33})V_4^{(p)} + S_3{}^*/Y_{33}V_3{}^*(p) \Leftarrow \end{aligned}$$

$$\begin{aligned} V_4^{(p+1)\Leftarrow} = & -(Y_{41}/Y_{44})V_1 - (Y_{42}/Y_{44})V_2^{(p+1)\Leftarrow} \\ & -(Y_{43}/Y_{44})V_3^{(p+1)} + 0 + S_4{}^*/Y_{44}V_4{}^*(p) \Leftarrow \end{aligned}$$

Note that as each new estimate becomes available, it is used in the succeeding equations. Being iterative, the process of convergence can usually be assisted by the use of 'acceleration'. If $\Delta V_k^{(p)} = V_k^{(p+1)} - V_k^{(p)}$, then a new estimate can be obtained from $V_k^{(p+1)} = V_k^{(p)} + \omega \Delta V_k^{(p)}$; here ω is an accelerating factor, optimally a complex number but usually taken as real, with typical values in the range 1.0–1.6.

To terminate the successive-estimation process, various convergence tests are applied. The simplest is to examine the difference between successive voltage estimates and to stop when the maximum of $|V_k^{(p+1)} - V_k^{(p)}|$ for $k = 1, 2, \ldots, n$ is less than ϵ, a suitable small number such as 0.000 01 p.u. However, the preferred test is

$$V_k^{(p)\Leftarrow}\Sigma_m Y_{km}{}^* V_m{}^*(p) - S_k < \gamma\,\psi$$

where γ is a measure of the maximum allowable apparent-power mismatch at any bus-bar, with a value typically 0.01 p.u.

During iteration, other calculations (e.g. voltage magnitude corrections at generator bus-bars and changes in transformer tap settings) can be included. If bus-bar 2 in the example above is a generator bus-bar, then the reactive power Q_2 can be assigned an initial value, say $Q_2{}^{(0)} = 0$, and $V_2{}^{(1)}$ obtained therefrom. The voltage estimate can then be scaled to agree with the specified magnitude $|V_2|$, and $Q_2{}^{(1)}$ immediately calculated, prior to proceeding to the next equation.

The Gauss–Seidel procedure is well suited to implementation on a microcomputer, in which core space is limited. However, matrix-inversion techniques (as required in the Newton–Raphson procedure following) for large networks demand too much core space.

3.3.2.4 *Newton–Raphson procedure*

The Newton–Raphson procedure is at present the most generally adopted method. It has strong convergence characteristics and suits a wide range of problems. The method employs the preliminary terms in the Taylor series expansion (Section 1.2.4) of a set of functions of variables $\boldsymbol{V}$. The kth function is defined as

$$\mathbf{f}(V) = \mathbf{f}[V^{(p)} + \mathbf{J}(V^{(p)})(V - V^{(p)}]$$

The true set of values $\boldsymbol{V}$ is taken as given by $\boldsymbol{V} = \boldsymbol{V}^{(p)} + \gamma^{(p)}$, i.e. by the sum of an approximate set $\boldsymbol{V}^{(p)}$ and a set of error terms $\gamma^{(p)}$. Then, taking the first two terms of the Taylor expansion,

$$\mathbf{f}(V) = \mathbf{f}[V^{(p)} + \mathbf{J}(V^{(p)})(V - V^{(p)})] \Leftarrow$$

whence

$$\mathbf{J}(V^{(p)})\gamma^{(p)} = -\mathbf{f}(V^{(p)}) \Leftarrow$$

Matrix $\mathbf{J}(V^{(p)})$ is the Jacobian matrix of first derivatives of the functions $\mathbf{f}(V)$. Voltage estimates $V^{(p)}$ are used to evaluate specific matrix elements of $\mathbf{J}$; and $\gamma^{(p)}$ is the column matrix of voltage differences $\Delta V^{(p)}$ to be evaluated, these being the difference between the true and approximate values of the voltages V. Likewise, the term $-\mathbf{f}(V^{(p)})$ is the set of per-unit apparent-power differences $\Delta S^{(p)}$ between specified and calculated values, where

$$\Delta S_k{}^{(p)} = S_k - V_k{}^{(p)\Leftarrow}\Sigma_m Y_{km}{}^* V_m{}^*(p) \Leftarrow$$

The load-flow equation to be solved becomes

$$\mathbf{J}(V^{(p)})\Delta V^{(p)} = \Delta S^{(p)}$$

When $\Delta V^{(p)}$ is determined, the voltages are updated to $V^{(p+1)} = V^{(p)} + \Delta V^{(p)}$.

The polar form of the equations is most usually employed, so with $V_k = |V_k| \angle\delta_k$ and $Y_{km} = |Y_{km}| \angle{\Leftarrow}_{km}$ the function $\mathbf{f}_i = 0$ becomes

$$\begin{aligned} & (V_i Y_{i1} V_1 \cos\beta_{i1} + V_i Y_{i2} V_2 \cos\beta_{i2} + \cdots - P_i) \Leftarrow \\ & \quad + \mathrm{j}(V_i Y_{i1} V_1 \sin\beta_{i1} + V_i Y_{i2} V_2 \sin\beta_{i2} + \cdots - Q_i) = 0 + \mathrm{j}0 \end{aligned}$$

where $\beta_{i1} = \delta_i - \delta_1 - {\Leftarrow}_{i1}$ and similarly for $\beta_{i2}, \ldots$. Partial differentiation to form the terms of the Jacobian matrix, and then separation of the real and imaginary parts, gives the matrix equations to be solved. For generator bus-bars the voltage magnitudes are fixed, so that only equations in reals are needed to evaluate $\angle\,\delta$.

Generator bus-bars are often termed 'P, V' and load bus-bars referred to as 'P, Q' bus-bars, reflecting the values specified.

The Jacobian equations for the Newton–Raphson method are thus of the form

$$\begin{pmatrix} \partial \mathbf{f}_1/\partial\delta_1 & \partial \mathbf{f}_1/\partial\delta_2 & \cdots & \partial \mathbf{f}_1/\partial V_1 & \partial \mathbf{f}_1/\partial V_2 & \cdots \\ \partial \mathbf{f}_2/\partial\delta_1 & \partial \mathbf{f}_2/\partial\delta_2 & \cdots & \partial \mathbf{f}_2/\partial V_1 & \partial \mathbf{f}_2/\partial V_2 & \cdots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ \partial \mathbf{f}_1/\partial\delta_1 & \partial \mathbf{f}_1/\partial\delta_2 & \cdots & \partial \mathbf{f}_1/\partial V_1 & \partial \mathbf{f}_1/\partial V_2 & \cdots \\ \partial \mathbf{f}_2/\partial\delta_1 & \partial \mathbf{f}_2/\partial\delta_2 & \cdots & \partial \mathbf{f}_2/\partial V_1 & \partial \mathbf{f}_2/\partial V_2 & \cdots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \end{pmatrix} \begin{pmatrix} \Delta\delta_1 \\ \Delta\delta_2 \\ \vdots \\ \Delta V_1 \\ \Delta V_2 \\ \vdots \end{pmatrix} = \begin{pmatrix} \Delta P_1 \\ \Delta P_2 \\ \vdots \\ \Delta Q_1 \\ \Delta Q_2 \\ \vdots \end{pmatrix}$$

Written in abbreviated form, this is

$$\begin{pmatrix} \mathbf{J}_{11} & \mathbf{J}_{12} \\ \mathbf{J}_{21} & \mathbf{J}_{22} \end{pmatrix} \begin{pmatrix} \Delta\delta \\ \Delta V \end{pmatrix} = \begin{pmatrix} \Delta P \\ \Delta Q \end{pmatrix}$$

To save computer-memory space, it is usual to omit $\mathbf{J}_{12}$ and $\mathbf{J}_{12}$, an approximation that leaves two *decoupled* sets of equations. This approach, called the 'fast decoupled Newton–Raphson loadflow', is in wide use.

For any set of estimates of the voltage, the elements of the Jacobian matrix are evaluated, and the set of equations solved (using space-saving sparse-matrix programming techniques) for $\Delta\delta$ and ΔV; the values of V, ΔP and ΔQ are updated, and so on. Convergence is achieved for most networks in a few iterations.

A further development is to extend the Taylor series to the second-derivative term, when the series will terminate if Cartesian co-ordinates are employed. Iteration is more lengthy, but the convergence characteristics are more powerful. This 'second-order Newton–Raphson procedure' is gaining popularity.

3.3.3 Fault-level analysis

The calculation of three-phase fault levels in large power networks again involves solution of the nodal-admittance equations $\boldsymbol{YV} = \boldsymbol{I}$ subject to constraints.

3.3.3.1 System representation

Representation of generators and loads by fixed P, $|V|$ and P, Q requirements is not valid because of the large and sudden departure of the bus-bar voltages from their nominal values.

Passive loads are usually represented by a constant admittance, implying that the load power is proportional to the square of the bus-bar voltage. Relations $P \propto |V|^{1.2}$ and $Q \propto |V|^{1.6}$ would be more likely, but the $|V|^2$ proportionality affords a measure of demand variability and is more easily represented in the admittance matrix $\boldsymbol{Y}$.

Synchronous machines such as generators and motors are represented by a voltage source in series with an appropriate admittance. For example, at bus-bar k where the node voltage is V_k, the current could be represented by $I_k = y_k''(E_k'' - V_k)$ using the subtransient e.m.f. and admittance. The term $y_k'' V_k$ can be transferred to the other side of the nodal-admittance equation in such a way that y_k'' joins any load-admittance term in the diagonal element Y_{kk}.

The network equations have now been modified to the form

$$\boldsymbol{Y}''\boldsymbol{V} = \begin{pmatrix} \boldsymbol{y}''\boldsymbol{E}'' \\ \boldsymbol{0} \end{pmatrix}$$

where $\boldsymbol{Y}''$ is the admittance matrix $\boldsymbol{Y}$ with diagonal elements supplemented by equivalent load admittances and machine subtransient admittances, and the right-hand-side elements are either of the type $y_k''E_k''$ or zero.

3.3.3.2 Method of solution

In three-phase short-circuit conditions the voltages V will differ from the steady-state load values, but the right-hand-side elements will, with the exception of the element corresponding to the faulted bus-bar, remain constant. Let bus-bar m be short circuited; then $V_m = 0$. Solving the remaining equations for the voltages V_i (with $i = 1, 2, \ldots, n$, $i \neq m$) by the Gauss–Seidel procedure and then substituting the voltage values obtained in the mth equation yields a new right-hand-side value of one or other of the forms

$$y''_m E''_m + I_{m\text{sc}} \quad \text{or} \quad 0 + I_{m\text{sc}}$$

Here $I_{m\text{sc}}$ is the three-phase per-unit short-circuit current injected into bus-bar m to make $V_m = 0$. The fault level (in megavolt-amperes (MV-A)) is then

$$V_{m(\text{prefault})} \times I_{m\text{sc}}^{*} \times \text{MV-A(base)}$$

A preferred alternative uses the superposition theorem (Section 3.2.2.1). The injection of $I_{m\text{sc}}$, when acting alone, superposes a change ΔV_m $(= -V_{m(\text{prefault})})$ at bus-bar m. The equations to be solved become

$$\boldsymbol{Y}'' \begin{pmatrix} \Delta V_1 \\ \vdots \\ \Delta V_m \\ \vdots \\ \Delta V_n \end{pmatrix} = \begin{pmatrix} 0 \\ \vdots \\ I_{m\text{sc}} \\ \vdots \\ 0 \end{pmatrix} \quad \text{or} \quad \boldsymbol{Y}''\Delta\boldsymbol{V} = \boldsymbol{I}_{\text{sc}}$$

Inversion of $\boldsymbol{Y}''$ gives $\boldsymbol{V} = \boldsymbol{Z}''\boldsymbol{I}_{\text{sc}}$, in which the mth equation is known to be $-V_m = Z_{mm}''I_{m\text{sc}}$. If we assume nominal prefault voltage, i.e. $V_m = 1 + \text{j}0$ p.u., then the value of the three-phase short-circuit current at bus-bar m is $I_{m\text{sc}} = -1/Z_{mm}''$. The voltage at any other bus-bar k can then be found from

$$V_k = V_{k(\text{prefault})} + \Delta V_k = V_{k(\text{prefault})} + Z_{km}''I_{m\text{sc}}$$

By shifting the short circuit from bus-bar to bus-bar, the fault level for each can be found from the inverses of the appropriate diagonal elements of matrix $\boldsymbol{Z}''$.

3.3.4 System-fault analysis

The analysis of *unbalanced* faults in three-phase power networks is an important application of the *symmetrical-component* method (Section 3.2.12). The procedure for given fault conditions is as follows.

(1) Obtain the sequence impedance values for all items of the plant, equipment and transmission links concerned.
(2) Reduce all ohmic impedances to a common line-to-neutral base and a common voltage.
(3) Draw a single-line connection diagram for each of the sequence components, simplifying where possible (e.g. by star–delta conversion, see Section 3.2.4.5).
(4) Calculate the z.p.s., p.p.s. and n.p.s. currents, tracing them through the network to obtain their distribution with reference to the particular values sought.

Impedance in the neutral connection to earth, and in the earth path itself, must be multiplied by 3 for z.p.s. currents when the z.p.s. connection diagram is being set up in (3) because the three z.p.s. component currents are co-phasal and flow together in the z.p.s. path.

In general, a network offers differing impedances Z_+, Z_-⇐ and Z_0 to the sequence components. In static plant (e.g. transformers and transmission lines) Z_- may be the same as Z_+, but Z_0 is always significantly different from either of the other impedances. The presence of z.p.s. currents implies that a neutral connection is involved.

3.3.4.1 Sequence networks

As an example, *Figure 3.43* shows transmission lines 4 and 5–6 linking a generating station with generators 1 (isolated neutral) and 2 (solid-earthed neutral) to a second station with generator 3 (neutral earthed through resistor R_n). The numerals are used to indicate position: e.g. Z_{1+} is the p.p.s. impedance per phase of generator 1, and Z_{60} is the z.p.s. impedance of line 6 between generator 3 and a fault at F.

The *p.p.s. network* is identical with the physical set-up of the original network (which operates with p.p.s. conditions when normally balanced and unfaulted). Each generator is a source of p.p.s. voltages only. It is here assumed that all the generators develop the same e.m.f. E_a.

The *n.p.s. network* is similar in configuration (but not usually in impedance values) to the p.p.s. system. There are, however, no source e.m.f.s: the n.p.s. voltages are 'fictitious' ones developed by the fault.

The *z.p.s. network* is radically different from the other two, being concerned with neutral connections and earth faults. The effective line impedance is that of three conductors sharing equally the total n.p.s. current. To this must be added the earth-connection and earth-path impedances multiplied by 3, to give the z.p.s. impedance Z_0.

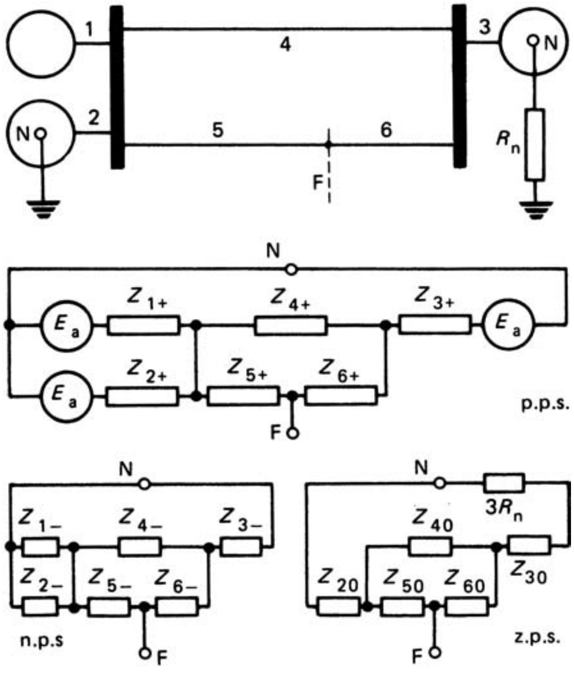


Figure 3.43 Phase-sequence networks

Typical values of Z_- and Z_0 in terms of Z_+ are

Ratio	*Generator*	*Transformer*	*Transmission link*	
			Overhead	*Cable*
Z_-/Z_+⇐	0.6–0.7	1.0	1.0	1.0
Z_0/Z_+⇐	0.1–0.8	1.0 or ∞⇐	3–5	1–3

The value of Z_0 for a synchronous generator depends on the arrangement of the stator winding.

To evaluate the system when faulted, it is necessary to determine the fault currents and the voltage of the sound line(s) to earth. If the voltages and currents at the fault are V_a, V_b, V_c and I_a, I_b, I_c, respectively, the following expressions always apply:

$$V_a = E_a - I_+Z_+ - I_-Z_- - I_0Z_0; \quad I_a = I_+ - I_- + I_0$$

$$V_b = \alpha^2E_a - \alpha^2I_+Z_+ - \alpha I_-Z_- - I_0Z_0; \quad I_b = \alpha^2I_+ + \alpha I_- + I_0$$

$$V_c = \alpha E_a - \alpha I_+Z_+ - \alpha^2I_-Z_- - I_0Z_0; \quad I_c = \alpha I_+ + \alpha^2I_- + I_0$$

where α is the 120° rotation operator (see Section 3.2.12). From the boundary conditions at the fault concerned it is possible to write three equations and to solve them for the symmetrical components I_+, I_- and I_0.

Sequence networks for some of the many transformer connections are shown in *Figure 3.44*. Further sequence networks are given in reference 1.

3.3.4.2 Boundary conditions

Three simple cases are shown in *Figure 3.45*. It is assumed that only fault currents are concerned, and that in-feed to the fault is from one direction.

(*a*) *Earth fault of resistance R_f on line A*—at the fault, $V_a = I_aR_f$ and $I_b = I_c = 0$. This leads to $I_{a0} = I_{a+} = I_{a-}$, so that the three sequence currents in phase A are identical. It follows that I_b and I_c are zero, as required. From the basic equations

$$I_{a+} = I_{a-} = I_{a0} = \tfrac{1}{3}I_a = E_a/(Z_+ + Z_- + Z_0 + 3R_f) = E_a/Z$$

and the fault current is $I_a = 3E_a/Z$. The three sequence networks are, in effect, connected in series. The component currents, and the voltages V_b and V_c, are obtained from those in phase A by application of the basic relations in Section 3.3.4.1. Each sequence current divides in the branches of its network in accordance with the configuration and impedance values.

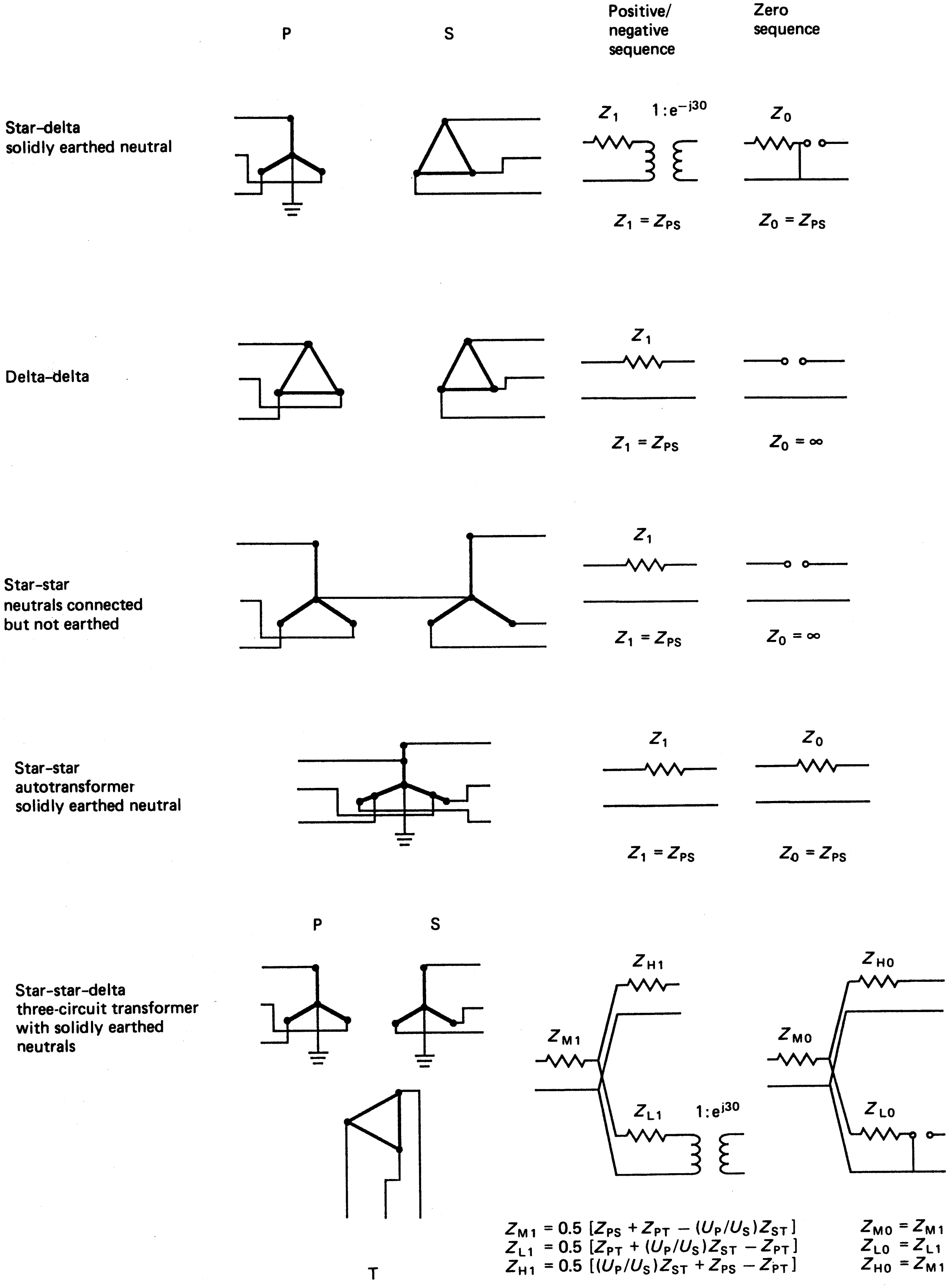
(*b*) *Short circuit between lines B and C*—the boundary conditions are $I_a = 0$, $I_b = -I_c$ and $V_b = V_c$. As there is no connection to earth at the fault, the z.p.s. network is omitted. The p.p.s. and fault currents are

$$I_{a+} = E_a/(Z_+ + Z_-) = E_a/Z$$
$$I_b = -I_c = -\mathrm{j}\sqrt{3}E_a/Z$$

where $Z = Z_+ + Z_-$ is the impedance of the p.p.s. and n.p.s. networks in series. The voltages to neutral at the fault are

$$V_a = E_a - I_+Z_+ - I_-Z_- = 2E_a(Z_-/Z)⇐$$
$$V_b = V_c = E_a(Z_-/Z)⇐$$

(*c*) *Double line–earth fault on lines B and C*—here the boundary conditions are $I_a = 0$, and $V_b = V_c = 0$. The sequence components of the fault current are


P
S
Positive/ negative sequence
Zero sequence
Star–delta solidly earthed neutral
Z_1
$1:e^{-j30}$
Z_0
$Z_1 = Z_{PS}$
$Z_0 = Z_{PS}$
Delta–delta
Z_1
$Z_1 = Z_{PS}$
$Z_0 = \infty$
Star–star neutrals connected but not earthed
Z_1
$Z_1 = Z_{PS}$
$Z_0 = \infty$
Star–star autotransformer solidly earthed neutral
Z_1
Z_0
$Z_1 = Z_{PS}$
$Z_0 = Z_{PS}$
P
S
Star–star–delta three-circuit transformer with solidly earthed neutrals
T
Z_{H1}
Z_{M1}
Z_{L1}
$1:e^{j30}$
Z_{H0}
Z_{M0}
Z_{L0}

$$Z_{M1} = 0.5\,[Z_{PS} + Z_{PT} - (U_P/U_S)Z_{ST}] \qquad Z_{M0} = Z_{M1}$$
$$Z_{L1} = 0.5\,[Z_{PT} + (U_P/U_S)Z_{ST} - Z_{PT}] \qquad Z_{L0} = Z_{L1}$$
$$Z_{H1} = 0.5\,[(U_P/U_S)Z_{ST} + Z_{PS} - Z_{PT}] \qquad Z_{H0} = Z_{M1}$$

Figure 3.44 Transformer equivalent sequence networks. U_P and U_S and the 3-phase MVA ratings of winding P and S respectively

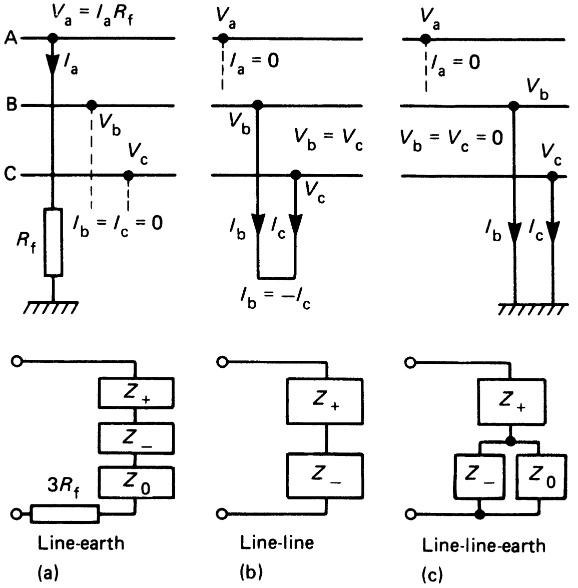


Figure 3.45 Boundary conditions at line faults

$$I_{a+} = E_a/[Z_+ + Z_-Z_0/(Z_- + Z_0)]\Leftarrow$$
$$I_{a-} = -I_{a+}[Z_0/(Z_- + Z_0)]\Leftarrow$$
$$I_{a0} = -I_{a+}[Z_-/(Z_- + Z_0)]\Leftarrow$$

The sequence networks are connected in series–parallel.

Figure 3.46 shows the interlinked phase-sequence networks where both ends feed the fault F. Conditions in (*a*), (*b*) and (*c*) correspond to those in *Figure 3.45*. Networks for a broken-conductor condition are shown in (*d*) and (*e*): the former is for a case in which both ends at the break remain insulated, while the latter applies where the conductor on side 2 falls to earth, the additional constraint involving ideal 1/1 transformers in side 2 of the combined sequence network. In more complicated cases, ideal transformers with phase-shift or with ratios other than 1/1 may be required. Evaluation of these, and of conditions involving simultaneous faults at different points and/or phases, requires matrix analysis and computer programs.

3.3.5 Phase co-ordinate analysis

The same techniques developed for the analysis of balanced networks, i.e. load-flow and three-phase fault-level analysis, can be used to develop the analysis of unbalanced faults or loads on either balanced or unbalanced networks using a phase representation of the system. The symmetrical component theory, which has formed the basis of unbalanced power system network analysis, was developed by several authors between 1912 and 1918. Although forming the basis of most modern computer-aided fault analyses, the method is limited by both balanced network assumptions and the difficulties associated with finding equivalent network models in the space described by the set of (transformed) 0, 1, 2 variables, e.g. for simultaneous faults.

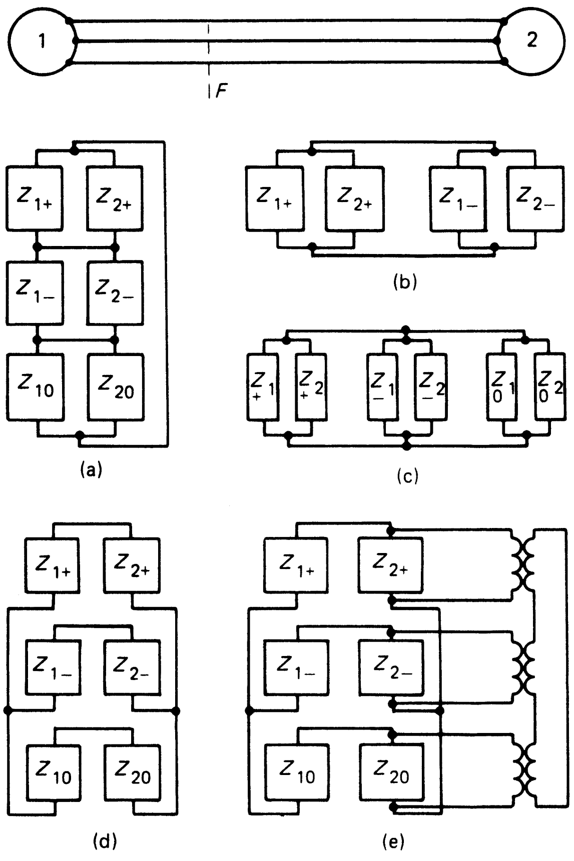


Figure 3.46 Interconnected phase-sequence networks

The phase co-ordinate method described here uses only the primary phase *a*, *b*, *c* variables and computer-based matrix computational methods to solve the resulting equations subject to the necessary constraints imposed by load-flow or fault analyses.$^{2–6}$

3.3.5.1 Element representation

For brevity, when referring to the sequence frame of reference, only the zero-, positive- and negative-sequence components usually associated with Fortescue will be discussed.

General element The general network element of the power system shown in *Figure 3.47* for a three-phase system may be described in matrix form by

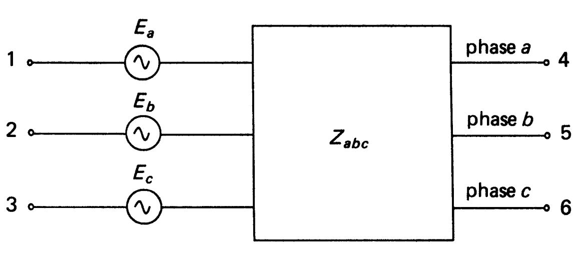


Figure 3.47 General three-phase system element

$$\boldsymbol{V}_{abc} = \boldsymbol{E}_{abc} + \boldsymbol{Z}_{abc}\boldsymbol{I}_{abc}$$

where $\boldsymbol{V}_{abc}$ represents the per-phase series voltage drops $(V_1 - V_4)$, $(V_2 - V_5)$ and $(V_3 - V_6)$, $\boldsymbol{E}_{abc}$ is the matrix of the equivalent voltage sources per phase, $\boldsymbol{I}_{abc}$ is the matrix of currents flowing per phase between nodes 1 and 4, 2 and 5, and 3 and 6, respectively, and $\boldsymbol{Z}_{abc}$ represents the passive three-phase mutually coupled network.

Defining the transformation relating the phase to the sequence components by

$$\boldsymbol{V}_{abc} = \boldsymbol{T}\boldsymbol{V}_{012}$$

where $\alpha = 1\angle 120°$ and, for the usual Fortescue components V_0, V_1 and V_2, will be used instead of the rather cumbersome notation V_0, V_+ and V_-; thus

$$\boldsymbol{T} = \begin{pmatrix} 1 & 1 & 1 \\ 1 & \alpha^2 & \alpha \\ 1 & \alpha & \alpha^2 \end{pmatrix}$$

Noting that $\boldsymbol{T}^{-1} = \frac{1}{3}\boldsymbol{T}^*$, upon transformation the equivalent sequence relationship becomes

$$\boldsymbol{T}\boldsymbol{V}_{012} = \boldsymbol{T}\boldsymbol{E}_{012} + \boldsymbol{Z}_{abc}\boldsymbol{T}\boldsymbol{I}_{012}$$

or

$$\boldsymbol{V}_{012} = \boldsymbol{E}_{012} + \boldsymbol{Z}_{012}\boldsymbol{I}_{012}$$

where

$$\boldsymbol{Z}_{012} = \tfrac{1}{3}\boldsymbol{T}^*\boldsymbol{Z}_{abc}\boldsymbol{T} \tag{3.3}$$

and likewise

$$\boldsymbol{Y}_{012} = \tfrac{1}{3}\boldsymbol{T}^*\boldsymbol{Y}_{abc}\boldsymbol{T}$$

For the customary linear device $\boldsymbol{Z}_{abc}$ is symmetric, but without further assumptions $\boldsymbol{Z}_{012}$ is unsymmetric.

Transmission lines The series relationships along a transmission line may be represented by the partitioned equations

$$\begin{pmatrix} \boldsymbol{V}_1 \\ \boldsymbol{V}_2 \\ \boldsymbol{0} \end{pmatrix} = \begin{pmatrix} \boldsymbol{Z}_{11} & \boldsymbol{Z}_{12} & \boldsymbol{Z}_{13} \\ \boldsymbol{Z}_{21} & \boldsymbol{Z}_{22} & \boldsymbol{Z}_{23} \\ \boldsymbol{Z}_{31} & \boldsymbol{Z}_{32} & \boldsymbol{Z}_{33} \end{pmatrix} \begin{pmatrix} \boldsymbol{I}_1 \\ \boldsymbol{I}_2 \\ \boldsymbol{I}_3 \end{pmatrix}$$

where $\boldsymbol{V}_1$ and $\boldsymbol{V}_2$ are column submatrices representing the series voltages along conductors and subconductors, respectively, carrying the currents in the submatrices $\boldsymbol{I}_1$ and $\boldsymbol{I}_2$. $\boldsymbol{I}_3$ is a submatrix representing the currents in any earth wires present, which are assumed to be solidly earthed at each end. This latter assumption is normal, but unnecessary, for the solution of the system equations, because extra equations representing earth wires can be included for solution.

A matrix reduction (completion of bundling process and earth-wire removal) will then yield the equations of the equivalent phase conductors, thus giving

$$\boldsymbol{V}_{abc} = \boldsymbol{Z}_{abc}\boldsymbol{I}_{abc}$$

The 3×3 matrix $\boldsymbol{Z}_{abc}$ represents the series impedance of the transmission line, including all the effects due to unbalanced configuration and the use of bundled conductors and earth wires. In a three-phase representation of the network this matrix may be treated in the same way as the single series impedance of the more usual one-line diagram.

By similar reasoning, the potential coefficients of the transmission line (phase and earth wires) per unit of length may be computed for any configuration. From these coefficients a shunt-reactance matrix is found, and then, by reduction, the 3×3 matrix $\boldsymbol{Z}_{shunt}$ representing the total distributed capacitive reactance over the length of the line. Inverting this matrix gives $\boldsymbol{Y}_{shunt}$ from which the Maxwell coefficients, and hence capacitances, can be determined if necessary.

Three-phase transmission-line models follow as extensions of the nominal or distributed π and T representations. In the π representation one-half of the transmission-line equivalent capacitance is connected to each end of the line.

The nodal voltages and currents injected into the busbars at each end of the transmission line are then related by

$$\begin{pmatrix} I_1 \\ I_2 \\ I_3 \\ I_4 \\ I_5 \\ I_6 \end{pmatrix} = \begin{pmatrix} \boldsymbol{Y}_{abc} + \frac{1}{2}\boldsymbol{Y}_{shunt} & -\boldsymbol{Y}_{abc} \\ -\boldsymbol{Y}_{abc} & \boldsymbol{Y}_{abc} + \frac{1}{2}\boldsymbol{Y}_{shunt} \end{pmatrix} \begin{pmatrix} V_1 \\ V_2 \\ V_3 \\ V_4 \\ V_5 \\ V_6 \end{pmatrix}$$

The similarity between the three-phase and single-phase admittance matrices is evident—each element in the single-phase matrix is replaced by the appropriate 3×3 admittance sub-matrix.

Machine representation The general element shown in *Figure 3.47* may be taken to represent any synchronous or induction machine with a star-connected stator winding when nodes 1, 2 and 3 are short circuited. For a balanced design the voltage sources would be displaced 120° from each other and be equal in magnitude.

The phase-impedance matrix $\boldsymbol{Z}_{abc}$ can be found by simple transformation of the normally available sequence impedance matrix, as indicated in equations (3.3) and (3.4). The sequence impedance matrix here reflects the structure of the machine and the purpose of the study in the selection of the positive sequence reactance. A typical salient-pole diesel generator, for example, with segmented dampers, 14 poles and 90.7% winding pitch rated at 1340 kV-A, 3.3 kV, 50 Hz has negative and zero reactances of $X_2 = 0.274$ p.u. and $X_0 = 0.067$ p.u., respectively. In a fault-level study the machine is represented in the positive-sequence circuit by the subtransient reactance $X_d{}' = 0.227$ p.u. $= X_1$. The phase-impedance matrix would then be

$$\boldsymbol{Z}_{abc} = \frac{1}{3}\boldsymbol{T}\boldsymbol{Z}_{012}\boldsymbol{T}^* =$$

$$\begin{pmatrix} 0+\mathrm{j}0.1893 & 0.0138-\mathrm{j}0.0612 & -0.0138-\mathrm{j}0.0612 \\ -0.0138-\mathrm{j}0.0612 & 0+\mathrm{j}0.1893 & 0.0138-\mathrm{j}0.0612 \\ 0.0138-\mathrm{j}0.0612 & -0.0138-\mathrm{j}0.0612 & 0+\mathrm{j}0.1893 \end{pmatrix} \tag{3.4}$$

Note that, with $X_1 \neq X_2$, the phase-impedance matrix elements contain both positive and negative real parts, although no resistances were included in the sequence impedances.

Such a representation neglects the effective winding capacitances. These capacitances could be added, if necessary, in the same manner as the transmission-line capacitances in the three-phase representation in delta and shunt connections (a four-terminal lattice network with one terminal earthed). Neglected also is any representation of a machine neutral node. Instead of adding an appropriate row or column to $\boldsymbol{Z}_{abc}$, the effect of any earthing reactance X_g can be included in the value of X_0, i.e. $X_0 + 3X_g$.

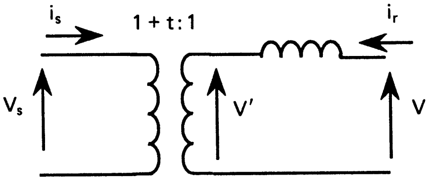


Figure 3.48 Per-unit schematic representation of a single-phase transformer

Transformer representation The third type of network element is the transformer in one-, two- and three-phase circuits with all the associated possible variations of construction and connection.

(*a*) *Derivation of equivalent-circuit model.* A single-phase representation of a transformer in per-unit form illustrated schematically in *Figure 3.48* by an ideal transformer of turns ratio $(1+t)$:1 with an equivalent leakage admittance of y per unit.

From the relationships across the ideal transformer an equivalent circuit model is as shown in *Figure 3.49* and thus represents a single-phase tapped transformer.

If nodes k and q are earthed as in the one-line diagram representation of a balanced three-phase system, the lattice circuit reduces to the equivalent π-representation, as shown in *Figure 3.50.*

The lattice equivalent of *Figure 3.49* provides an adequate model for single-phase variable-turns-ratio transformers and in combinations for three-phase star–star banks with tapped windings, but can only be used with care in banks containing delta-connected windings. In a star–delta bank of single-phase transformer units, for example, with normal turns ratio, a value of 1.0 p.u. voltage on each leg of the star winding produces under balanced conditions 1.732 p.u. voltage on each leg of the delta winding (rated line to neutral voltage as base). The structure of the bank requires in the per-unit representation an effective tapping at $\sqrt{3}$ times nominal turns ratio on the delta side, i.e. $1+t=1.732$ or $t=0.732$.

For a delta–delta or star–delta transformer with taps on the star winding, the equivalent circuit of *Figure 3.49* would have to be modified to allow for effective taps to be represented on each side. This, the general symmetrical lattice equivalent circuit of a single-phase transformer where both primary and secondary windings may have either actual or equivalent variable turns ratios α and β or both is shown in *Figure 3.51.* This single-phase transformer model can be used to assemble equivalent circuits of polyphase transformer banks, some of which are shown below

(*b*) *Star–star transformer.* For a two-circuit three-phase transformer or autotransformer connected in a star–star arrangement, the equivalent circuit is as shown in linear-graph form in *Figure 3.52.* Parallel transformer windings are taken to represent equivalent single-phase transformers. The circuit is constructed from the simple connection of three of the general circuits shown in *Figure 3.51* with taps on both windings. In practice, of course, either α or β, or both, would be 1.0 p.u.

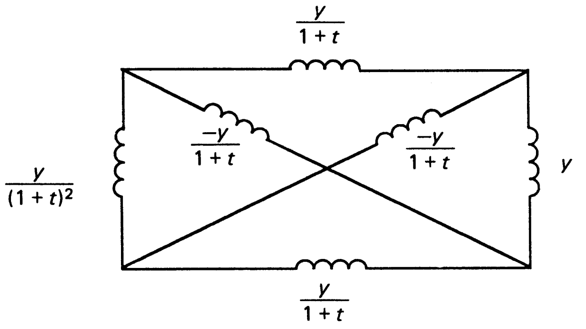


Figure 3.49 Symmetrical lattice equivalent circuit of a single-phase transformer with a variable turns ratio

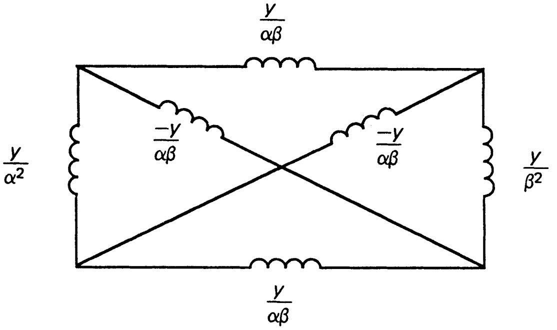


Figure 3.51 General transformer symmetrical lattice equivalent circuit with primary and secondary equivalent turns α and β per unit

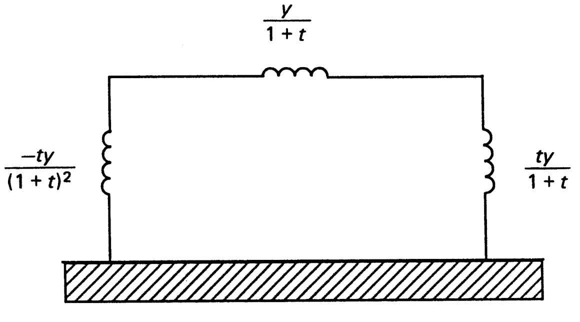


Figure 3.50 Single-phase equivalent π representation when nodes k and q are earthed

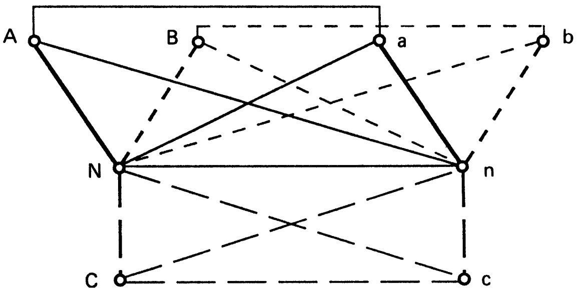


Figure 3.52 A three-phase equivalent circuit of a star–star connected transformer

Table 3.6 Connection table for the star–star transformer equivalent circuit shown in *Figure 3.52* ($\alpha = 1 + t_\alpha$ p.u.; $\beta = 1 + t_\beta$ p.u.)

Admittance	*Between nodes*
$y/\alpha^2\psi$	N–A, N–B, N–C
$y/\beta^2\psi$	n–a, n–b, n–c
$y/\alpha\beta\psi$	A–a, B–b, C–c
$-y/\alpha\beta\psi$	n–A, n–B, n–C; N–a, N–b, N–c
$3y/\alpha\beta\psi$	N–n

In a more concise form, the equivalent circuit of *Figure 3.52* may be described by the connection table given in *Table 3.6* where, for example, an admittance of value $y/\alpha^2\psi$ is connected between N and A, also N and B, etc. If the neutrals are earthed or connected together either solidly or through an impedance, the appropriate additions or deletions can be made to the circuit and corresponding terms changed in the connection table. From inspection of the circuit, the corresponding admittance matrix can be assembled with or without rows and columns for the neutral nodes, depending on the earthing arrangements.

(*c*) *Delta–delta transformer.* With the same convention (that parallel windings may be considered to represent single-phase transformers) the equivalent circuit of a delta–delta transformer may be constructed by the same principles, as shown in *Figure 3.53*. The corresponding connection table is shown in *Table 3.7*, where both windings have variable ratios. With taps on one winding only either $\alpha\psi$ or $\beta\psi$ would have the value $\sqrt{3}$, the tap value $t_{\alpha\psi}$ or $t_{\beta\psi}$ being zero accordingly.

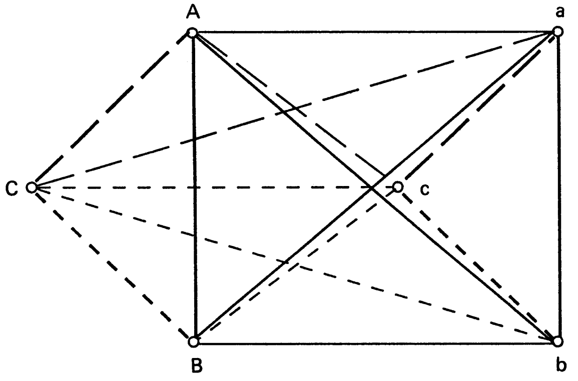


Figure 3.53 A three-phase equivalent circuit of a delta–delta connected transformer

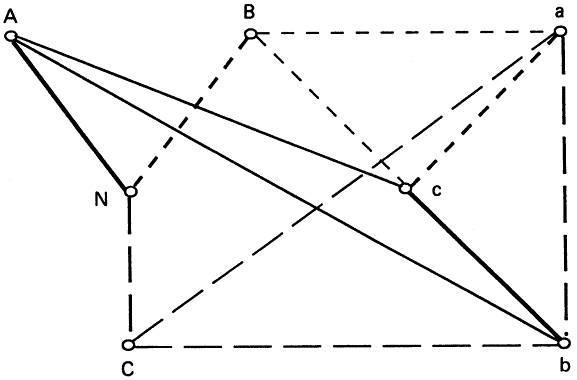


Figure 3.54 A three-phase equivalent circuit of a star–delta connected transformer

(*d*) *Star–delta transformer.* Using the same techniques, the three-phase equivalent circuit model of a star–delta transformer may be assembled and is shown in *Figure 3.54*. The convention used for numbering nodes and thus identifying opposite sides of the symmetrical lattice networks is as follows:

A–N/c–b; B–N/a–c; C–N/b–a.ψ

The connection table for *Figure 3.54* is given in *Table 3.8*. Again, for taps on one side of the transformer only, either $t_{\alpha\psi}$ or $t_{\beta\psi}$ is zero. The neutral node *N* in the table can be identified with the reference earth node if solidly earthed or extra terms can be added to the table or not, according to the earthing arrangements

(*e*) *Three-winding transformers.* With the same assumptions the analysis can be extended to three-circuit transformers and autotransformers and to any multiwinding transformer regardless of the number of circuits. Consider, for example, a star–star–delta transformer with solidly earthed neutrals. Let the star primary and secondary winding (P and S) terminals be labelled A, B, C, N and A′⇐B′, C′⇐ N′, respectively, with the delta tertiary winding (T) terminals labelled a, b, c, as shown in *Figure 3.55*.

If y_{PS}, y_{PT} and y_{ST} are the short-circuit per-unit admittances of the two windings indicated by the subscripts with the third winding open, a three-phase equivalent circuit can be assembled from paralleling one star–star and two star–delta equivalent circuits in turn. The circuit line diagram is too complex to illustrate conveniently, but with the same convention concerning the matching of parallel sides in the identification of the single-phase units, A–N with A′–N′⇐ A–N with c–b and A′–N′⇐with c–b, etc., the connection

Table 3.7 Connection table for the delta–delta transformer in the equivalent circuit shown in *Figure 3.53* ($\alpha = \sqrt{3}(1 + t_\alpha)$ p.u.; $\beta = \sqrt{3}(1 + t_\beta)$ p.u.)

Admittance	*Between nodes*
$y/\alpha^2\psi$	A–B, B–C, C–A
$y/\beta^2\psi$	a–b, b–c, c–a
$2y/\alpha\beta\psi$	A–A, B–b, C–c
$-y/\alpha\beta\psi$	A–b, B–c, C–a; a–B, b–C, c–A

Table 3.8 Connection table for the star–delta transformer in the equivalent circuit shown in *Figure 3.54* ($\alpha = 1 + t_\alpha$ p.u.; $\beta = \sqrt{3}\ (1 + t_\beta)$ p.u.)

Admittance	*Between nodes*
$y/\alpha^2\psi$	A–N, B–N, C–N
$y/\beta^2\psi$	a–b, b–c, c–a
$y/\alpha\beta\psi$	A–c, B–a, C–b
$-y/\alpha\beta\psi$	A–b, B–c, C–a

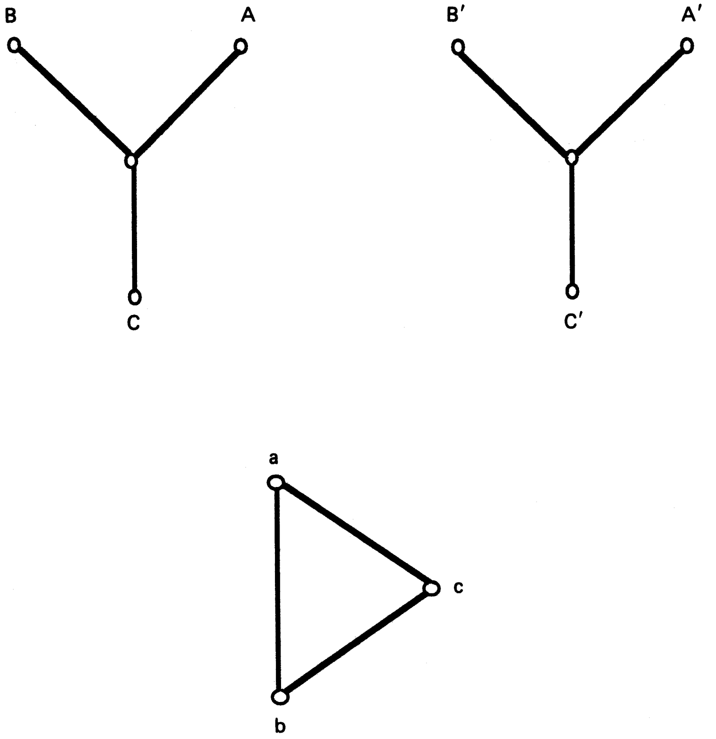


Figure 3.55 A star–star–delta transformer showing terminal markings

table will be as in *Table 3.9*, where $\alpha\psi$ represents the turns ratio of winding P, $\beta\psi$ of winding S, and $\gamma\psi$ of winding T. Neutrals N and N′⇐ are solidly earthed.

Note that, because two symmetrical lattice networks are connected to any two nodes on the same winding, a and b for example, the total admittance between these nodes is the sum of the corresponding admittances belonging to each of the two lattices between these nodes, e.g.

$$y_{PT}/\gamma^2\psi + y_{ST}/\gamma^2\psi = (y_{PT} + y_{ST})/\gamma^2\psi$$

(*f*) *Open-delta transformer.* The validity of the equivalent-network model in representing unbalanced transformer designs is demonstrated in the analysis of the open-delta transformer. *Figure 3.56* shows a schematic circuit diagram of the transformer with each delta open opposite nodes A and a, respectively. Connecting parallel branches of the windings by symmetrical-lattice equivalent circuits yields the connection table shown in *Table 3.10*.

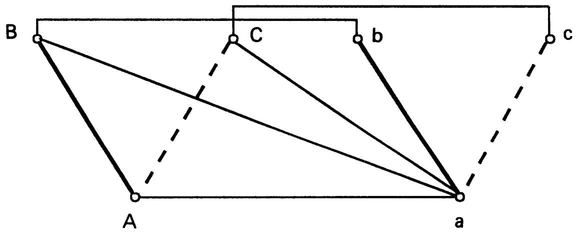


Figure 3.56 Equivalent circuit of an open-delta transformer

Table 3.9 Connection table for a three-winding star–star–delta transformer ($\alpha = 1 + t_p$ p.u.; $\beta = 1 + t_s$ p.u.; $\gamma = \sqrt{}\ 3(1 + t_T)$ p.u.)

Admittance	*Between nodes*
$(y_{PS} + y_{PT})/\alpha^2\psi$	N–A, N–B, N–C
$(y_{PS} + y_{ST})/\beta^2\psi$	N′–A′, N′–B′, N′–C′⇐
$y_{PS}/\alpha\beta\psi$	A–A′, B–B′, C–C′⇐
$-y_{PS}/\alpha\beta\psi$	N′–A, N′–B, N′–C; N–A′, N–B′,⇐
$(y_{PT} + y_{ST})/\gamma^2\psi$	N–C′⇐
$y_{PT}/\alpha\gamma\psi$	a–b, b–c, c–a
$y_{ST}/\beta\gamma\psi$	A–c, B–a, C–b
$-y_{PT}/\alpha\gamma\psi$	A′–c, B′–a, C′–b
$-y_{ST}/\beta\gamma\psi$	A–b, B–c, C–a
	A′–b, B′–c, C′–a

Equivalent circuits for other unbalanced transformers and for transformers with different numbers of primary and secondary phases, i.e. *m*-to-*n* transformers such as Scott, Zig-zag, Vee transformers, etc., can be found in reference 5.

(*g*) *Effects of magnetising impedances.* The preceding transformer models do not account for the effects of core structure and saturation. In particular, it is noted that the phase circuits give the same representation as transformer sequence impedance circuits with a series impedance of equal value in each of the positive-, negative-, and zero-sequence circuits. More accurate representations include the transformer magnetising impedances in three-phase transformers.

Table 3.10 Connection table for the open-delta transformer with each winding open opposite nodes A and a respectively ($\alpha = \sqrt{3}(1 + t_\alpha)$ p.u.; $\beta = \sqrt{3}(1 + t_\beta)$ p.u.)

Admittance	*Between nodes*
$y/\alpha^2\psi$	A–B, C–A
$y/\beta^2\psi$	a–b, c–a
$y/\alpha\beta\psi$	B–b, C–c
$-y/\alpha\beta\psi$	A–b, A–c, B–a, C–a
$2y/\alpha\beta\psi$	A–a

In terms of the sequence quantities, these impedances are of particular importance in the zero-sequence networks. High values of zero-sequence voltage in shell-type transformers, and the effects of the tank walls and out-of-core return paths for zero-sequence fluxes in three-legged core-type transformers, give magnetising impedances of the same order of magnitude as the system impedances.

The transformer equivalent sequence networks can be modified to become T networks, with the magnetising impedance sequence components in the legs of the Ts. In phase coordinates the shunt impedance branch can be added likewise to the transformer single-phase model, as indicated in *Figure 3.57*. The dotted line indicates the part of the transformer short-circuit impedance placed in the lattice networks; the other part could be incorporated into a lattice network if equivalent taps are required on both sides. The equality of the series admittances has no significance.

Starting from the usual measures of these magnetising admittances in sequence terms, the three-phase equivalent circuit can be developed as follows where, for a balanced design,

$$\boldsymbol{Y}_{abc} = \tfrac{1}{3}\boldsymbol{T}\boldsymbol{Y}_{012}\boldsymbol{T}^*$$

$$= \tfrac{1}{3}\begin{pmatrix} Y_0 + Y_1 + Y_2 & Y_0 + \alpha Y_1 + \alpha^2 Y_2 & Y_0 + \alpha^2 Y_1 + \alpha Y_2 \\ Y_0 + \alpha^2 Y_1 + \alpha Y_2 & Y_0 + Y_1 + Y_2 & Y_0 + \alpha Y_1 + \alpha^2 Y_2 \\ Y_0 + \alpha Y_1 + \alpha^2 Y_2 & Y_0 + \alpha^2 Y_1 + \alpha Y_2 & Y_0 + Y_1 + Y_2 \end{pmatrix}$$

This phase-admittance matrix represents an equivalent delta network with connections to earth at each node where $y_1 = y_2$ as shown in *Figure 3.58*.

(*h*) *Phase-shifting transformers.* Phase-shifting transformers may be represented in a similar manner from an assembly of single-phase elements, but here the single-phase elements have to be derived.

(1) *Single-phase equivalent-circuit model*—following the method described in Section 3.3.5.2, the single-phase representation may be illustrated approximately as in *Figure 3.59*, where the ideal transformer now represents an ideal phase-shifting transformer. The invariance of the product $V_1{}^*$ across the ideal transformer requires a distinction to be made between the turns ratios for current and voltage; thus

$$V_\mathrm{s} = (1 + t_\mathrm{v})V'^{\Leftarrow}$$
$$i_\mathrm{r} = -(1 + t_1)i_\mathrm{s}$$

where $1 + t_\mathrm{v} = 1 + t + \mathrm{j}q$ and $1 + t_\mathrm{I} = 1 + t - \mathrm{j}q$.

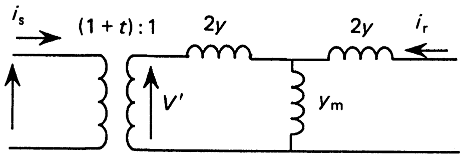


Figure 3.57 Per-unit schematic representation of a single-phase transformer with core representation

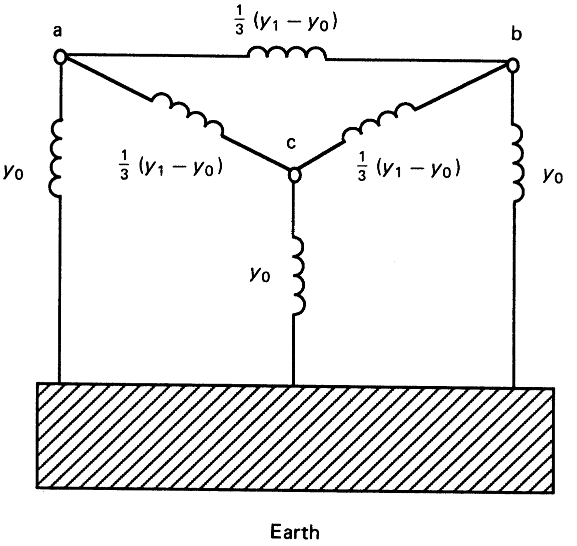


Figure 3.58 Phase representation of core branch impedances

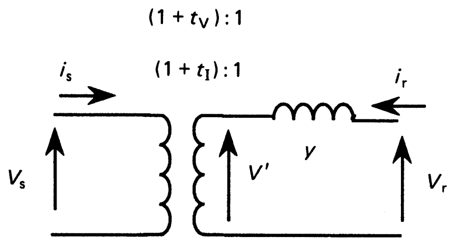


Figure 3.59 Per-unit schematic representation of a single-phase phase-shifting transformer

Following the same procedure as before yields the 4×4 phase admittance matrix $\boldsymbol{Y}$ of the equivalent phase-shifting transformer. For a phase-shifting transformer, however, although an equivalent lattice network corresponding to the admittance matrix $\boldsymbol{Y}$ can be constructed, it is no longer a bilinear network because of asymmetry in $\boldsymbol{Y}$. The equivalent circuit of a single-phase phase-shifting transformer is thus of limited value, and the transformer is best represented algebraically by its admittance matrix.

(2) *Three-phase phase-shifting transformers*—For both non-phase-shifting and phase-shifting transformers the phase-admittance matrix for any polyphase bank can be built up from the single-phase admittance matrices by identification of the nodes of each single-phase unit with the three-phase terminations, A, B, C and a, b, c according to the winding connection.

The three-phase admittance matrix of the star–delta phase-shifting transformer is:

$$
\boldsymbol{Y}_{\boldsymbol{3phase}} = \begin{array}{c} \\ \mathrm{A} \\ \mathrm{B} \\ \mathrm{C} \\ \mathrm{N} \\ \mathrm{a} \\ \mathrm{b} \\ \mathrm{c} \end{array}
\begin{array}{c}
\begin{array}{ccccccc} \mathrm{A} & \mathrm{B} & \mathrm{C} & \mathrm{N} & \mathrm{a} & \mathrm{b} & \mathrm{c} \end{array} \\
\begin{pmatrix}
\frac{y}{\alpha_I\alpha_V} & 0 & 0 & \frac{-y}{\alpha_I\alpha_V} & 0 & \frac{y}{\alpha_I\beta} & \frac{-y}{\alpha_I\beta\psi} \\
0 & \frac{y}{\alpha_I\alpha_V} & 0 & \frac{-y}{\alpha_I\alpha_V} & \frac{-y}{\alpha_I\beta} & 0 & \frac{y}{\alpha_I\beta} \\
0 & 0 & \frac{y}{\alpha_I\alpha_V} & \frac{-y}{\alpha_I\alpha_V} & \frac{y}{\alpha_I\beta} & \frac{-y}{\alpha_I\beta\psi} & 0 \\
\frac{-y}{\alpha_I\alpha_V} & \frac{-y}{\alpha_I\alpha_V} & \frac{-y}{\alpha_I\alpha_V} & \frac{3y}{\alpha_I\alpha_V} & 0 & 0 & 0 \\
0 & \frac{-y}{\alpha_V\beta} & \frac{y}{\alpha_V\beta\psi} & 0 & \frac{2y}{\beta^2} & \frac{-y}{\beta^2} & \frac{-y}{\beta^2} \\
\frac{y}{\alpha_V\beta} & 0 & \frac{-y}{\alpha_V\beta\psi} & 0 & \frac{-y}{\beta^2} & \frac{2y}{\beta^2} & \frac{-y}{\beta^2} \\
\frac{-y}{\alpha_V\beta} & \frac{y}{\alpha_V\beta} & 0 & 0 & \frac{-y}{\beta^2} & \frac{-y}{\beta^2} & \frac{2y}{\beta^2}
\end{pmatrix}
\end{array}
$$

The neutral node N has been preserved in the matrix. If N was earthed through an earthing impedance, the appropriate admittance would appear in the element y_{NN} or, if earthed directly, the row and column corresponding to N would be removed.

3.3.5.2 Fault analysis

By representing polyphase network conditions in terms of their phase co-ordinates, i.e. phase voltages, currents and impedances, thereby preserving the physical identity of the system, instead of transforming the phase co-ordinates to symmetrical component co-ordinates, a generalised analysis of polyphase networks under all fault conditions can be developed.

General form of prefault equations It was shown in Section 3.3.3.3 that the general form of the nodal admittance equations $\boldsymbol{YV} = \boldsymbol{I}$ may be used to describe the three-phase system where each bus-bar in the one-line diagram of the balanced system is replaced by three equivalent separate-phase bus-bars. Each voltage and current element in equations $\boldsymbol{YV} = \boldsymbol{I}$ for the balanced system is replaced correspondingly by three phase-to-earth voltages and three currents, with each element of the nodal-admittance matrix being replaced by a three-phase element represented by a 3×3 nodal admittance submatrix. The same principle used in the assembly of the single-phase admittance matrix underlies the assembly of the three-phase admittance matrix.

The phase relationships at each bus-bar are, then, at bus-bar k, for example,

$$I_k = S_k^*/V_k^*$$

where I_k is the phase current injected into bus-bar k, V_k is the phase-to-earth voltage, and S_k is the phase power.

The network admittance matrix $\boldsymbol{Y}$ can be adjusted to the basic prefault form from which all calculations for the various faults commence by classifying the energy sources as active or passive, according to their behaviour during fault conditions. Some loads, for example, may be characterised by passive admittances per phase; thus the equation for node j becomes

$$I_{\mathrm{j}} = y'_{\mathrm{j0}} V_{\mathrm{j}}$$

The network nodal admittance equations can be modified accordingly by substituting for the load currents and then transferring the admittances across to supplement the diagonal elements of matrix $\boldsymbol{Y}$. If the load has unequal positive-negative-, and zero-sequence admittances, the admittance y_{j0}' may be replaced by an equivalent 3×3 phase admittance matrix, which in turn is transferred to supplement the appropriate block diagonal 3×3 submatrix of $\boldsymbol{Y}$. In this case, the substitution for the three corresponding phase currents is made simultaneously.

With appropriate node connections, active sources such as machines may be represented by the general network element shown previously containing voltage sources in series with a passive network. The equations $\boldsymbol{YV} = \boldsymbol{I}$ governing the current injected into the network from a star-connected machine connected to the three-phase bus-bars p, q, and r with a neutral earthed through an impedance y_{N0}, are:

$$
\begin{pmatrix}
Y_{\mathrm{pp}} & Y_{\mathrm{pq}} & Y_{\mathrm{pr}} & -y_0 \\
Y_{\mathrm{qp}} & Y_{\mathrm{qq}} & Y_{\mathrm{qr}} & -y_0 \\
Y_{\mathrm{rp}} & Y_{\mathrm{rq}} & Y_{\mathrm{rr}} & -y_0 \\
-y_0 & -y_0 & -y_0 & Y_{\mathrm{NN}}
\end{pmatrix}
\begin{pmatrix} V_1 \\ V_2 \\ V_3 \\ V_{\mathrm{N}} \end{pmatrix}
=
\begin{pmatrix}
y_1 & E_{\mathrm{p}} \\
y_1 & E_{\mathrm{q}} \\
y_1 & E_{\mathrm{r}} \\
-y_0 & E_{\mathrm{N}}
\end{pmatrix}
$$

where I_{N} is the current injected into the neutral node N, and is usually zero, E_{N} is the sum of the phase e.m.f.s, and is also usually zero ($E_{\mathrm{p}} + E_{\mathrm{q}} + E_{\mathrm{r}} = 0$), y_0 is the machine zero sequence admittance, y_1 is the positive-sequence admittance, and $Y_{\mathrm{NN}} = 3y_0 + y_{\mathrm{N0}}$. The values of the voltage sources E_{p}, E_{q} and E_{r} are the appropriate phase-displaced transient or subtransient values according to the purpose of the study.

Substituting again for the respective currents in equation $\boldsymbol{YV} = \boldsymbol{I}$ and transferring the product terms $y_{\mathrm{ij}}V_{\mathrm{j}}$ to the left-hand side, the modified network admittance equations become

$$\boldsymbol{Y}'\boldsymbol{V} = [\boldsymbol{y_g E}] = \boldsymbol{I}' \tag{3.5}$$

where $\boldsymbol{Y}'$ is the supplemented phase admittance matrix and $[\boldsymbol{y_g E}]$ or $\boldsymbol{I}'$ is a column matrix the elements of which are of the form y_1E, or zero. The similarity is evident between the three-phase prefault equations and the familiar equivalent one-line diagram or single-phase prefault equations established for three-phase fault studies of balanced systems described in Section 3.3.3.

Solution of equations for various types of fault condition All the various types of fault condition can be analysed by means of simple modifications to, and the solution of, these equations.

(*a*) *Single phase-to-earth fault.* If a single phase-to-earth short circuit occurs at bus-bar k, the bus-bar voltage V_k will be constrained to be zero, the value of the earth reference voltage. To obtain the extra degree of freedom for this constraint to be valid within a consistent set of equations, the current on the right-hand side of the kth equation in equation (3.5) must be unspecified, and allowed to take its value according to the solution of the remaining set of $(n-1)$ equations in the $(n-1)$ unknown voltages, V_{j}, $i = 1, 2, \ldots, n$, $(i \neq k)$, with $V_k = 0$.

Letting the value of this kth current be I''_k, where

$$I''_k = \Sigma Y'_{km} V_m$$
$$= I'_k + I_{SC}$$

I'_k is the prefault value and Y'_{km} ($m = 1, 2, \ldots, n$) are the elements of the kth row of the matrix $\boldsymbol{Y}'$, then I_{SC} is the current in the short-circuit connection to earth.

This calculation is exactly the same as that for a three-phase fault in existing computer programs based on a one-line diagram and three-phase apparent-power base.

The numerical methods of solution are the standard methods developed for such problems in linear algebra, namely matrix-inversion methods or, for large systems, iterative methods such as the Gauss–Seidel technique, and can be found in the usual numerical-analysis textbooks.

(*b*) *Multiphase faults.* These are described in (1)–(4) below.

(1) *Phase-to-phase short circuits*—for bus-bars short-circuited together (zero-impedance connection) either the modified admittance equations can be solved subject to a number of additional constraints, or the admittance matrix $\boldsymbol{Y}'$ can be modified, and the number of equations reduced for solution without constraints.

In the first approach, where the number of nodes remains constant, the appropriate bus-bar voltages are constrained to be equal. To obtain the extra degrees of freedom for these constraints to be valid within a consistent set of equations, the currents on the right-hand sides of the corresponding equations must be unspecified and allowed to take their values accordingly. These currents are, at bus-bar k, for example,

$$I''_k = I'_k + \Sigma I_{SC}$$

where ΣI_{SC} is the total short-circuit current injected into bus-bar k from all other bus-bars short circuited to bus-bar k.

For a phase-to-phase short circuit occurring between bus-bars j and k, for example, the equations are solved with voltages $V_j = V_k$ and currents I''_j and I''_k being unknown where

$$I''_j = I'_j + I_{kj}$$

and

$$I''_k = I'_k + I_{jk}$$

I_{kj} $(= -I_{jk})$ is the short-circuit current passing between bus-bars j and k, I_{kj} that being injected into bus-bar j, and I_{jk} being that injected into bus-bar k.

(2) *Phase-to-phase faults via impedances*—if the faults between phases occur through impedances the equations are solved for the voltages $\boldsymbol{V}$ with matrix $\boldsymbol{Y}'$ supplemented by the appropriate admittances.

(3) *Phase-to-phase-to-earth faults*—earthed multiphase faults follow as extensions of the previous sections. If several phases are short circuited to earth simultaneously, the equations are solved with all the respective voltages having a value of zero. The total fault current to earth follows from the appropriate sums of the respective fault currents. If impedances are present in the fault, the corresponding admittances are included in the matrix $\boldsymbol{Y}'$ and the fault currents are evaluated depending on the fault impedance configuration.

(4) *Simultaneous faults*—using phase co-ordinates, any two- or three-phase fault may be considered to be a multiple fault, in the sense that more than one represented bus-bar is involved. The techniques of solution above may be applied, therefore, without restriction, to any number of simultaneous faults, regardless of their type or geographical location. As noted previously, the solution involves the solving of a set of simultaneous linear algebraic equations with or without constraints on the appropriate voltages, depending on the modifications made to the original network connection table.

(*c*) *Open conductors.* Open conductors present no difficulty other than that of introducing an extra bus-bar or bus-bars into the network, depending on the number of open circuits. The appropriate changes are made in the connection table, and the admittance or impedance matrices are modified accordingly.

The source transformation method of solution The above methods of solution may be referred to as 'distributed source methods', in which the various equivalent-current sources retain their identity and are generalisations of the existing methods of solution in the standard three-phase fault-level analysis described in Section 3.3.3.2.

An alternative approach would be to use Norton's theorem, or superposition methods, commencing from the supplemented nodal impedance matrix $\boldsymbol{Z}'$ $(= \boldsymbol{Y}'^{-1})$. The method depends on knowing the voltage drop caused by the fault current, and hence determines this current for an equivalent current source acting alone at the point of fault. The fault conditions are then obtained by superposition.[3]

3.3.5.3 Polyphase loadflow analysis

For polyphase load flow, several questions are raised which are not encountered in balanced one-line diagram analysis. If the phase voltages are unbalanced then the characteristics of the loads under such conditions should be known. An admittance matrix representation may well be better than a specified $P + \mathrm{j}Q$ demand. A further problem is the lack of knowledge of the active and reactive power distribution between the phases of the generators.

A synchronous generator model avoiding this latter difficulty and using only the total output power of the machine is, for a machine connected to the three-phase bus-bars p, q and r:

$$\begin{pmatrix} & & & -y_1 & -y_0 \\ & \boldsymbol{Y_{abc}} & & -\alpha^2 y_1 & -y_0 \\ & & & -\alpha y_1 & -y_0 \\ -y_1 & -\alpha y_1 & -\alpha^2 y_1 & 3y_1 & 0 \\ -y_0 & -y_0 & -y_0 & 0 & (3y_0 + y_{N0}) \end{pmatrix} \begin{pmatrix} V_p \\ V_q \\ V_r \\ E_a \\ V_N \end{pmatrix}$$
$$= \begin{pmatrix} I_p \\ I_q \\ I_r \\ \Sigma S^*/E^*_a + y_{N0}|V_N|^2/E^*a \\ 0 \end{pmatrix}$$

To a first approximation the term $y_{N0}|V_N|^2/E^*a$ can be neglected in comparison with the magnitude of the term $\Sigma S^*/E^*_a$.

For each generator the machine admittance matrix $\boldsymbol{Y_{abc}}$ can be added to the polyphase admittance matrix representing the network of lines, transformers and admittance loads. By any of the usual load-flow techniques these equations can then be solved iteratively. At generator bus-bars where a $P|V|$ specification is given, the value of P is the total active power output of the machine and $|V|$ can be either E_a or V_p or any other controlled bus-bar voltage. Several existing conventional (one-line diagram) load-flow programs already allow the generator terminal voltage magnitude to be unknown and to be adjusted according to the voltage value of a remote-controlled bus-bar. In principle, the same situation and adjustments characterise the phase co-ordinate load-flow analysis

3.3.6 Network power limits and stability

Steady-state a.c. power transfer over transmission links has limitations imposed by terminal voltages and link impedance. Transient conditions for stable operation are dynamic, and more complicated.

3.3.6.1 Steady-state conditions

Two typical cases concern a link transferring power from a sending-end generator at bus-bar voltage V_s to a receiving end of voltage V_r where there is either (1) a static load only or (2) a generator. Case (2) is the more important, as loss of synchronism is possible.

(1) Load stability The power taken by a static load of constant power factor is proportional to the square of the voltage. As the load power is increased the voltage falls, at first slightly but subsequently more rapidly until maximum power is attained. Thereafter both load voltage and power decrease, but the system is still stable, though overloaded. The condition could occur following the clearance of a system fault.

The load is rarely purely static: it usually contains motors. With induction motors the reactive-power requirements increase as the voltage falls, and beyond the maximum-power conditions the machines will stall, and will draw heavy 'pick-up' currents after a restoration of the voltage.

(2) Synchronous stability The receiving-end active and reactive powers in terms of V_s and V_r, and the parameters ***ABCD***, are given for a transmission link in Section 3.2.13.1. For a short line, $\boldsymbol{A} = 1\angle 0°$⇐and $\boldsymbol{B} = Z\angle\beta$, where $Z = R + jX$ and $\beta = \arctan(X/R)$, conditions shown in *Figure 3.27.*

If the resistance R can be neglected (as is often the case, especially where the link includes terminal transformers), the receiving-end active power P_r and its maximum P_{rm} become

$$P_r = (V_s V_r / X)\sin\theta\psi$$
$$P_{rm} = V_s V_r / X$$

To attain maximum active power, the receiving end must also accept a leading reactive power $Q_r, = V_r^2/\mathrm{X}$.

Interpreting the angle $\theta\psi$between V_s and V_r as that between the generator rotor (indicated by the e.m.f. E_t) and V_r, and including the appropriate generator reactance in X, the angle is now the *load angle* δ, and maximum active power transfer will occur for a load angle $\delta = \pi/2$ rad (90°⇐ electric). The relation for normal conditions is marked N in *Figure 3.60.*

Although a system does not operate under continuous steady-state conditions with a system fault, the power–angle relation is important in the assessment of transient stability. The network for which the curve is calculated is obtained by connecting a 'fault shunt' Z_f at the point of fault. The value of Z_f is in terms of Z_- and Z_0, respectively, the total impedance to n.p.s. and z.p.s. currents up to the point of fault. These values are given below for line-line (LL), single-earth (LE), double-earth (LLE) and three-phase (3P) faults, while the corresponding power-angle relations are shown in *Figure 3.60.*

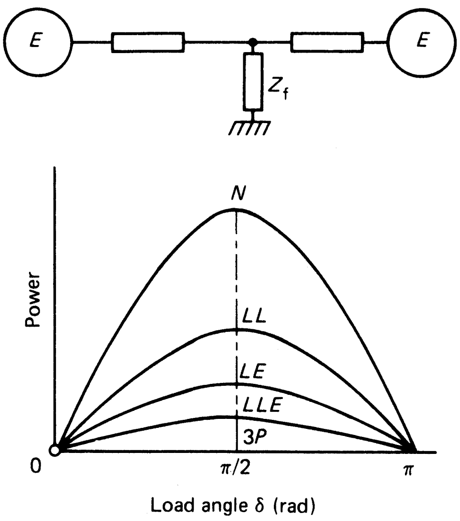


Figure 3.60 Power–angle relationships

Fault:	LL	LE	LLE	3P
Z_f	Z_{-}⇐	$Z_- + Z_0$	$Z_- Z_0/(Z_- + Z_0)$	Zero

3.3.6.2 Transient conditions

If a system in a steady state is subjected to a sudden disturbance (e.g. short circuit, load change, switching out of a loaded circuit) the power demand will not immediately be balanced by change in the prime-mover inputs. To restore balance the rotors of the synchronous machines must move to new relative angular positions; this movement sets up angular oscillations, with consequent oscillations of current and power that may be severe enough to cause loss of synchronism. The phenomenon is termed *transient instability.*

Rotor angle For a single machine connected over a transmission link to an infinite bus-bar, the simple system shown in *Figure 3.61* applies. The mechanical input P_m is, in the steady state, balanced by the electrical output for the angle δ_0 on the full-time power–angle relationship. If an electrical disturbance occurs such that the power–angle relation is suddenly changed to that indicated by the broken curve, the angle cannot

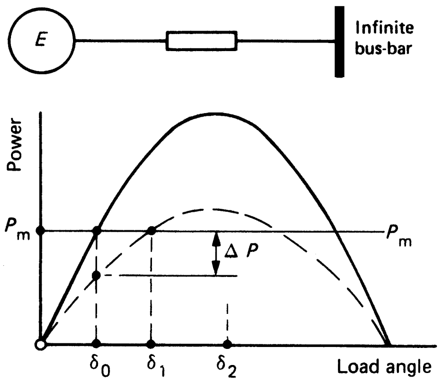


Figure 3.61 A single-machine system

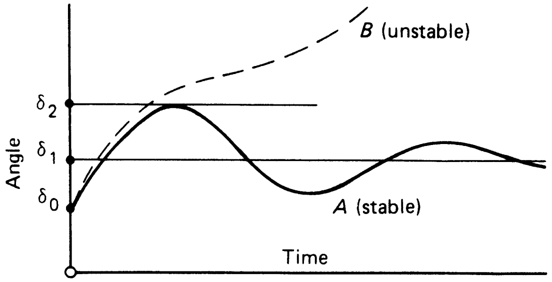


Figure 3.62 Swing curves

immediately change because of the inertia of the generator and prime-mover rotors. The electrical power drops, and a power difference ΔP appears, accelerating the rotating members towards the new balancing angle δ_1. Overshoot takes the rotor angle to δ_2. If the disturbance is not severe, the rotor assumes the angle δ_1 after some rapidly decaying oscillations of frequency 1 or 2 Hz. The angle–time relationship is that shown as curve A in *Figure 3.62*. However, if ΔP is large, the overshoot may cause loss of synchronism—the unstable curve B. A comprehensive investigation of stability thus involves the calculation of *swing* curves for the machines concerned.

Equation of motion The equation of motion for a *single* machine is

$$M(\mathrm{d}^2\delta/\mathrm{d}t^2) + K_1(\mathrm{d}\delta/\mathrm{d}t) + K_2 = \Delta P$$

where M is the angular momentum. If the damping coefficients K_1 and K_2 are ignored, the equation of motion reduces to $\mathrm{d}^2\,\delta/\mathrm{d}t^2 = \Delta\,P/M$.

A mass of inertia J rotating at angular speed ω_1 stores a kinetic energy $W = \frac{1}{2}J\omega^2$. The momentum $M = J\omega_1$ can be usefully related to the machine rating S by the *inertia constant*

$$H = W/S = \tfrac{1}{2}M\omega_1/S = \tfrac{1}{2}J\omega_1^2/S \simeq 20Jn_1^2/S$$

in which ω_1 is the synchronous angular speed (rad/s) and $n_1 = \omega_1/2\pi = f/p$ is the corresponding rotational speed (revolutions per second) for a machine with $2p$ poles operating at a frequency f. The magnitude of H (in joules/volt-ampere, or MJ/MV-A, or seconds) has the typical values given in *Table 3.11*.

A direct solution of the equation of motion is not normally possible, and a *step-by-step* process must be adopted. For this, a succession of time intervals (e.g. 50 ms) is selected and the rotor acceleration ($\mathrm{d}^2\delta/\mathrm{d}t^2$) is calculated at the beginning of each. Assuming the acceleration to be constant throughout a time interval, the angular velocity and the movement $\delta\psi$ during the interval can be found. At the end of the first interval, the new ΔP is obtained from the power–angle curve and used to calculate the acceleration during the second interval, and so on. The complete *swing curve* can thus be obtained. The method can be extended (if the relevant data are available) to include damping, changes in excitation, saliency, prime-mover governor action and other factors that affect the swing phenomenon.

Table 3.11 Inertia constants of 50 Hz synchronous machines

Machine	*n* (rev/s)	*H* (s)
Turbogenerators	50	3–7
	25	5–10
Compensators	—	1–1.25
Motors	—	2–2.25
Hydrogenerators	8.3	2–4
	5	2–3.5
	2.5	2–3
	1.7	15–2.5

Multi-machine system *A two-machine* system with a transmission link can be represented by a single machine feeding an infinite bus-bar and having an equivalent momentum $M = M_1M_2/(M_1 + M_2)$.

A *group* of machines 1, 2, …, paralleled on the same bus-bar can be treated as a single machine of rating $S = S_1 + S_2 + \ldots$, and of equivalent momentum $M = (S_1/S)M_1 + (S_2/S)M_2 + \ldots$.

For a *multi-machine* network, a separate equation of motion must be set up for each generator and a step-by-step solution undertaken. Determination of ΔP for each machine at the end of a time interval involves a comprehensive load-flow calculation by computer.

Equal-area criterion Neglecting damping, governor action and changes in excitation, the stability of a simple generator/link/infinite-bus-bar system can be checked graphically using power–angle relationships. Consider the system shown in *Figure 3.63*, with the generator operating at a load angle δ_0 on the power–angle curve P_2 with a prime-mover input P_m and both transmission links intact. A fault occurs on one link, changing the power–angle relationship to P_f and giving a power difference ΔP between P_m and the electrical output. As a result the rotor accelerates until the angle δ_s is reached, when the faulted link is switched out. The kinetic energy acquired by the rotor during this period is represented by area A. At δ_s the power–angle relationship becomes P_1 corresponding to a single healthy link. This reverses ΔP and the rotor decelerates. At the angle δ_2 such that area B

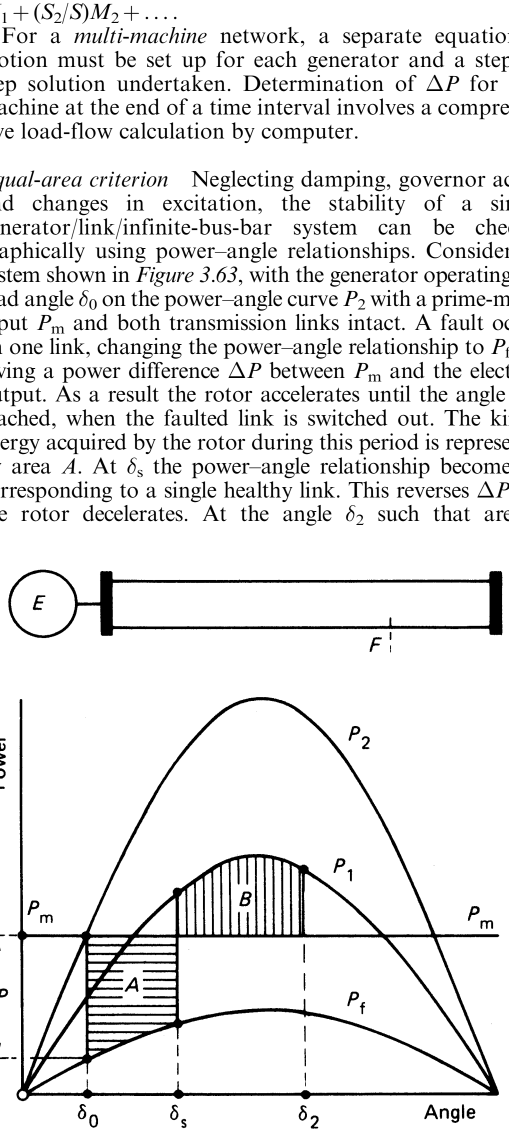


Figure 3.63 Equal-area stability criterion

(representing kinetic energy returned from the rotor) is equal to *A*, the rotor speed is again synchronous. However, ΔP is now reversed and the rotor will begin to swing back. The range of rotor-angle excursions is stable, but there is a critical value of the fault-clearance angle δ_s that, if exceeded, will result in instability. If all the power–angle relationships are true sinusoids, the critical angle can be found analytically.

Conditions other than that shown in *Figure 3.63* can be dealt with if the relevant power–angle relationships can be drawn. It is to be noted that a *swing* curve may be required to relate rotor angle to *time*, as it is the time of fault clearance (or other event)—a quantity based on the delay of switch opening—that is normally specified.

References

1 *Electrical Transmission and Distribution Reference Book*, Westinghouse Electric Corporation, Pittsburgh, USA (1950)

2 LAUGHTON, M. A., Analysis of unbalanced polyphase networks by the method of phase co-ordinates. Part 1. System representation in phase frame of reference, *Proc. IEE*, **115,** 1163–1172 (Aug. 1968)

3 LAUGHTON, M. A., Analysis of unbalanced polyphase networks by the method of phase co-ordinates. Part 2 Fault analysis, *Proc. IEE*, **116**, 857–865 (May 1969)

4 LAUGHTON, M. A. and SALEH, A. O. M., Unified phase co-ordinate loadflow and fault analysis of polyphase networks, *Elec. Power Energy Systems J.*, **2**(4), 181–192 (Oct. 1980)

5 SALEH, A. O. M., LAUGHTON, M. A. and STONE, G. T., *M*- to *N*-phase transformer models in phase co-ordinates, *Proc. IEE, Part C*, **132**, 41–48 (Jan. 1985)

6 SALEH, A. O. M. and LAUGHTON, M. A., Phase co-ordinate and fault analysis program, *Elec. Power Energy Systems J.*, **2**(4), 193–200 (Oct. 1980)

Section B

Materials & Processes

4

Fundamental Properties of Materials

G R Jones PhD, DSc, CEng, FIEE, MInstP
University of Liverpool

Contents

4.1 Introduction 4/3

4.2 Mechanical properties 4/3

4.3 Thermal properties 4/3

4.4 Electrically conducting materials 4/3

4.5 Magnetic materials 4/4

4.6 Dielectric materials 4/6

4.7 Optical materials 4/7

4.8 The plasma state 4/8

4.1 Introduction

The properties of materials relevant to electrical engineering applications are mechanical, thermal and electromagnetic in nature. The mechanical properties which are of interest relate to the strength and deformation of materials, the latter being either reversible and governed by three elastic moduli or irreversible because of plastic changes or fracture. The thermal properties relate to heat capacity, conduction and expansion. The interaction of mechanical and thermal properties with the electrical properties leads to electromechanical and thermoelectric properties.

Traditionally, the electromagnetic properties are describable in terms of three main parameters: the electrical conductivity (σ), the magnetic permeability (μ), and the electric permittivity (ϵ). The relative magnitude of these parameters determines whether the material is, respectively, electrically conducting, magnetic or dielectric. The electromagnetic transmittance of the material (i.e. the optical transparency) is governed by the refractive index which may be complex in nature and thus lead to complicated electromagnetic behaviour.

4.2 Mechanical properties

The mechanical properties of materials relate to their strength, rigidity and ductility. Materials may be deformed elastically, by which it is meant that the deformation is reversible and that the stress (force per unit area) and resulting strain (fractional size change) are proportional (*Hooke's law*). The constant of proportionality is known as the elastic modulus of which there are three types:

(1) *Young's modulus* (M) relates tensile stress and strain;
(2) *Shear modulus* (G) relates shear stress and strain; and
(3) *the bulk modulus* (K) relates bulk stress and volume strain.

During elastic deformation, not only is there a longitudinal elongation but also a cross-sectional decrease. The ratio of longitudinal to transverse strain is *Poisson's ratio*, γ.

If the materials are polycrystalline, the mechanical properties may be regarded as isotropic so that the four elastic constants may be interrelated by

$$M = 3K(1 - 2\gamma) \qquad (4.1)\Leftarrow$$

$$M = 2G(1 + \gamma) \qquad (4.2)\Leftarrow$$

The inherent strength of a material is limited by plastic deformation or fracture. Plastic deformation is a permanent deformation caused by a shearing of a few crystal planes (slip). Such deformation is affected by various types of dislocations within the material structure.

The fracture of a material is the separation into two or more parts by an applied stress and may be classified as either brittle or ductile. The former occurs after little or no plastic deformation, whereas ductile fracture occurs after extensive plastic deformation.

Materials may be strengthened but usually only at the expense of their ductility which may cause fabrication problems. Methods of increasing mechanical strength include work hardening (progressive application and removal of increasing stress), solution hardening (addition of an alloying element to produce internal strains), precipitation hardening (which involves careful heating and solution treatment), and dispersion hardening (diffusion of gases into the solid to produce hard particles).

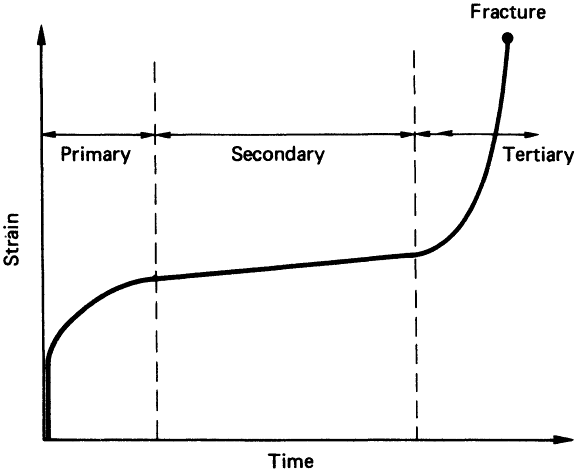


Figure 4.1 Creep behaviour under constant applied stress

The stress and strain do not necessarily occur simultaneously in real materials but rather the material continues to deform indefinitely under the influence of a constant applied stress (creep). Such creep behaviour consists of four main phases (*Figure 4.1*). An initial instantaneous strain (which may be partly elastic and partly plastic) is followed by a primary creep region having a decreasing creep rate, a secondary creep region of constant minimum flow rate and a final region of accelerating strain ultimately leading to fracture.

4.3 Thermal properties

The basic thermal properties of materials include thermal capacity, conduction and expansion. The thermal capacity (the thermal energy required to raise the temperature by one degree Kelvin) is due to the storage of energy in the motion and oscillation of atoms and electrons. Such energy may be conducted through atomic and electronic collisions between regions at different temperatures. The thermal stability of the material is governed by the magnitudes of the thermal conduction and capacity since these parameters govern the temperature rise experienced by the material on the dissipation of power within the material. Such temperature rises can lead to technical difficulties, not only because of the eventual destruction of the material, but also due to its expansion which results from the increased amplitude of oscillation of the atoms.

4.4 Electrically conducting materials

Electrically conducting materials may be identified from fundamental electromagnetic theory as those for which the conduction current density j_c is considerably greater than the displacement current density j_d where

$$j_c = \sigma E \qquad \text{(Ohm's law)} \qquad (4.3)\Leftarrow$$

$$j_d = \delta D/\delta t \qquad (4.4)\Leftarrow$$

where D is the electric flux density.

For j_c to dominate the electrical conductivity $\sigma\varsigma$needs to be large. Since the electrical conductivity is approximately given by classical theory as

$$\sigma\varsigma= (e^2/m)n\tau_c = enp \qquad (4.5)\Leftarrow$$

(where m and e are the electron mass and charge, respectively, and p is the mobility), a high concentration of conduction electrons (n) or a long time between collisions of electrons and the lattice ions of the material ($2\tau_c$) leads to the dominance of j_c.

Since both electrical and thermal conduction are due to the flow of electrons in metals, it should be anticipated that the electrical and thermal (K) conductivities should be related. The relationship is given by the *Wiedemann–Franz law*

$$\frac{\sigma\varsigma}{K}= LT \qquad (4.6)\Leftarrow$$

where L is the Lorentz Number and T is the temperature.

This has implications for the removal of Joule heat (which depends upon σ) by thermal conduction (which depends upon K).

Furthermore, the conduction of heat may be accompanied by the flow of electric current. This flow ceases when opposed by the electric field created by the resulting charge migration. The electromotive force (e.m.f.) generated is characterised by the thermoelectric power s which is the potential difference (dv) per degree temperature difference (dT)

$$s = \mathrm{d}V/\mathrm{d}T \qquad (4.7)\Leftarrow$$

Thus the potential difference between the ends of a material with thermoelectric power S_1 at different temperatures measured via junctions with a second material of thermoelectric power S_2 is governed by $(S_1 - S_2)$. This, the *Seebeck effect*, forms the basis of thermocouple operation. If, alternatively, a current is made to flow along or against a temperature gradient, the need for the conduction electrons to be in equilibrium with the crystal lattice leads to the heating of the cold end or the cooling of the hot end (*Thompson effect*), respectively. The flow of current through the junction between two materials can produce heating or cooling known as the *Peltier effect*, which can be used for temperature control.

The electrical conductivity of semiconductor materials incorporates both electrons (n_-) and holes (u_+) as carriers so that

$$\sigma\varsigma= en_-p_{n_-} + en_+p_{n_+} \qquad (4.8)\Leftarrow$$

In photoconductors the number of carriers n_-, n_+ is enhanced through the absorption of light energy to free the carriers for conduction.

For technical applications it is often preferred to use the reciprocal of conductivity, i.e. the resistivity, ρ. The resistivity of a pure, conventional conductor increases with temperature

$$\rho_\mathrm{T} = \rho_0(1 + \alpha T \ldots) \qquad (4.9)\Leftarrow$$

where $\alpha\varsigma$is the temperature coefficient of resistivity (in parts per million per degree Celsius). The temperature dependence of the resistivity of some conductors is shown in *Figure 4.2*.

The resistivity of an impure conductor is greater than that of the pure conductor (due to additional electron collisions with the impurity ions) according to *Matthiessen's rule*:

$$\rho\varsigma= \rho_\mathrm{T} + \rho_\mathrm{r} \qquad (4.10)\Leftarrow$$

In the case of a dilute alloy (*Nordheim's rule*)

$$\rho_\mathrm{r} = \mathrm{a}x(1 - x)\Leftarrow \qquad (4.11)\Leftarrow$$

where x is the impurity concentration.

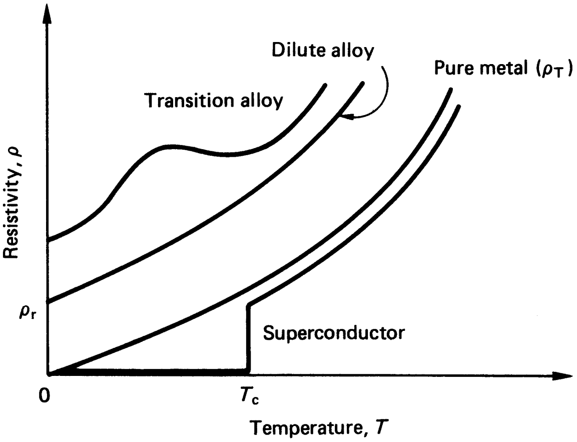


Figure 4.2 Temperature dependence of the resistivity of various types of conductors

However, for alloys of transition metals (e.g. nickel) the above rules do not apply. Instead the resistivity variation with temperature may possess a point of inflexion which has important implications for a temperature-independent resistivity over a limited range. Furthermore, some materials become superconducting below a critical temperature T_c whereby the resistivity decreases to zero. For such a superconducting state to be maintained the current density in the material needs to be kept below a critical value.

Under d.c. conditions the resistance of a bulk conductor depends on the cross-sectional area (A) and length (l)

$$R = (l/A)\rho\varsigma \qquad (4.12)\Leftarrow$$

With a.c. at an angular frequency $\omega\varsigma$the current conduction is confined to a peripheral annulus of effective width (*skin depth*)

$$S =\Leftarrow\sqrt{2\rho/\mathrm{c}\mu\omega)} \qquad (4.13)\Leftarrow$$

where $\mu\varsigma$is the permeability, so that the effective cross-sectional area A is reduced and the resistance increases with signal frequency.

Electrical conduction in thin films is also governed by the proximity of the film surface contributing to the impedance of the electron flow

$$R = sl/w \qquad (4.14)\Leftarrow$$

where w is the film width and s the sheet resistance.

4.5 Magnetic materials

The magnetic influence of materials is incorporated in electromagnetic theory via the relative permeability which relates the magnetic flux density, B, to the magnetic field strength, H, according to

$$B = \mu_0\mu_\mathrm{r}H \qquad (4.15)\Leftarrow$$

where μ_0 is the permeability of free space for which $\mu_\mathrm{r} = 1$. A measure of the magnetic influence of a material is given by the magnetic susceptibility which is the difference between the relative permeability of the material and that of free space

$$\chi_\mathrm{B} = \mu_\mathrm{r} - 1 \qquad (4.16)\Leftarrow$$

However, in material science the preference is for the material influence to be described additively rather than factorially according to

$$B = \mu_0 H + J_\mathrm{m} \tag{4.17}$$

where J_m is the magnetic moment per unit volume which is determined by the product of the magnetic moment of the elementary dipoles $\bar{\mu}_\mathrm{H}$ and their concentration N according to

$$J_\mathrm{m} \simeq N\bar{\mu}_\mathrm{H} \tag{4.18}$$

A number of physical processes may produce such elementary dipole moments and these determine the type of magnetism displayed by the material.

Diamagnetism occurs when the electron orbitals associated with the atoms of the material are distorted in the presence of an external magnetic field. The orbiting electrons may be regarded as forming circulating currents which, according to electromagnetic theory, have associated magnetic moments. On the application of a magnetic field the orbitals adjust to oppose the field. Diamagnetic susceptibility is temperature independent, of the order of 10^{-4} and is negative, implying a dilution of the magnetic flux within the material.

Paramagnetism is due to an inherent magnetic moment arising from an orbiting electron, a spinning electron and a spinning nucleus. In the latter two cases the magnetic moment may be regarded as originating from the spinning charge of the electron or proton constituting an effective circulating current. The nuclear effect is less pronounced than the electronic effect. Paramagnetic susceptibility is greater than diamagnetic susceptibility (ca. 10^{-3} to 10^{-4}), is positive and decreases with temperature (T) according to the *Curie law*:

$$\chi_\mathrm{p} = C/T \tag{4.19}$$

where C is the Curie constant.

Both diamagnetism and paramagnetism originate from the individual effects of atomic processes. However, in some materials, mainly the transition metals, the magnetic effects arising from individual atoms may be strongly coupled to give more pronounced magnetic properties to the bulk material. Such strong interatomic coupling leads to ferromagnetic, antiferromagnetic and ferrimagnetic effects

Ferromagnetism occurs when the magnetic moments of neighbouring atoms are aligned because of the strong interatomic coupling so that

$$J_\mathrm{m} = N_\mathrm{A}\mu_\mathrm{A} + N_\mathrm{B}\mu_\mathrm{B} \tag{4.20}$$

where the suffixes A and B designate the neighbouring atoms. To minimise the magnetic energy of the bulk material, regions of like-magnetic-moment alignments (*domains*) are formed with neighbouring regions having differently orientated magnetic moments.

Antiferromagnetism occurs when the magnetic moments of neighbouring atoms are misaligned so that for a pure material the bulk magnetic effects are zero:

$$J_\mathrm{m} = N_\mathrm{A}\mu_\mathrm{A} - N_\mathrm{B}\mu_\mathrm{B} = 0; \quad \mathrm{A} = \mathrm{B} \tag{4.21}$$

Ferrimagnetism arises when the neighbouring atoms in the misalignment have different magnetic moments so that there is a remnant magnetic effect in the bulk material:

$$J_\mathrm{m} = N_\mathrm{A}\mu_\mathrm{A} - N_\mathrm{B}\mu_\mathrm{B} \neq 0; \quad \mathrm{A} \neq \mathrm{B} \tag{4.22}$$

For both ferromagnetic and ferrimagnetic materials the magnetic susceptibility is enhanced and considerably greater than for both diamagnetic and paramagnetic materials. In all cases the magnetic susceptibility decreases with temperature according to

$$\chi = C/(T - T_\mathrm{c}) \tag{4.23}$$

where T_c is the *Curie point* for ferromagnetic materials and the *Neel temperature* for antiferromagnetic materials. The significance of T_c is that the material ceases to be strongly magnetic for temperatures above T_c and becomes instead paramagnetic in nature.

Incorporation of magnetic materials as part of devices in electrical networks (e.g. transformer cores), electronic circuits (e.g. magnetic pick-ups) or microwave systems (e.g. ferrite isolators) may produce power losses or waveform distortion. At power frequencies, the power losses are produced by hysteresis effects which arise from the inertia of the ferromagnetic domains in following changes in an applied magnetic field and which produces a hysteresis *B–H* loop (*Figure 4.3*). The power losses are given by the product of the energy product, the volume and the power frequency (*energy product* $= (\frac{1}{2}BH)_\mathrm{max}$). In addition, losses occur due to *eddy currents* flowing in the magnetic material. These depend on the resistivity of the material (ρ), the shape of the material (β), the dimension (d) perpendicular to the magnetic flux (B) and the power frequency (f) according to

$$\text{Eddy loss} = [(\pi B_\mathrm{max} f d^2)/(p\beta)](\text{Volume}) \tag{4.24}$$

Waveform distortion is produced by the non-linear relationship between B and H due to the hysteresis *B–H* loop because, according to Lenz's law, the output voltage from an inductive element is given by

$$V_\mathrm{out} \propto \frac{\mathrm{d}B}{\mathrm{d}t} = \mu_0 \frac{\mathrm{d}[F(H)]}{\mathrm{d}t} \tag{4.25}$$

At radio and microwave frequencies, losses may occur due to the absorption of power from the electromagnetic wave to increase the energy of the elementary atomic magnetic dipoles. Ferrite materials which are used in radio and microwave frequency devices are anisotropic in nature so that the relative permeability is replaced by a permeability tensor.

Materials which demonstrate superconductivity do so not only below a critical temperature, T_c, and critical current

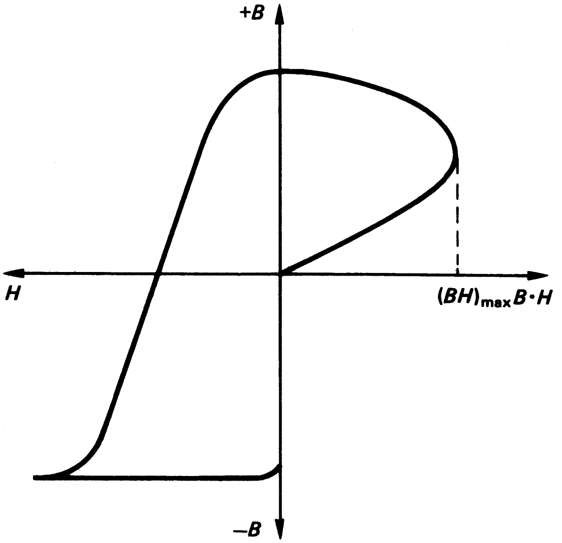


Figure 4.3 *B–H* hysteresis and the energy product $B \cdot H$

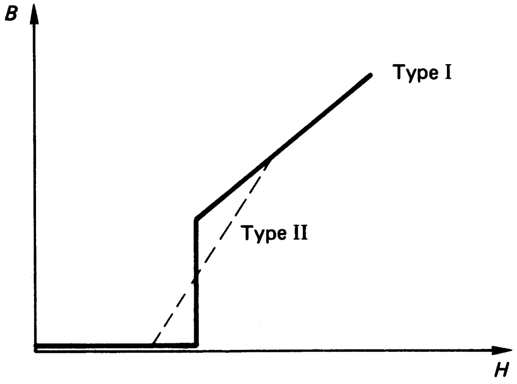


Figure 4.4 *B–H* curves for superconductors

density, but also below a critical magnetic field strength (H_c) which varies with temperature T according to

$$H_c = H_{c0}[1 - (T/T_c)^2] \qquad (4.26)\Leftarrow$$

where H_{c0} is the critical field at absolute zero of temperature.

For field strengths below H_c, superconductors become diamagnetic, implying a low magnetic flux density within the material (*Figure 4.4*). The transition to the diamagnetic state may occur abruptly at H_c (corresponding to type I superconductors which are pure materials) or more gradually (corresponding to type II superconductors which are more common and include impure materials).

The ratio of the field penetration into the material (the *London penetration depth*, Λ) to the growth in superconductivity (the *coherence length*, ξ), which is known as the *Ginsburg–Landau constant*, determines whether a material is of type I ($\Lambda/\xi < 0.71$) or type II ($\Lambda/\xi > 0.71$). In type II superconductors, in particular, the growth in current density may lead to 'flux jumping' whereby flux lines are dislodged leading to power dissipation which causes heating and hence loss of superconduction.

4.6 Dielectric materials

In electromagnetism, dielectric properties are incorporated via the relative permittivity ϵ_r which relates the electric flux density, D, to the electric field strength, E, according to

$$D = \epsilon_0 \epsilon_r E \qquad (4.27)\Leftarrow$$

where ϵ_0 is the permittivity of free space against which the permittivity of a dielectric is normalised. Thus, since for free space $\epsilon_r = 1$, the magnitude of dielectric effects may be considered in terms of the dielectric susceptibility

$$\chi_E = \epsilon_r - 1 \qquad (4.28)\Leftarrow$$

Alternatively, a relationship between D and E which is often preferred in fundamental materials studies is

$$D = \epsilon_0 E + P \qquad (4.29)\Leftarrow$$

where P is the polarisation per unit volume and

$$P = Np = N\alpha E_L \qquad (4.30)\Leftarrow$$

where p is the dipole moment of each atomic/molecular dipole, N is their concentration, E_L is the local field within the dielectric which results from the electrostatic fields between ions and electrons

$$E_L = E + P/(3\epsilon_0)\Leftarrow \qquad (4.31)\Leftarrow$$

and α is the polarisability which is produced by a number of atomic or molecular processes in the dielectric material.

Electronic polarisation (α_e) is produced when the electron orbital is distorted by the presence of an electric field (*Figure 4.5*(*a*)).

Ionic polarisation (α_i) is produced when the lattice structure of the material is distorted by the electric field (*Figure 4.5*(*b*)).

Orientational polarisation α_o arises when electric dipoles already exist in the dielectric and are re-aligned by an applied electric field (*Figure 4.5*(*c*)).

Other polarisation mechanisms which occur in some materials under special conditions are ion-jump polarisation and space charge polarisation (α_s).

Consequently, the total polarisation per unit volume is

$$P = E_L \Sigma N_a \alpha_a \qquad (4.32)\Leftarrow$$

where $a =$ e, i, o or s.

The electromagnetic and materials descriptions are interrelated via the Clausius–Mosotti equation:

$$\frac{\epsilon_r - 1}{\epsilon_r + 1} = \frac{1}{3\epsilon_0}\Sigma N_a \alpha_a \qquad (4.33)\Leftarrow$$

For situations in which the electric field is time varying in a periodic manner the polarisation P may lag behind the field E

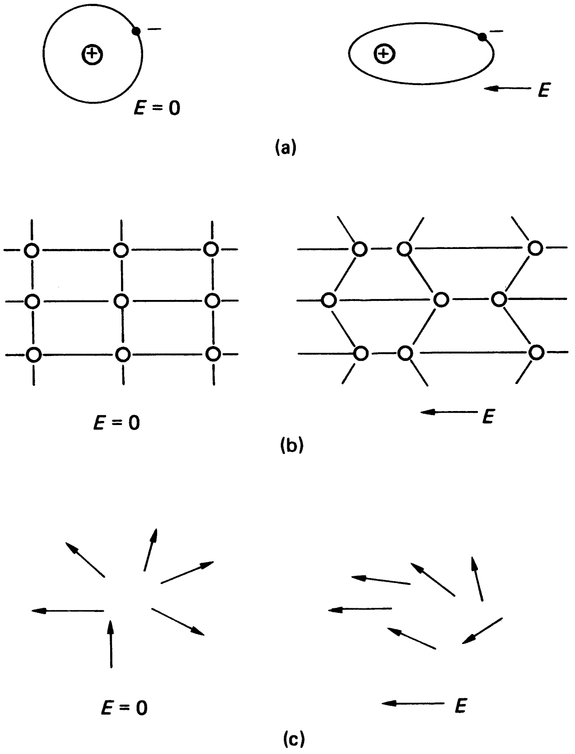


Figure 4.5 (a) Electronic polarisation (α_e). (b) Ionic polarisation (α_i). (c) Orientational polarisation (α_o)

due to the inertia associated with the physical processes described above. As a result, the relative permittivity becomes dependent on the E field angular frequency ω and on the relaxation time (τ) governing the inertia of the polarisation. As a result, the permittivity is complex in nature

$$\epsilon_r = \epsilon_\infty + \frac{N_0\alpha_0(\epsilon_\infty + 2)^2}{9\epsilon_0(1+\omega^2\tau^2)} - j\frac{N_0\alpha_0(\epsilon_\infty + 2)^2\omega\tau}{9\epsilon_0(1+\omega^2\tau^2)} \qquad (4.34)$$

$$= \epsilon' - j\epsilon''$$

Since the relaxation times of the space charge and orientational polarisations differ significantly, the effects lead to sequential changes in ϵ' and ϵ'' as the E field frequency increases. Ionic and electronic polarisation are governed by resonant effects (see Section 4.8) which also produce frequency-dependent complex permittivities (*Figure 4.6*).

Since the velocity of propagation of an electromagnetic wave depends upon ϵ'

$$\nu = c(\sqrt{\epsilon'})^{-1} \qquad (4.35)$$

the frequency dependence can lead to dispersion and waveform distortion. The imaginary component of ϵ_r leads to signal attenuation and power dissipation which accounts for the loss tangent of a capacitor

$$\tan\delta = \epsilon'/\epsilon'' \qquad (4.36)$$

The permittivity of a dielectric composed of two components whose individual permittivities, ϵ_1 and ϵ_2, are not excessively different is given by

$$\epsilon^m = \chi_1\epsilon_1^m + (1-\chi_1)\epsilon_2^m \qquad (4.37)$$

where χ_1 is the fractional volume concentration of the ϵ_1 component; m = 1 or − 1 depending upon whether the components are in parallel or series, respectively. For a random distribution of components such as in ceramics

$$\ln\epsilon = \chi_i \ln\epsilon_1 + (1-\chi_1)\ln\epsilon_2 \qquad (4.38)$$

The behaviour of anisotropic dielectrics requires the introduction of a susceptibility matrix to relate the polarisation to the electric field:

$$\begin{bmatrix} P_x \\ P_y \\ P_z \end{bmatrix} = \begin{bmatrix} \chi_{11} & \chi_{12} & \chi_{13} \\ \chi_{21} & \chi_{22} & \chi_{23} \\ \chi_{31} & \chi_{32} & \chi_{33} \end{bmatrix} \begin{bmatrix} E_x \\ E_y \\ E_z \end{bmatrix} \qquad (4.39)$$

which is simplified for uniaxial crystals by the relationship

$$\chi_{11} = \chi_{22} \neq \chi_{33} \qquad (4.40)$$

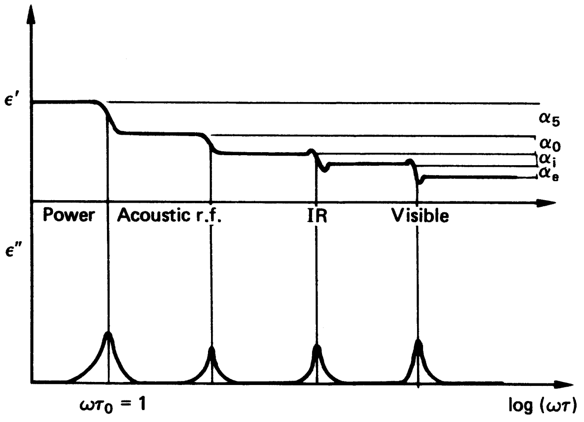


Figure 4.6 Variation of real and imaginary parts of the permittivity. (IR – infra-red; r.f. – radiofrequency)

Dielectrics which have an inherently asymmetric crystalline structure may be piezoelectric in nature whereby electrical polarisation is produced in response to a mechanical stress, and *vice versa*. (The effect is to be distinguished from electrostriction, whereby the mechanical strain direction is independent of field polarity.) Some piezoelectric materials are pyroelectric (spontaneously polarised in the absence of an E field and dependent upon heating), whilst some pyroelectric materials are ferroelectric (spontaneously polarised in the absence of an E field and polarisation switched in direction by E). The behaviour of piezoelectric materials is governed by

$$D = \epsilon_0\epsilon_r E + qT \qquad (4.41)$$

$$s = M^{-1}T + qE \qquad (4.42)$$

where s is the mechanical strain, T is the stress, M is Young's modulus and q the piezoelectric modulus. Such materials have important implications for electromechanical transduction which, in addition to their electronic applications, have power-orientated uses for spark ignition and vibration generation.

The electrical resistance of dielectrics is extremely high so that their insulating properties are enhanced. The equivalent resistance is composed of three components which are the bulk dielectric resistivity, the surface resistivity and an effective resistivity representing the imaginary part of the complex permittivity. The currents which flow in dielectrics are two-fold (equations (4.3) and (4.4))—the conduction current, j_c, being less than the displacement current, j_d. The insulating nature of dielectrics may be destroyed if the voltage across the dielectric exceeds a critical value. Such failure may occur via three processes: (1) thermal breakdown occurs when the Joule heating of the dielectric exceeds the thermal conduction; (2) dielectric breakdown occurs due to avalanche ionisation in the dielectric; and (3) gas breakdown occurs via voids in solid dielectrics producing chemical degradation of the dielectric. Since the breakdown voltage of the latter is less than that of the dielectric, gas occlusion into solid dielectrics leads to significant degradation of the electrical insulation properties.

4.7 Optical materials

The basic interaction which produces the optical properties of materials is that between the electromagnetic field of light and the electrical charges within the material.

Propagation of light through a transparent dielectric is governed by the refractive index η which is related to the relative permittivity according to

$$\eta = \sqrt{\epsilon_r} \qquad (4.43)$$

The signal frequency dependence of ϵ_r (equation (4.34)) produces dispersion which leads to signal distortion.

For transparent dielectrics which are anisotropic, the refractive index depends upon the plane of polarisation of the light and the material is *birefringent*. The implication is that the propagation velocity thus depends on the plane of polarisation of the light.

Birefringence may be induced in certain dielectrics through the application of either an external electric field (*electro-optics*) or mechanical stress (*photoelasticity*). Electro-optical effects occur due to the E field dependence of the susceptibility

$$P = \epsilon_0\chi(1 + aE + bE^2)E \qquad (4.44)\Leftarrow$$

Since $1 \gg a \gg$ b, high electric fields (e.g. high-power lasers) comparable to interatomic cohesive forces are needed for the effects to be observable. The quadratic dependence of the polarisation upon the E field corresponds to the *Kerr effect* and leads to the science of non-linear optics. Situations where the quadratic term is negligible correspond to the *Pockel effect*. Both effects may be utilised for power system voltage monitoring using optical fibre sensing systems.

Photoelastic effects arise through the dependence of the polarisation on mechanical stress which follows from the piezoelectric equations (equation (4.41)).

Some transparent materials also exhibit the *Faraday effect*, whereby the plane of polarisation of the optical signal is rotated through an angle $\theta\varsigma$in proportion to an applied magnetic flux B according to

$$\theta\varsigma= VBl = \frac{2\pi\varsigma}{\lambda\varsigma}(\eta_1 - \eta_2)l \qquad (4.45)\Leftarrow$$

where V is the Verdet constant, l the path length, $\lambda\varsigma$the optical wavelength, and η_1 and η_2 are the refractive indices.

Unlike the electro-optical effects, the magneto-optical effect leads to the rotation of the plane of polarisation being independent of the propagation direction. (The Faraday effect is also observable in certain ferrites at microwave frequencies.)

Reflection of light from the surface of a material occurs due to the incident light polarising charges near to the surface and a re-radiating of light from the oscillating dipole. The reflection coefficient for light at normal incidence depends on the refractive index according to

$$\mathrm{R} = [(n-1)/(n+1)]^2 \qquad (4.46)\Leftarrow$$

Light absorption occurs due to ionic polarisation (infrared wavelengths), electronic polarisation (visible–ultraviolet range) and excitation of conducting electrons. Ionic and electronic polarisation give rise to narrow-band absorption, whilst excitation of conducting electrons provides a sharp edged absorption band extending below a critical wavelength.

Light scattering occurs due to atomic-sized particles (*Rayleigh scattering*, $\alpha\lambda\bar{\varsigma}^{4}$) and micrometre sized irregularities (*Mie scattering*).

Light emission (*luminescence*) occurs via *fluorescence* (emission within 10 ns of energy absorption) or phosphorescence (emission delayed in some cases up to hours).

Energy to produce luminescence may be supplied by higher energy photons (photoluminescence), electron beams (cathodoluminescence), oscillating electric fields (electroluminescence), or by heating following activation by the above (thermoluminescence).

4.8 The plasma state

The plasma state may be formally defined as an ionised material in which the concentrations of ions and electrons are approximately equal and are relatively high. Examples of such plasmas are high-temperature gases such as those occurring in electrical discharges (e.g. electric arcs and high-power gas lasers), the ionosphere, or in certain semiconductors.

The electromagnetic behaviour of such materials is governed by the plasma frequency (ω_p) which is the natural frequency of the collective motion of the charges forming the plasma

$$\omega_\mathrm{p} = \frac{Ne^2}{\epsilon_0 m} \qquad (4.47)\Leftarrow$$

where N is the number density of the electrical charges. The significance of this parameter is that it governs the refractive index of the plasma according to

$$n^2 = 1 - (\omega_\mathrm{p}/\omega)^2 \qquad (4.48)\Leftarrow$$

where $\omega\varsigma$is the angular frequency of the electromagnetic wave. Thus the interaction of electromagnetic waves with a plasma depends upon the wave frequency. For $\omega < \omega_\mathrm{p}$, n is imaginary which implies that such waves are reflected. Thus waves of different radiofrequencies are reflected by different ionospheric layers; metals and semiconductors transmit at short electro-magnetic wavelengths and reflect longer wavelengths. The electromagnetic behaviour of such plasmas is, therefore, more complicated than that of conventional conductors, dielectrics or magnetic materials.

5

Conductors and Superconductors

A G Clegg MSc, PhD
University of Sunderland
(Sections 5.1 and 5.2)

N G Dovaston BSc, PhD, MPRI.
University of Sunderland
(Section 5.1.3)

Contents

5.1 Conducting materials 5/3
  5.1.1 Copper 5/3
  5.1.2 Aluminium 5/4
  5.1.3 Carbon 5/5

5.2 Superconductors 5/9
  5.2.1 Low temperature superconductors 5/9
  5.2.2 Stabilisation of magnet conductors 5/10
  5.2.3 Applications of superconducting magnets 5/10
  5.2.4 Power transmission 5/11
  5.2.5 Electronic devices 5/11
  5.2.6 High temperature superconductors 5/12

5.1 Conducting materials

5.1.1 Copper

A major part of the world's production of copper is used in the unalloyed form, mainly in the electrical industries. Copper has the highest electrical and thermal conductivity of the common industrial metals with good mechanical properties, resistance to corrosion, easy jointing, ready availability and high scrap value.

5.1.1.1 Conductivity

Pure copper has a volume resistivity at 20°C of $1.697 \times 10^{-8}\ \Omega$-m, lower than any known material except silver. In 1913 the International Electrochemical Commission established the International Annealed Copper Standard (IACS) by which the conductivity of all other grades and purities of copper and its alloys should be measured. The standard chosen was an annealed copper wire of length 1 m and cross-sectional area 1 mm^2, having a resistance of 0.17421 Ω. The corresponding volume resistivity at 20°C was assigned at $0.17421 \times 10^{-8}\ \Omega$-m representing 100% IACS. The percentage IACS for any other material can then be calculated as

$$\%\ \text{IACS} = \frac{1.7241}{\text{Volume resistivity}} \times 100$$

Since the standard was adopted in 1913, higher purity copper is now commonly produced, explaining why electrical conductivities of up to 101.5% are frequently quoted.

The relative conductivity at 20°C of other metals compared to that of copper (=100) is silver 104, aluminium 60, nickel 25, iron 17, platinum 16, tin 13 and lead 8.

All impurities tend to lower the conductivity of copper but the worst effects are produced when solid solutions are formed. Precipitation by heat treatment can sometimes be used to minimise these effects. The worst impurities include phosphorus, arsenic, antimony and nickel which are commonly found in some grades of copper. Silver, cadmium and zinc produce only a marginal decrease in conductivity whilst improving mechanical properties considerably. *Figure 5.1* shows the effect of a range of impurities on the electrical resistance of copper. The requirements for copper wires for various electrical purposes are given in BS 4109 and BSEN 12166. More detailed information about standards is given at the end of the chapter.

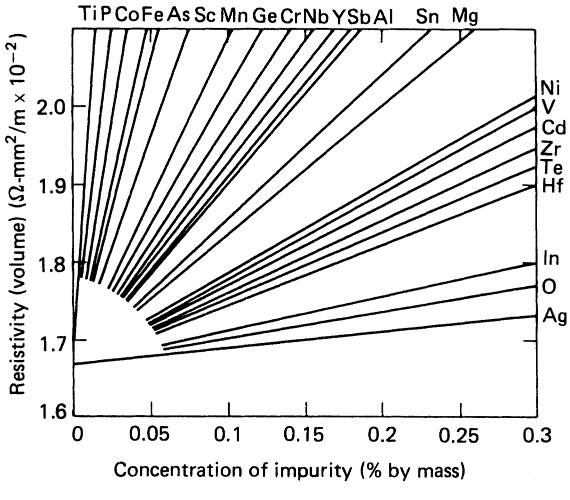


Figure 5.1 Effect of added elements on the electrical resistivity of copper. (Courtesy of the Copper Development Association)

The resistivity of copper, like that of all pure metals, increases with increasing temperature. The temperature coefficient of resistivity, α_T, for pure copper is 3.95×10^{-3}/°C. For accurate work the temperature coefficient must take account of dimensional changes due to thermal expansion. However omission of this correction for copper leads to errors of less than 0.5%.

5.1.1.2 Mechanical properties and electrical conductor applications of copper and its alloys

Copper, as cast, has a tensile strength of 150–170 MN/m^2. Subsequent rolling, drawing or other hot and cold working can raise the tensile strengths to 230 for the annealed material and to a maximum of 450 MN/m^2 for hard-drawn wire. Over this range of strengths, tensile moduli increase from 110 to 130 GN/m^2, Vickers hardness values increase from 50 to 110–130 and ductilities decrease from 45–60% to 5–20%. Cold drawn materials start to recrystallise and lose their strength at temperatures in the range 110–200°C. The increase in strength due to cold working is associated with a loss of conductivity. For hard-drawn material of tensile strength T in the range 300–400 MN/m^2, there is a loss of conductivity of 0.007T% relative to IACS copper.

For low-strength, low temperature applications, fire refined tough pitch (TP) copper is used. Small amounts of oxygen (0.04–0.05%), present as finely dispersed copper oxide, produce some strengthening and act to concentrate harmful impurities such as bismuth, preventing them from forming brittle intergranular films. Because oxygen has negligible solubility in the copper matrix (less than 0.002%) the conductivity is only slightly reduced (less than 1% IACS). Tough-pitch copper is unsuitable for intricate castings or applications where gas welding or brazing is involved. Reaction of oxides with the gas flame leads to hydrogen embrittlement. For these applications, the slightly more expensive oxygen free high conductivity copper (OFHC) containing less than 0.001% oxygen, is used. These two materials are most widely used for wire and strip conductors, for windings of a.c. and d.c. motors and generators and for transformers. Heavier gauge bar, strip and channel section material is used for bus-bars. Large quantities of high-conductivity wire and strip are used in telephone and power cables and as the outer sheathing of mineral insulated copper clad cable for fire and abrasion resistant applications.

Alloying of copper with cadmium, chromium, silver, beryllium and zirconium is used to improve mechanical properties and resistance to wear, especially at high temperatures. Such improvements are always at the expense of some increase in resistivity. Hardening of contacts by dispersions of refractory oxides improves resistance to ablation, fusion and wear.

Copper alloys containing 0.7–1.0% cadmium have greater strengths under both static (up to 750 MN/m^2) and alternating stresses and greater resistance to wear, making them useful for contacts and telephone wires, although there is little improvement in strength retention at elevated temperatures. Conductivity is between 80 and 97% of that of IACS material, depending on the degree of cold work.

Alloys containing 0.77% chromium can be heat treated to retain their enhanced hardness and tensile strength (up to 480 MN/m^2) even after exposure to more than 1000 h at

340°C when the tensile strength of OFHC or tough–pitch copper would be reduced to 170–200 MN/m^2. Both tensile strength and conductivity depend on the heat treatment, high-temperature (solution) treated materials having typical values of 230 MN/m^2 and 45% IACS while for annealed (precipitation hardened) materials the values are 450 MN/m^2 and 80%. These materials are used in electrical engineering for welding electrodes and for light current-carrying springs. Copper-beryllium and copper-zirconium alloys have similar properties and applications with superior notch-fracture resistance.

For applications where the highest conductivity is essential and enhanced creep strength at high temperature is required (e.g. for the rotor conductors in large turbogenerators and for components which have to be tinned, soldered or baked during fabrication), copper alloys containing up to 0.15% silver are used. These have the same conductivity as IACS copper and retain their mechanical properties to 300°C.

Alloys containing tellurium (0.3–0.7%), sometimes with small amounts of nickel and phosphorus, have machining properties approaching those of free-cutting brass and retain their tensile strength (275 MN/m^2) to 315°C with improved oxidation resistance. Tellurium additions alone produce only small reductions in conductivity as the solubility of tellurium in copper is only about 0.003% at 600°C. Copper-sulphur (0.4%) and copper lead (0.8%) alloys are also finding application because of their easy machining and electroplating properties.

For castings, electrolytically refined copper is sometimes used as the raw material, but more commonly, deoxidised tough-pitch copper is employed. The usual deoxidant for copper is phosphorus which produces large increases in resistivity (as little as 0.04% will reduce the electrical conductivity to about 75% of that for pure copper) and hence more expensive deoxidants such as silicon, lithium, magnesium, beryllium, calcium or boron are required. Cadmium-copper and chromium-copper alloys are also used for castings.

5.1.1.3 Other applications of copper and its alloys

Springs The material selected depends on whether the spring itself is required to carry current. For low conductivity spring materials, phosphor bronze (3.75–6.75% Sn, 0.1% P) and nickel silver (10–30% Ni, 10–35% Zn) are widely used and have conductivities in the ranges 12–27% and 6–8% IACS, respectively. Beryllium-copper alloys are used in more critical applications; a 2% beryllium alloy can be cold worked and heat treated to give tensile strengths of 1000–1500 MN/m^2, while retaining a conductivity of 25–35% IACS. With other compositions and heat treatments even higher conductivities can be obtained.

Resistance and magnetic materials Copper alloys containing manganese (9–12%), aluminium (0–5%) and nickel (0–10%) are widely used as resistance materials because they have high resistivities ($38–48 \times 10^{-8}$ and $44–175 \times$ ⇐ 10^{-8} Ωm, respectively), low or zero temperature coefficients of resistance (-0.03 to $+1.4 \times 10^{-3}$/°C over the range 0–100°C) and low thermal e.m.f.s relative to copper (see page 2/9 section 2.2.2.6). Copper is also a constituent of many magnetic materials, both hard and soft.

Contacts Palladium-copper and silver-copper (7.5–50% copper) alloys are suitable for light current applications. For heavy-duty contacts, sintered copper-tungsten materials are tough and durable; similar materials are used for electric discharge machining electrodes.

Electroplating alloys Copper alloys containing nickel (7–23%) and zinc (10–35%), known as German or nickel silvers, are widely used for electroplating as they form durable corrosion-resistant coatings with reflectivities equivalent to that of standard silver.

Heat-transfer materials Copper alloys are extensively used in nuclear and fossil fuel steam plant for electric power generation for heat exchangers such as feedheaters and condensers, although there is an increasing tendency towards the substitution of mild steel for the former and titanium for the latter. The principal condenser alloys are Admiralty brass (70% Cu, 30% Zn with small arsenic and zinc additions to prevent dezincification) and aluminium brass (76% Cu, 2% Al, 30–10% Ni with 1–2% Fe, 1–2% Mn). These are used where improved corrosion resistance is required. In more erosive conditions caused by suspended solids in the cooling water, cupronickel alloys (70–90% Cu, 30–10% Ni, with 1–2% Fe, 1–2% Mn) are used at higher initial capital cost and with some penalty in heat transfer. For feedheater applications cupronickel alloys are most widely used in older plant.

Memory-effect alloys Some copper-zinc-aluminium alloys (8–14% Al, 0–14% Zn, 0.3% Ni) have the useful property of existing in two distinct shapes above and below a critical transformation temperature (within the range -70°C to $+130$°C) due to a structural change. These alloys are finding applications in temperature-sensitive actuating devices, replacing bimetallic strip or thermistor controlled relay devices.

Superconducting alloys Dendritic copper-niobium alloys (20–30% Nb) may be plated or diffused with tin and reacted *in situ* to form fine superconducting Nb_3Sn filaments intimately incorporated in a copper matrix. Such materials have high critical current densities in the 8–14 K temperature range and an improved tolerance to strain.

5.1.2 Aluminium

Aluminium and its alloys are widely used in the electrical industry because of their good electrical and thermal conductivity, generally excellent mechanical properties and corrosion resistance, ease of fabrication, low density and non-magnetic properties. World production of aluminium has steadily increased and has overtaken that of copper which it has replaced in many electrical applications. Weight for weight aluminium is a cheaper and better conductor than copper.

5.1.2.1 Resistivity

Pure aluminium has a resistivity of 2.64×10^{-8} Ωm at 20°C with a mean temperature coefficient over the range 0–100°C of 4.2×10^{-3}/°C. Thus it has about 66% of the conductivity of pure copper or 66% of that of the International Annealed Copper Standard (IACS) at 20°C. The density of aluminium is 2.7 compared with 8.9 for copper and hence, weight for weight, the conductivity of aluminium is 2.1 times that of copper and exceeds that of all known materials except the alkali metals.

5.1.2.2 Mechanical properties and electrical conductor applications of aluminium and its alloys

The mechanical properties of aluminium depend upon the purity as well as the degree of cold work. Ordinary aluminium of commercial purity contains 99.2% aluminium whilst superpurity grades contain 99.99%. For special purposes aluminium may be zone refined to give a purity of 99.9995%. Tensile strengths and hardness are at a minimum in the high purity 99.99% annealed grades giving values of 59 MN/m^2 and 15 Brinell hardness, respectively. Commercial purity, 99.2% aluminium has tensile strength of about 91 MN/m^2 and Brinell hardness of 22. Cold working will increase these values by a factor of 2. Low strength and hardnesses mean that in most applications aluminium is used in the alloyed form.

Overhead-line conductors All-aluminium alloy conductors (AAAC), made from alloys containing 0.3–1.0% silicon, 0.4–0.7% manganese and small amounts of iron and manganese, are being used increasingly for high-voltage ($\geq$400 kV) overhead lines. These alloys can be precipitation hardened to give tensile strengths in the range 310–415 MN/m^2, while retaining conductivities of 52–60% of that of IACS copper.

All-aluminium (AAC) and steel-cored aluminium conductors (ACSR) for overhead lines are constructed to IEC 60207,9 (BS 215) with a maximum permitted resistivity for the aluminium of 2.83×10^{-8} Ωm at 20°C and a constant-mass temperature coefficient of resistivity of 4.03×10^{-3}/°C. All-aluminium alloy conductors (AAAC) are constructed to IEC 60208 (BS 3242) which specifies a maximum resistivity of 3.28×10^{-8} Ωm at 20°C and a constant-mass temperature coefficient of resistivity of 3.6×10^{-3}/°C.

Reinforced cables Where cables are steel reinforced the requirements for the steel core wires are given in IEC 60888 (BS 4565) and for aluminium in BSEN 1172 and BSEN 1652-4. In order to prevent galvanic corrosion, all steel wires and fittings must be galvanised—suitable specifications are given in IEC 60888 (BS 4565) and ISO 1459-61 (BS 729). Complete greasing of the central steel core wires and of the inner aluminium wires is necessary to minimise corrosion in marine or industrially polluted environments.

Because of the widely different mechanical properties of aluminium and steel, special jointing and anchoring systems have been designed. These generally take the form of compression fittings applied using hydraulic or mechanical crimping tools. Aluminium conductor alloy reinforced (ACAR) cables overcome many of the shortcomings of steel-reinforced cables in terms of corrosion and jointing, they offer a good combination of conductivity, strength and weight.

Bus-bars Bus-bars are commonly constructed from aluminium and aluminium alloys of similar compositions to those employed in overhead-cable manufacture. Specification in terms of strength is lower, permitting lower resistivity requirements of less than 3.133×10^{-8} Ωm to be achieved. Alloy materials have lower temperature coefficients of resistivity than pure aluminium so that performance in terms of current-carrying capacity improves relative to pure aluminium as the temperature increases. Aluminium bus-bars of the same current-carrying capacity as copper bus-bars have the added advantage of greater heat dissipation and only half the weight.

Jointing of aluminium or aluminium alloy bus-bar material requires more care than copper. Aluminium brazing, soldering or inert-gas welding is employed. Alternatively, mechanical jointing may be employed. Where steel or copper fittings are involved, precautions similar to those used with steel-reinforced cables should be used to avoid galvanic corrosion. Sealing with greases, tapes, mastics or paints containing corrosion inhibitors will reduce the need for maintenance. Anodised or aluminium clad alloys are often employed in severe environments.

5.1.2.3 Other applications of aluminium and its alloys

Switchgear, generators, motors and transformers Aluminium casting alloys containing 0–12% silicon with small additions of iron, copper, magnesium and nickel are widely used for switchgear, transformers, rotors for small induction motors and generator castings where lightness is essential. Tensile strengths range up to 345 MN/m^2 and conductivities to 50% of that of IACS copper.

Aluminium strip and wire are also used for transformer windings and for field coils in starter motors for automotive applications.

Heat sinks and heat exchangers Aluminium and its alloys with copper, silicon, magnesium and other metals are widely used as heat-sink and heat-exchanger materials. The thermal conductivity of pure aluminium is 2.38 W/m/K, which is 56% of that of pure copper. Again on an equal weight basis, the thermal conductivity of aluminium is almost twice that of copper.

Capacitors High-purity aluminium is widely used as a capacitor-plate material. High-capacity, low-volume capacitors can be fabricated by anodising aluminium electrolytically; the oxide layer produced forming the dielectric. Alternatively, plastic film materials may be aluminised on both sides forming the electrodes of film capacitors for a range of electronic applications.

5.1.2.4 Standards

Each country has in the past had its own standards for materials. Over the past twenty years or so there has been a movement towards international standards, which for electrical materials are prepared by IEC (International Electrotechnical Commission). More general standards are prepared by ISO (International Standards Organisation). When an international standard is produced the member countries copy this standard and issue it under their own covers. This now applies to BSI, ASTM, DIN, and the French standards organisation. Where appropriate, standards are also issued as applying to all the European Union Countries.

Leading standards for electrical materials are shown as follows in *Table 5.1*.

5.1.3 Carbon

Carbon occurs in nature in two crystalline forms, graphite and diamond, the former being thermodynamically more stable at ambient temperature and pressures. Synthetic carbons prepared from coke or other precursors may be predominantly graphitic or amorphous depending on the conditions of manufacture. Diamonds may be prepared from graphite using high-pressure techniques but the resulting products are usually suitable only for industrial applications; new techniques have been reported for the production

Table 5.1 International and British Standards for Copper and Aluminium

Copper		
BS 1432,3 & 4: 1970–1991	IEC 60356	Specifications for copper for electrical purposes.
BS 4109: 1991	BSEN 12166 (replaces BS 2873) ASTM B1 – B3	Specifications for copper for electrical purposes. Wire for general electrical purposes and for insulated cables and flexible cords. (Includes wire tables with diameter, resistance (Ω/km), and mass (kg/km).)
BS 7884: 1997		Specifications for copper and copper-cadmium stranded conductors for overhead electric traction and power transmission systems.
	ASTM B258	Standard nominal diameters and cross-sectional areas of AWG sizes of solid round wires used as electrical conductors.

Further information may be obtained about copper and its alloys from The Copper Development Association, Verulam Industrial Estate, 224 London Road, St Albans, Herts, AL1 1AQ, UK. Phone 0172 773 1200, or The Copper Development Association, 260 Madison Avenue, New York, NY 10016, USA. Phone 212 251 7200

Aluminium		
BS 215 parts 1 & 2: 1970–1985	IEC 60207	Aluminium stranded conductors. (BS part 1 includes physical properties and wire tables including diameter, resistance (Ω/km) and breaking load.)
BS 2627: 1985	ASTM B396 & B398	Wrought aluminium for electrical purposes—wire.
BS 3242: 1970	IEC 60208	Aluminium alloy stranded conductors for overhead power transmission.
BS 3988: 1970	IEC 60121	Specification for aluminium for electrical purposes—solid conductors for insulated cables.
BS 6360: 1991	IEC 60228 and 60228A	Specifications for conductors in insulated cables and cords.
BS 4565: 1990	IEC 60888	Specifications for galvanised steel for aluminium conductors, steel reinforced.
BSEN 1301, 2 & 3		Aluminium and aluminium alloys. Drawn wire. Inspection, mechanical properties and tolerances.
BSEN 60889	IEC 60889	Hard drawn aluminium for overhead line conductors.
BS 729: 1994	ISO1459-61	Specification for hot dip galvanised coatings, on iron and steel articles.
BSEN 61232	IEC 61232	Aluminium-clad steel wires for electrical purposes.

Further information may be obtained about aluminium and its alloys from The Aluminium Federation Ltd, Broadway House, Calthorpe Road, Birmingham, B15 1TN, UK. Phone 0121 456 1103, or Aluminium Association, 900 19th Street, NW, Washington, DC 20006, USA, phone 202 862 5178.

of gem-quality stones by vapour deposition onto small specimens.

The engineering applications of carbon exploit the following properties:

(1) it is thermally stable and in the absence of an oxidising atmosphere retains most of its mechanical strength to a temperature of 3500°C at which, under atmospheric pressure, it sublimes;
(2) its oxides are gases and leave no surface film;
(3) it is dimensionally stable, does not swell in water and can be machined to close tolerance;
(4) it has a low expansion coefficient and low density (each about one-quarter that of steel);
(5) it is a good conductor of heat, with a high specific heat and a great resistance to thermal shock;
(6) it is not wetted by molten metals;
(7) it is chemically inert;
(8) it is self-lubricating under normal atmospheric conditions; and
(9) it is electrically conductive and has high contact resistance with metals.

Electrical resistivity is lowest in the graphitic materials, which in the form of graphitised carbon black finds wide use in the manufacture of electrical carbons. Resistivity is dependent upon origin, composition and manufacture; typical values are in the region of 10^{-5} Ωm.

5.1.3.1 Carbon brushes

Current collection in moving contacts forms one of the most important electrical applications of graphitic carbon. Brushes must carry heavy current without excessive overheating or wearing of the parts contacted. Low friction, high contact resistance and infusibility are amongst the most desirable characteristics. *Table 5.2* gives data on some widely used brush grades.

Commutators Contact resistance is of major importance in commutation. High contact resistance reduces losses due to high circulating currents, minimising problems due to over-heating and sparking. Commutator machines use non-metallic 100% carbon brushes, except in low-voltage applications where metal–graphite grades are necessary to reduce voltage drop and hence losses. Current density and surface speed must be considered in selection of suitable grades. Low-friction materials and good design serve to reduce chatter, ensuring good contact between brush and commutator bars. Friction characteristics may be affected by a number of factors, including the chemical and mechanical properties of the commutator metal as well as humidity, temperature, contaminants and abrasives. The ash content of the brush material is important in determining friction and abrasion characteristics and in some types of commutator may serve to wear down insulation between commutator segments. The performance of a brush on a commutator machine is influenced by its position on the commutator, i.e. the circumferential and axial stagger.

Slip-rings When commutation phenomena are absent, brush grades with low contact drops, particularly those in the metal–graphite class, can be employed. Again, current density and surface speed must be taken into account; at the highest surface speeds it is necessary to select a grade from the natural graphite class. For applications in instrumentation, such as pick-ups for thermocouples and strain gauges, silver–graphite (SM) brushes on pure silver rings are needed to give minimal and constant contact drop.

Wear The rate of wear of brushes is not directly related to the hardness of the brush material but more to the grade, the current density and quality of mechanical features (surface roughness, eccentricity, stability of brush holders and brush arms and angle of brush relative to the pick-up surface).

Spring pressure The pressure specified for a particular brush grade is determined from laboratory and field tests to give the optimum performance and life, and should be carefully adhered to. Pressures are normally of the order of 14–21 $\mathrm{mN/mm^2}$, but in conditions of considerable vibration (as in traction and in aircraft) pressures may be as high as 28–50 $\mathrm{mN/mm^2}$.

Brush materials Brushes using lampblack base are widely used for medium- and high-voltage machines, d.c. motors and generators and universal motors. These materials possess high contact resistance which is necessary for good commutation. Graphitised petroleum coke is used for non-metallic slip-ring brushes where high contact resistance is not required. Metal powders, copper or silver, blended with graphite may be used for slip-ring or low-voltage commutating machinery. The bond in most non-metal brush grades is carbon from pyrolysis of the coal-tar pitch binder. Some graphite grades are bonded with synthetic resins. Final properties are controlled by use of impregnating agents combined with heat treatments which are designed to control contact resistance, filming action and friction characteristics.

Brush design For standards of brush design, reference should be made to BS 4999 and to IEC Publication 136, on which the British Standard is based.

5.1.3.2 Linear current collection

This is the reverse of the machine condition in that the conductor is stationary and the collector (equivalent to the brush) is moving. Current collection of this type is found on rail and trolley traction systems, cranes and line process plants. In many applications the collector and conductor are open to aggressive environments including adverse weather conditions. Arcing which results from such situations roughens the conductor and accelerates wear in the collector. Carbon as a contact material provides the real benefits of a very low rate of conductor wear, long collector life and good contact stability. The carbon element, which needs no applied lubrication, is very hard and strong. Carbon shoes are in competition with brass or bronze wheels and copper shoes. Radio interference is minimised by use of carbon shoes; the grades most generally used are Link CY and Link MY, the latter being metal impregnated.

5.1.3.3 Carbon contacts

Electrical contacts include components for switches, circuit-breakers, contactors, relays and sliding contacts. The properties of carbon and graphite of importance in such applications include: self-lubricating and filming characteristics; non-welding; unwetted by molten metals; and high thermal stability. These properties make carbon the ideal contact material, particularly where arcing may occur under severe contact-bounce conditions. Carbon requires higher voltages to maintain arcing conditions than do many metals and the effects of arcing on carbon are less significant as carbon sublimes rather than melts at temperatures in excess of 3500°C. The material does not tarnish, giving constant contact resistance in the absence of any surface films.

Selection of carbon contact grade for such applications is in accordance with the required contact resistance, the metal-impregnated and metal–graphite classes giving the lowest and the carbon class the highest values.

5.1.3.4 Resistance brazing and welding

Using the relatively high resistance characteristics of carbon, it is possible to obtain a heating effect that can be used for the joining of metals. The major application is in resistance brazing where one component in the system is melted, or a low-melting alloy solder is introduced to complete the bond.

The absence of melting of carbon at high temperatures prevents sticking or welding of the carbon to the workpieces and prevents distortion of the tips of the tool even at white heat under pressure. Design of the carbon tips, in terms of shape and area of contact, is used to control the amount of heat generated in the joint. The lower strength of carbon, compared to the alternative water-cooled metal electrode, limits its use in resistance-welding and spot-welding applications; plain carbon or electrographites are used, the latter generally giving better life.

5.1.3.5 Arc welding

Electrodes of carbon are very suitable for arc welding. The weld is achieved by either a fusion process (i.e. fusion of butting edges) or by feeding a weld rod into the arc formed between the workpiece and the electrode. Carbon arcs find limited use in metal cutting, typical applications being in the cutting of risers from castings or the grooving of metal plates. Carbon arcs are widely used for brazing of thin mild-steel sheet.

5.1.3.6 Granules

Carbon granules for microphone applications are made in sizes of 60–700 μm, the largest being suitable for maximum

Table 5.2 Carbon-brush grades

Grade	*Morganite Grade coding*
C/S	For commutator/slip-ring application
C*	For traction commutator
J	Normal current density, mA/mm^2
p	Normal pressure, mN/mm^2
v	Contact volt drop per brush
μς	Coefficient of friction
u	Normal maximum operating speed

The characteristics *v*, *μς*and *u* are indicated by numbers:

Number	*Contact voltage drop* (V/brush)	*Coefficient of friction*	*Surface speed* (m/s)
1	<0.4	<0.1	20
2	0.4–0.7	0.1–0.15	30
3	0.7–1.2	0.15–0.2	50
4	1.2–1.8	>0.2	60
5	>1.8	—	>60

Grade	*C/S*	*J*	*p*	*v*	*μς*	*u*	*Properties*
Electrographite class							
EGO	C	100	18	3	2	1	Low voltage, high current
	S	115	18	3	2	3	Steel and bronze slip-rings
EG3	C	85	14	3	2	1	Small and medium industrial motors
	C*	85	20–50	3	2	2	Traction a.c. motors
EG260	C	100	21	3	2	2	Wide range of d.c. machines, and for corrosive atmospheres
	S	115	21	3	2	2	
EG8101	C	85	21	4	1	2	Difficult a.c. commutating conditions: fractional motors
EG12	C	100	18	3	3	3	For prolonged overloading
EG14D	C*	100	20–50	3	1	3	For d.c. and $16\frac{2}{3}$ Hz a.c. traction
EG17	C	100	21	3	2	4	Long life: stable current collection
EG116	C*	110	20–50	4	1	3	A.c. and d.c. traction motors
EG133	C	100	18	3	3	3	For high-rated welding generators
EG224	C	95	21	3	2	3	For rolling-mill motors
EG236	C	110	18	4	2	4	For rolling-mill motors and mine-hoist generators
EG236S	C	110	21	2	3		For industrial d.c. machines and traction generators
EG251	C	95	21	4	1	3	For peaky overloads, light-load running
EG6345	C	95	18	3	2	2	For d.c. machines in ships
EG6749	C*	95	20–50	3	2	3	For d.c. traction motors
Natural graphite class							
HM2	C	100	14	3	2	3	Low friction
HM6R	C	100	14	3	3	4	Low inertia, good collection
	S	100	14	3	3	5	Suitable for turbogenerator rings
HMl00	C	100	14	3	3	4	Good load sharing
	S	100	14	3	3	5	For helical-grooved steel rings
Resin-bonded class							
IM6	C	25	21	5	1	2	High contact drop: for d.c. and universal fractional motors
IM23	C	95	21	5	1	2	For Schrage motors
Carbon and carbon–graphite class							
A2Y	C	70	14	3	3	1	For fractional machines with flush mica
B	C	85	14	3	4	1	High conductivity, for easy commutation conditions
C4	C	65	14	3	4	1	Dense, for machines with hard mica
H1OO	C	85	84	3	1	2	Contains MoS$_2$; quiet, low friction; for car generators
PM9	C	65	50–20	4	4	1	For fractional machines with flush mica
PM70	C	40	21	4	4	1	For difficult fractional machines with recessed mica
Metal–graphite class							
CM + O	C	170	130–70	1	3	1	Low loss; for car starter motors only
CM	S	230	14	1	3	1	Very low loss; for heavily loaded slip-rings
CM3H	C	125	21	1	2	2	High graphite content: for car starters and d.c. machines up to 12 V
	S	155	21	1	2	2	
CM5B	C	115	21	2	1	2	For d.c. machines up to 50 V
	S	140	21	2	1	2	For totally enclosed slip-rings, car accessory motors
CM5H	C	115	21	2	1	2	For d.c. machines up to 30 V, induction-motor slip rings
Metal-impregnated graphite class							
DM4A	C	115	—	2	2	1	Short-period current densities up to 2300 mA/mm^2; pressures according to application; for automobile and battery vehicle motors, and contacts
DM5A	C	115	—	3	2	1	

response but developing more background noise. The smallest granules give better quality but a lower response. For normal telephone work, granules of the smaller range of sizes are employed to give good response and acceptable noise. Carbon is superior to metal powders in these applications due to high specific and contact resistance and the absence of any tendency to oxidise or tarnish. The result is a stable and reproducible performance essential to audio applications.

Telephone granular carbon is a crushed product made from petroleum coke; particle size and shape are important in determining performance characteristics. Special non-ageing grades are available.

5.1.3.7 Fibres

Fibres are the latest form in which carbon is manufactured; they possess a near-perfect orientation of the graphitic structure parallel to the fibre axis. Fibres have high strengths and elastic moduli which may be controlled by the maximum temperature attained during heat treatment. Low density confers high specific strength and stiffness of particular value in aerospace applications. Selective use of carbon fibres in composites is employed to maximise properties in desired directions. Carbon fibres are graphitic structures produced by carefully controlled pyrolysis of polyacrylonitrile or cellulose fibres under tension. More recently, petroleum or coal-tar pitch have been used.

5.2 Superconductors

Superconducting materials lose all electrical resistance below a temperature called the *critical temperature* (T_c) which is different for different materials. The phenomenon was first observed in mercury in 1911. Engineering interest in superconductors became really significant in the early 1960s when materials capable of carrying high current densities (up to $10^9\,\mathrm{A/m^2}$) in high magnetic fields (several tesla) were discovered. These opened the way for high-field electromagnets. The interest was reinforced by advances (partly spurred by superconductor development) in the technology of large-scale helium refrigeration (hundreds of watts cooling at 4 K) which could produce cold gaseous or liquid helium for cooling purposes. The discovery of the Josephson effect (1962) led to small, low-field, superconducting electronic devices (see Section 5.2.5). More recently the interest has multiplied with the discovery of the high temperature superconductors in 1986 and there are now power transformers and power cables at the prototype stage.

Two groups of materials have superconducting properties. The 'classical' or low-T_c superconductors (LTS) are metals or alloys all of which show superconducting properties below 23 K. Virtually all industrial applications use materials from this category, in particular, NbTi or Nb_3Sn. The second group are metal oxide compounds based on copper oxide (CuO_2) subunits. These are known as ceramic or high temperature superconductors (HTS) and the compound with the highest known critical temperature (HgBaCaCuO; T_c 133 K (−140°C)) belongs to this group. The coolant for LTS material is liquid helium which is liquid below 4.2 K and for HTS material the coolant is liquid nitrogen which is liquid below 77.3 K (−195.8°C).

Once cooled below the critical temperature, superconductors can carry resistanceless current up to a maximum determined by the ambient magnetic field and the temperature. If the limiting combination of current, magnetic field and temperature is exceeded, the material reverts to an electrically 'normal' condition. The temperature, magnetic field and current density are independent variables, the critical values of which characterise the material. The capability to carry large lossless currents in magnetic fields of several tesla is a primary requirement of any superconductor if it is to be used in electrical plant.

5.2.1 Low temperature superconductors

For engineering purposes there are two varieties of superconductor—surface superconductors and bulk superconductors—determined as much by the way they are used as by their intrinsic properties. Surface superconductors carry superconducting current only in a surface layer and only in low magnetic fields. The magnetic field at the surface is equal to the current per unit width of the surface. In general, the resistance of superconductors to a.c. is not zero, though small; that of surface superconductors can be virtually zero. The small power losses that do then occur arise mainly from factors such as surface irregularities.

Bulk superconductors can carry currents throughout their section. Losses under changing current are relatively high and bulk superconductors are difficult to use at power or high frequencies. The loss is effectively a hysteresis loss. Both the high-T_c superconductors and ultrafilamentary superconductors (see Section 5.2.2) may ameliorate this problem. Some bulk superconductors, often referred to as 'hard' superconductors, can carry large current densities in strong magnetic fields (see *Table 5.3*). Materials with good superconducting properties tend to have high normal resistivities.

Physicists divide superconductors into types I and II. Type I materials are perfectly diamagnetic below the critical magnetic field, they carry surface currents only and exclude all magnetic flux from their bulk. The depth of current and field penetration is determined by quantum mechanical considerations and is typically 10^{-7} m say 1000 atoms. Above the critical field type I materials lose all diamagnetic properties. Type II materials also show diamagnetism but lose the property gradually as the field is raised above a first critical field. The magnetic field penetrates the bulk in quantised flux bundles or fluxons (2×10^{-15} Wb/bundle) which can be detected with sensitive techniques. Electrical resistance returns above the critical field.

The quantum theory of superconductivity envisages a low-temperature condensed state of electrons in which they interact, through the atomic lattice, to form temporary pairs. Each member of the pair distorts the lattice in such a way as to attract the other. Because of their energy state, these pairs are not readily scattered and flow unimpeded. This mechanism is well established for metallic superconductors but the exact coupling mechanism which generates the pairs in high-T_c materials is uncertain. Without detailed prior knowledge of the electronic and atomic structure of a material, the theory is not sufficiently quantitative to predict whether or not the material will exhibit superconducting properties, let alone what the properties will be. Analogy with known superconductors has proved the best guide to new ones.

Typical low T_c materials used for engineering purposes are shown in *Table 5.3*. Niobium is the surface superconductor most commonly envisaged for use in a.c. equipment such as power cables. Lead is cheaper but not usually preferred unless for special reasons a low-field capacity is satisfactory. Bulk superconductors can be used in low fields as surface superconductors, e.g. niobium-tin (Nb_3Sn). Of the bulk superconductors, niobium-titanium is a ductile alloy,

Table 5.3 Some low T_c superconducting materials and their properties

Material	*Critical Temperature** (K)	*Upper critical magnetic field at 4.2 K†* (T)	*Critical current density (in superconductor) at 4.2 K and 5 T* (GA/m^2)
Lead	7.2	0.06	—
Niobium	9.2	0.4	—
Niobium-tin (Nb_3Sn)	18.1	23	5–10
Niobium-germanium (Nb_3Ge)	23	37	5–15

*All the superconductors listed are type II except lead (type I). Critical current densities are very dependent on the way in which the material is prepared.
†4.2 K is the boiling point of liquid helium.

which needs to be cold-worked to produce good current carrying capacity. Niobium-tin is a compound and very brittle; it is usually made by diffusion of tin into niobium, sometimes from a bronze, sometimes in ribbon form to reduce mechanical stresses during handling. The diffusion is often produced by heat treatment of the magnet coil after it has been wound.

5.2.2 Stabilisation of magnet conductors

Bulk superconductors suffer from electrothermal instability, especially when used near their critical conditions. A small increase in temperature lowers the critical current density, which can lead to a further increase in temperature, ultimately driving the material into its normal resistive state; heat generation then becomes very rapid and the whole coil becomes 'normal'. Stabilisation can be achieved in several ways. All methods use a composite conductor in which ordinary conductors, frequently of copper, are in intimate contact with the superconductor. The copper has a much lower electrical resistance than the superconductor in its normal state. The most robust stabilisation system is the so-called cryostatic method. The design provides plenty of copper in parallel with the superconductor and good cooling of the composite. It envisages some agent causing the temperature to rise and current to transfer into the low-resistance copper over a short length of conductor. The thermal conditions are designed to be such that the composite conductor will cool so that once more the superconductor element regains its superconducting capacity. This method requires coolant access to most of the winding.

A more compact scheme for coils aims to limit localised temperature rises. Magnetically induced electrical losses in the superconductor during current changes are reduced by using a composite of fine strands of superconductor co-processed in a matrix, often of copper. The diameter of the superconducting filaments can be as small as 100 nm and each composite wire may contain upwards of half a million strands. Such ultra-fine filamentary conductors have very low a.c. losses and may potentially be used at power frequencies. Although normally the superconducting filaments have diameters in excess of 1 μm, strands can be twisted and the matrix incorporates components with high resistance, such as cupronickel, to control circulating currents and losses. The need for additional metal means that current density in terms of the total conductor cross-section is up to 10 or more times lower than that in the superconductor itself. This is especially so for Nb_3Sn and Nb_3Ge. Perhaps most importantly, the winding can be encapsulated in resin to prevent frictional heating as the winding takes up mechanical loads; preventing cracking of the resin is an important part of the technology.

The thermal instability is a consequence of the very small specific heat of most materials at liquid helium temperatures and, consequently, a localised energy input can easily lead to a significant increase in temperature. However, at liquid nitrogen temperatures the specific heat is some orders of magnitude greater and thereby the problem is considerably eased.

5.2.3 Applications of superconducting magnets

To produce magnetic fields greater than about 2T using conventional copper windings can require powers of the order of megawatts. The outstanding success of superconductors has been in providing strong fields for medical imaging, laboratory experiments and for high energy nuclear physics equipment without a massive power burden. The power to drive the helium refrigerators or liquefiers is a factor of 100 or even 1000 less. This advantage can extend to lower fields, especially when the field is required over a large volume. Magnets generating fields in excess of 15 T are readily obtainable with sufficient stability and homogeneity for use in magnetic resonance experiments.

5.2.3.1 Magnetic resonance imaging

The main commercial market for low-T_c superconductors is as magnets for medical imaging systems based on the magnetic resonance of hydrogen and other nuclei. More than 1500 magnetic resonance imaging (MRI) units are installed world-wide. A superconducting solenoid with a large room temperature is the heart of the system and represents about 25% of the total cost. Although these MRI units have to be topped up with liquid helium, with the latest sophisticated designs, this is usually required only annually. They have been further improved by the use of high temperature superconductor tape for the leads which reduces the heat leaking to the liquid helium by 90%. These MRI instruments provide a non-invasive diagnostic technique without the dose restrictions of conventional X-ray tomography. MRI can provide information concerning the structure of tissue and the flow of fluids in the body by the measurement of various relaxation times which, combined with its anatomical imaging ability allows detailed diagnostic characterisation. Although whole-body scanners using superconducting solenoids generating fields of more than 2 T are in use, a typical system has a room temperature bore diameter of 1.2 m with a central field of less than 1.5 T, and a homogeneity of about 10 ppm over half the bore. An example of a MRI system illustrating the superconducting magnet is shown in *Figure 5.2*.

5.2.2.2 High energy physics

Many superconducting magnets are used for particle accelerators, for beam handling and for analysing particle

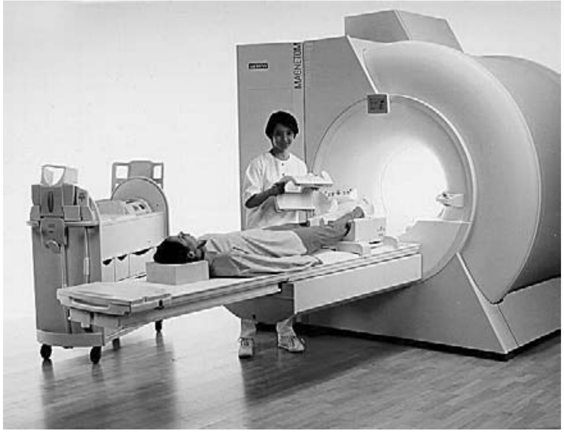


Figure 5.2 A magnetic resonance imaging (MRI) system with a superconducting magnet operating at 4.2 K. (Courtesy of Oxford Magnet Technology and Siemens)

reactions in bubble chambers. Beam-bending dipoles are made in large numbers and when the US superconducting supercollider (SSC) is built it will represent a market of a size similar to that of the MRI industry. The electron-proton colliding beam facility at Deutches Electronen-Synchrotron (DESY) in Hamburg, Germany has for example 416 dipoles each 8.8 m long with a central field of 4.7 T. Radiofrequency cavities employing surface superconductors are also used in some accelerators. The good performance of all these devices and the associated refrigeration plant has shown that equipment using superconductors can run economically and reliably. If magnetically confined thermonuclear fusion is to be a future power source, the windings needed will have to be superconducting. Such magnets have already been built for experimental investigation.

5.2.3.3 Magnetic separators

Superconducting magnets can be used for extracting weakly magnetic materials. This is a very good application for low temperature superconductivity magnets which are readily available. A fine magnetic particle will move in a magnetic field gradient and will be attracted onto the magnet providing the gradient. This technique has been used throughout the world for removing iron oxide from china clay to give a good white colour and to prevent low temperature localised melting during firing of the clay.

The strong magnetic fields and field gradients are also used for separating quite weakly magnetic materials such as ferro-manganese and ferro-titanium oxides. It is also used for seperating haemoglobin from blood.

5.2.3.4 Traction and machines

A pilot development for trains uses superconducting coils mounted underneath the carriages. The field from the coils provides lift from a conducting surface placed between the rails when the train is moving fast enough. The principle offers a possible efficient alternative to the wheels used on very high speed trains.

Homopolar d.c. motors using superconducting exciter coils have been built and work satisfactorily. A generator and motor operated back-to-back can form a quiet efficient ‘gearbox’ for ship propulsion.

An a.c generator with a superconducting exciting winding on the rotor is more efficient than one with a conventional copper winding: about 0.5–1% more output is achievable for a given mechanical input. Prototype superconducting rotors have been built and tested up to 400 MVA using NbTi technology. Continued development indicates that these machines will eventually become available.

5.2.3.5 A.c. power switches and current limiters

At the moment a superconductor becomes ‘normal’ it very quickly regains its resistance. The transition phenomena can be used as the basis of a simple power switch, but such applications have been limited by the poor a.c. performance of conventional superconductors. High-T_c superconductors should improve the position.

The conventional transductor intended for fault current limitation in electric power systems has a d.c. bias winding to saturate the core of an inductor, and so reduce the a.c. impedance of a coil on the same core. The core can come out of saturation on one half-cycle if the a.c. current attempts to rise excessively; the impedance of the transductor then increases. Superconductors offer an attractive method of providing the d.c. bias.

5.2.3.6 Other power applications

If hydrodynamic generators are to be used to produce electricity efficiently, their windings will have to be superconducting.

For energy storage, gigantic inductors holding perhaps 500 MWh are being discussed. Such a store could retain electricity generated cheaply at night for use during the day. A small superconducting energy store for controlling stability in a power system is being used on an experimental basis.

Because superconductors that carry a.c. do not function satisfactorily in strong fields, they give no benefit to power transformer design. However, ultra-fine filamentary NbTi has a sufficiently improved a.c. performance that consideration of such a design is possible. A small 70 kV-A transformer using such wires has been constructed.

5.2.4 Power transmission

High power superconducting cables, either a.c or d.c. appear to be more economical than conventional ones for power transfers in excess of a few gigawatts; such power transfers in single circuits are not at present required. A.c. cables would use superconductors in a surface mode to avoid excessive hysteresis losses.

5.2.5 Electronic devices

When two pieces of superconductor are connected together by a ‘weak’ link, many intriguing phenomena occur which form the basis for a variety of electronic devices. Niobium and lead are commonly used and the link, rather poorly conducting or even insulating, may consist of a point contact, a fine constriction or a thin (2–5 nm) layer of oxide. The resulting junction can carry supercurrents (typically up to a few hundred microamps) by tunnelling of electron pairs through the barrier without any voltage drop. This is the d.c. Josephson effect. If the current exceeds some critical value, which depends on the magnetic flux in the barrier, the junction switches to a resistive state with a voltage drop of typically a few millivolts. When there is a voltage across the junction, the supercurrent component oscillates at a very high frequency determined by the voltage (483.6 GHz/mV). This is the a.c. Josephson effect: it can be used as a precise voltage standard.

5.2.5.1 SQUIDs

An extremely sensitive measuring device may be made from a pair of Josephson junctions connected in parallel by a superconductor; this is known as a SQUID (superconducting quantum interference device). The total critical current then depends periodically on the flux (the period being the flux quantum) within the loop formed by the two junctions. Field changes as low as 10^{-14} T can be detected. Careful screening, together with differential techniques, are needed to eliminate unwanted responses. Small voltages in low-resistance circuits (e.g. 10^{-10} V in 1 Ω) and small currents may be measured by using the SQUID to determine the associated magnetic field. Sometimes just one junction is used in a superconducting loop with the flux being sensed by its effects on an inductively coupled radiofrequency bias circuit. With sufficiently high bias frequencies, measurements into the microwave region are possible.

Instruments incorporating SQUIDs are finding application particularly in geophysics (studies of rock magnetism, exploration based on anomalies, determination of deep earth conductivity) and in biomagnetism (e.g. investigation of electrical activity in the heart, brain and muscles).

5.2.5.2 Computer elements

Many groups are exploring the possibility of using SQUIDs and related devices in computers. The bistable characteristics, small size, ultra-short switching times (as low as 10 ps) and very low power consumption make SQUIDs attractive components for large high-speed computers which must be compact to reduce signal-transit times. The need for liquid helium cooling is not seen as a serious disadvantage.

As a logic gate, currents in control lines overlying the SQUID induce transitions between the zero voltage and resistive states by altering the flux linking the device. For memory cells, superconducting loops may be used with the binary values 1 and 0 being represented by the presence or absence of a persistent circulating current. One Josephson junction in series with the loop and one inductively linked to it are required to 'write' and 'read', respectively. Arrays of junctions may be deposited on a substrate using many of the processes developed for fabricating silicon microcircuits. Using such technology with niobium superconductor random-access-memory (RAM) devices of 1-kbit capacity and access times of less than 0.5 ns can be made.

5.2.5.3 Electronic devices

Devices other than digital circuits have been fabricated. Infrared and millimetre wavelength receivers based on superconducting-insulator-superconducting (SIS) quasi-particle heterodyne mixers and on bolometric detection are readily available. Real-time signal processing with bandwidths approaching 10 GHz are required in many front-end radar and communication applications. Superconducting analogue signal-processing components with bandwidths exceeding 2 GHz have been achieved in striplines, microstrips and co-planar devices. They exploit the unique property of superconducting thin films, i.e. low radiofrequency surface resistance. This property allows the fabrication of delay lines or high-Q resonators in compact planar structures. The component is manufactured in a similar manner to conventional microwave stripline, etc., except for the substitution of a superconducting film for copper.

5.2.5.4 Outlook

While many of the applications above are in everyday use it appears that eventually the low T_c superconductors will be replaced by high T_c superconductors. This may take some time particularly for MRI instruments which are part of a many million dollar industry. The coming use and application of high T_c superconductors is discussed below.

5.2.6 High temperature superconductors

Although certain metal oxides were known to superconduct at a few degrees kelvin for many years, they were only of academic interest until 1986 when an LaSrCuO compound was discovered which had a T_c around 30 K, a full 7 K higher than any previously known T_c. This discovery has generated world-wide activity and a number of other compounds have been discovered. The principal superconducting compounds to emerge are compounds of YbaCuO (YBCO), BiSrCaCuO (BISCO), TlBaCaCuO and HgBaCaCuO, all of which have critical temperatures significantly above the boiling point of liquid nitrogen

Table 5.4 Properties of some high T_c superconductors

Material	*Critical temperature* T_c* (K)	*Upper critical magnetic field* (T)		*Critical current density at 77.4 K and zero field* (GA/m^2)	
		4.2 K	*77 K*	*Bulk*†⇐	*Thin film*‡⇐
$YBa_2Cu_3O_7$	92	90	15	0.01	20–50
$Bi_2Sr_2CaCu_2O_8$	86	—	—	—	—
$(Bi,Pb)_2Sr_2Ca_2Cu_3O_{10}$	106	140	60	0.01	10–30
$Tl_2Ba_2Ca_2Cu_3O_{10}$	122	220	100	0.01	20–40
$HgBa_2Ca_2Cu_3O_{10}$	133	—	—	—	—

*Critical temperature depends on the exact stoichiometry.

†Sintered untextured material.

‡Thin film deposited on MgO or $SrTiO_3$ substrates—as indication of potential.

(77.4 K). They each form homologous series, the superconductivity depends on the exact stoichiometry of the particular compound. *Table 5.4* lists some of the more important compounds together with their properties. The use of liquid nitrogen as the coolant is a major breakthrough because it simplifies the cryogenic engineering, it gives a 50-fold reduction in the refrigeration power requirement, and, being about 1/30 of the cost of liquid helium, it gives a substantial reduction in the cryogenic costs. In addition, significantly better thermal properties at liquid nitrogen temperatures improve the superconductor's stability (see section 5.2.2). Below 30 K certain classes of compounds (e.g. BiSrCaCuO) have very good current-transport properties and a remarkable ability to carry current in fields up to 25 T, out-performing the best conventional metallic superconductors.

5.2.6.1 Applications

Developments to replace the conventional conductors copper or aluminium by high temperature superconductors (HTS) are very active in many parts of the world including USA, Europe and Japan. There are prototypes of power transformers, underground power cables, large motors and generators and fault current limiters in active development and in use. The electricity supply of the city of Geneva in Switzerland is completely provided by power transformers wound with HTS conductors. It is expected that there will be definite power savings with the use of HTS and this will provide a considerable contribution to the reduction of environmental pollution. All of the applications for low temperature superconductors mentioned in sections 5.2.3.2 to 5.2.5.3 are candidates for replacement by HTS.

These HTS materials are better than LTS for 50/60 Hz a.c. current and can even be used for radiofrequency and microwave applications. Here the main requirement is a low radiofrequency surface resistance for antenna arrays and microwave resonators and the HTS tape has this property.

5.2.6.2 Tape

High temperature superconductors are very brittle and this has provided a considerable challenge for the production of suitable conductors. This is gradually being mastered and narrow composite tape of the HTS is currently being produced. There are a number of variations of the substrates but the following is typical. A nickel tape is used as the base and a buffer layer of palladium followed by ceria and yttria-zirconia layers are applied to the tape. This is the basic substrate and the high temperature superconductor is then deposited onto the tape using pulsed laser deposition. The tape is 3 mm wide and cables up to 100 metres in length are currently under test at various locations. The production of this tape is essential to all the power applications and the production rates are increasing and the cost is decreasing.

5.2.6.3 Outlook

The outlook for HTS is very good with many prototypes of applications at the test stage in utilities and elsewhere. While there is need for cheaper tape, this will inevitably come as production increases. The highest T_c has remained at 133 K for the last few years however it has been found that the T_c can be increased by pressure. This may provide a lead for producing materials with higher values of T_c. The dream of a T_c higher than room temperature may remain unfulfilled but we must remember that the highest T_c remained at 23 K for over 20 years and we can always hope for a further break through.

My thanks are due to the staffs of The City of Sunderland College and the University of Sunderland for their assistance with this section.

References

Low temperature superconductors

FONER, S. and SCHWARTZ, B. B. (Eds), *Superconducting Applications: SQUIDs and machines*, *(NATO Advanced Study Institutes Series, vol. 21)*, Plenum Press, New York (1977)

ROSE-INNES, A. C. and RHODERICK, E. H., *Introduction to Superconductivity*, 2nd edition, Pergamon Press, Oxford (1997)

WILSON, M. N., *Superconducting Magnets*, Oxford University Press, Oxford (1983)

High temperature superconductors

BEDNORZ, J. G. and MULLER, K. A., Perovskite-type oxides—the new approach to high T_c superconductivity, *Reviews in Modern Physics*, **60**, 585 (1988)

SHEAHEN, T. P., *Introduction to High Temperature Superconductivity*, Plenum Press, New York (1994)

VARIOUS AUTHORS, *Superconductivity in electric power: a special report, IEEE Spectrum*, July 1997, pp. 18 to 49

SEEBER, B., ed, *Handbook of Applied Superconductivity* 2 vols. Institute of Physics, London (1998)

LAWRENCE, L. R., COX, C. and BROMAN, D., *High Temperature Superconductivity: The Products and their benefits 2000 edition.* ORNL/Sub/450006921.

6

Semiconductor, Thick and Thin-Film Microcircuits

W Eccleston MSc, PhD, CEng, FIEE
University of Liverpool
(Section 6.1)

D Grieve BSc
Welwyn Components, Bedlington
(Section 6.2)

Contents

6.1 Silicon, silicon dioxide, thick- and thin-film technology 6/3
  6.1.1 Materials for integration 6/3
  6.1.2 Silicon and impurities 6/3
  6.1.3 Mobility and the materials limitations of device size 6/4
  6.1.4 Materials for large-area circuits 6/5

6.2 Thick- and thin-film microcircuits 6/5
  6.2.1 Thick-film materials 6/5
  6.2.2 Thin-film materials 6/7
  6.2.3 Types of hybrids, and their applications 6/7

6.1 Silicon, silicon dioxide, thick- and thin-film technology

Silicon is unrivalled as a semiconductor material for active devices such as rectifiers, metal oxide semiconductor field effect transistors (MOSFETS), bipolar transistors, thyristors and a wide variety of other structures particularly designed for power applications. Few of the properties of silicon are ideal but none are so far from ideal as to exclude its use in all but a small fraction of current devices. Perhaps the most important application where it is excluded is in the generation of electromagnetic radiation, optical infra-red or microwave. Such emitters usually require a material where electrons can lose energy directly with emission of light rather than by interaction with the lattice of semiconductor atoms. For this application a variety of compound semiconductor materials are available such as gallium arsenide and gallium phosphide. Much research is, and has been, focused on producing optically emitting alloys of silicon with other elements such as germanium which can be integrated on a silicon substrate using the existing highly developed technology.

6.1.1 Materials for integration

Quite fundamental to silicon technology is the existence of a natural oxide which can be produced thermally in oxygen or steam and which is hard, uniform in thickness, has an acceptable value of relative permittivity (3.9) and high electrical breakdown field strength ($2 \times 10^7\,\mathrm{Vcm}^{-1}$). Many semiconductor devices have at the surface electron densities of between 10^{11} and $10^{12}\,\mathrm{cm}^{-2}$. It is important that these carriers remain mobile, rather than being trapped at spare or dangling bonds at the interface between the single-crystal silicon and the amorphous oxide with its much more randomly oriented atoms. Less than 1% of the silicon atoms should have such a bond if an acceptable number of electrons is to remain mobile. This can be achieved through hydrogen treatment at temperatures below 650°C. Too high a level of trapping leads to devices of low gain and high electrical leakage.

Silicon dioxide is near ideal for carrying thin-film conductors from and to adjacent regions of active devices. When it contacts silicon, however, to make an ohmic contact it will take it into solid solution, essentially etching the silicon. Spikes of aluminium, with dissolved silicon, occur which may short out the device. For this reason aluminium already doped to the level of solid solubility with silicon is used for the starting material. Contact to the package is made with aluminium wire which is ultrasonically welded to thick aluminium pads on the silicon dioxide. The bond-wire is also doped with silicon. Polycrystalline silicon also serves as a current carrier on integrated circuits. Impurities such as phosphorus and arsenic are used to dope the polysilicon and this renders it much more conducting. Its conductivity can also be improved by converting it to one of the silicides of metals such as platinum, cobalt or tungsten. Silicon dioxide must be etched to allow the conductors on its surface to reach the underlying active regions. These holes in the silicon dioxide are required to be as small as possible, of 'minimum feature size', and methods of etching it employing either gaseous or aqueous fluorides have been developed.

There is considerable advantage in having an insulating layer buried beneath the transistors. Device capacitance is reduced as is a wide range of spurious transistor latch phenomena which can destroy the transistors. Silicon dioxide is used in the form of a thin film, often implanted into the silicon. This kind of implantation requires very high current densities, followed by annealing at temperatures of 1350°C, when there is a condensation of the oxygen ions to produce a layer of not very dense silicon dioxide. Wafer bonding is also used for this application. Two oxidised wafers are bonded together by heating the two oxidised surfaces in contact. One of the slices is then etched back to produce a thin film suitable for devices. An intriguing technology is that of porous silicon. The slice is anodised in hydrofluoric acid, when the selected regions acquire a large number of microscopic holes which are later oxidised to produce silicon dioxide. Porous silicon also has possible applications for light-emitting devices. The presence of pores substantially modifies the electronic properties of silicon, the confined electrons being able to fall through large energy differences without interaction with the lattice of silicon atoms. A particular problem of this type of technology is the very poor thermal conductivity of the oxide. Power dissipated in the silicon islands containing the devices produces temperature rises of the order of 150°C. An earlier form of 'silicon on insulator' employed sapphire as substrate, partly because of its crystalline structure but also because of its high thermal conductivity.

6.1.2 Silicon and impurities

The devices produced in silicon depend for their properties on the addition of very low concentrations of impurities coming from group 3 or 5 of the periodic table. The silicon atom belongs to group 4 and it bonds in the single crystal in a tetrahedral configuration with each atom having four nearest neighbours. This configuration gives the material its brittleness which can be a problem when it is being handled. The lattice constant is 5.4 Å which is the length of the side of the cubic unit cell. Group 5 elements such as phosphorus and arsenic have five outer electrons available for bonding, leaving one electron unused and in orbit. The application of Bohr theory to such an orbiting electron is justified as long as the effect of the crystal is taken into account via the use of the effective mass of the electron and the relative permittivity (12) of the silicon. The calculated radius of orbit is very much larger than the lattice constant and the electron is only very weakly held. Thermal vibrations of the lattice can easily shake the electron loose at room temperature and it can then contribute to the conductivity. Suppose $10^{16}\,\mathrm{cm}^{-3}$ of phophorus atoms is added to an otherwise pure crystal (an impurity level of 1 ppm). The electron concentration is increased from 1.5×10^{10} to $10^{16}\,\mathrm{cm}^{-3}$ and the conductivity is increased by approximately 10^6. This remarkable change is the basis of the usefulness of impurities in producing semiconductor devices.

In order to consider the effects of substituting group 3 atoms for silicon it is necessary to understand the nature of electrical conductivity in single-crystal solids. Bohr theory describes the properties of the isolated atoms in terms of electrons in orbit around a positive nucleus. When isolated atoms condense to produce a solid we must also apply the Pauli exclusion principle which states that the electrons must each have a different orbit. Instead of a single energy level of the Bohr model we have bands of levels each containing the same number of energy levels as the number of atoms. The free or conduction electrons fit a similar model. When an applied field acts on an electron it will increase its kinetic energy and move it to a higher level, but it must still obey the exclusion principle. A consequence is that no conductivity can take place in a full band. In a semiconductor the highest filled band at 0 K is called the *valence band*; the first unfilled band is the *conduction band*. At higher temperatures some of the electrons in the valence band will be able to jump across the energy gap leaving behind a few vacancies in the valence band and giving some free electrons in the conduction band. At 300 K the number of such electrons in the silicon is the *intrinsic carrier concentration*. The presence of vacant energy levels in the valence band provides an additional form of conductivity since the sea of electrons in this band can experience an overall gain in energy and, because the band has an electron deficiency, it can be ascribed a positive charge. The total conductivity of all the electrons is equivalent to that of a single positively charged particle. We term this a *hole*. The intrinsic hole concentration is therefore equal to the concentration of electrons in the conduction band since, as expected, the material remains electrically neutral.

When a group 3 atom substitutes for silicon it is short of an electron to complete all the bonds to the four nearest neighbour silicon atoms. This electron is gained from the valence band, leaving a vacancy and the remaining electrons can contribute to the conduction. A hole is therefore created.

Two types of material can therefore be created by the addition of impurities. Group 5 elements become positively charged fixed centres which release electrons making the material *n type*. Group 3 elements take up an electron from the valence band leaving behind a mobile positive hole and the material is then said to be *p type*. The intersection of a p-type material with an n-type one gives rise to a p–n junction which is a rectifier but is also the building block of almost all other semiconductor devices.

Impurities can be added to the melt during the growth of the single-crystal material. This process involves pulling a seed of the crystal from the melt whilst also rotating it. A long single crystal is so produced up to 8 in. diameter. This is sawn many times along the same well-defined crystal plane and each resulting slice is ground and polished on one of its two faces. The doping level and, therefore, the resistivity can be controlled to within a few per cent of the desired value. In most cases later doping during device fabrication will overdope this background level. Such impurity atoms can be introduced from a solid or gaseous source in a furnace at temperatures in the region of 1000°C. Patterning silicon dioxide guarantees that the surface of the slice will only be doped in the areas required. This stage is termed the *predeposition stage*. In order to get the atoms on the desired silicon sites it is necessary to heat the silicon for a prolonged period in the absence of the dopant. This is termed the *drive-in stage* and must be carried out in an oxidising atmosphere in order to produce a layer of silicon dioxide on the surface of the silicon, sealing in some of the impurity. Even with this precaution a large amount of impurity is lost. The final depth of the p–n junction will be much deeper than that after the predeposition. A major difficulty with this technology is that the spread in the resistance can be as high as 20%, which is unacceptable for the more advanced processes. Better control is achieved by ion implantation. Here a beam of suitably charged ions delivers a closely monitored number of ions to the surface. The ions are energetic and much surface damage is caused. The ions sit interstitially in the silicon and must be activated by prolonged annealing. Control is greatly improved but the cost of a commercial implanter is high. The process is essentially serial and, therefore, time consuming, in contrast to diffusion where many slices of silicon can be loaded into a furnace for simultaneous predeposition. Control of the temperature–time cycle is difficult with simple furnace processing and a ramped temperature cyle is increasingly used. This factor is of great importance in very low dimensional structures since any unnecessary heating during the warm-up period produces an unwanted lateral spread of the impurities and loss of dimensional control. For this reason a well-controlled temperature cycle is employed giving precise control of geometry. This is a single-slice process and is often termed *limited reaction processing*. Batch processing is less popular since it can, on equipment failure, lead to the loss of expensive part-processed wafers.

In a small but significant number of cases it is necessary to produce a layer which is less heavily doped than the background of the crystal. This is done using epitaxy. Silicon tetrachloride, dichlorosilane or silane is decomposed at high temperature (1200°C down to 900°C). The resulting silicon atoms which have high mobility on the silicon surface are able to migrate to appropriate crystal sites, reproducing the orientation of the underlying silicon. Gaseous dopant is added to the gas stream and can be varied in type and concentration during the growth. The high temperatures required in the reactor guarantee high atom mobility and good crystal quality; however, they do lead to short growth times and relatively poor thickness and doping control, particularly across a slice. The reactor consists of a cold-wall vessel which prevents unwanted deposition of material, and heating either with radiofrequency in small-scale systems or lamps in large commercial apparatus.

6.1.3 Mobility and the materials limitations of device size

Throughout the history of silicon circuit development there has been a particular emphasis on making devices, and hence circuits, smaller. This need is associated with the statistics of defect creation, both in single-crystal silicon and in the associated silicon dioxide and conductors. The fraction of working circuits increases in an exponential fashion with decreasing area of silicon and the cheapness of producing the finished chip follows, therefore, a similar very sharp function.

Producing the windows for selective addition of impurities is a critical process which utilises a photosensitive liquid spun on the surface, and after heat treatment is exposed through a photomask to produce the desired pattern. The silicon dioxide can then be etched to leave windows of exposed silicon. Alternatively, a material can be applied to the surface which is chemically sensitive to electron beams. A scanning electron beam can be used to define features in this 'resist' layer, which masks the surface in selected areas during a subsequent manufacturing process. As the physical dimensions of transistors are reduced, the electric-field strength in the silicon and the silicon dioxode becomes very high. At low field, the drift velocity of the electrons, or holes, depends linearly on the applied field. The rate of change of velocity with field is the *mobility*. Reducing the chip supply voltage in proportion to the device dimension should be possible without degradation of speed of operation. Unfortunately, as the intensity of the fields becomes higher the velocities of the electrons become field independent, mobility falls and circuit designers are forced to drive the circuits harder by maintaining a high supply voltage, typically 5 V. The silicon shows carrier multiplication through hot electrons or holes exciting other electrons and holes, by collision, across the energy gap. Modern circuits are often so complex that adjacent n and p regions form unwanted four-layer thyristors which can switch into a state where very high currents can pass. The same carrier multiplication can generate white light. Long-term degradation of silicon dioxide is known to be associated with the generation of very energetic carriers in very small transistors where the electric field strengths are very high.

6.1.4 Materials for large-area circuits

The prime limitation on the production of large-area circuits is the very strong relationship between the fraction of working circuits and the defect density. Attempts at wafer-scale integration (WSI) have been made. This is an attractive possibility in that it reduces packaging costs and the overall size and weight of equipment. Clearly, to overcome the problems of the relationship between yield and area it is necessary to build in large amounts of redundancy through the duplication of power lines and circuits. There is little materials technology which is specific to this kind of circuit.

Although single-crystal silicon slices have increased in area and have fallen in cost over the years, they are still impractical for some applications. In some cases it is possible to get satisfactory performance from devices made in silicon on glass or other cheap flat substrates. Silicon deposited at room temperature is usually amorphous. It has randomly oriented atoms and thus a large number of dangling bonds which can be taken up by treating the film with hydrogen. Dopants can be introduced during growth and relatively poor p–n junctions are obtained. The energy difference between conduction and valence bands is larger than for single-crystal silicon and this reduces leakage currents. A major course of such leakage is the excitation of electrons from the valence band to the conduction band which becomes less probable with a larger energy gap. Amorphous silicon can be very photosensitive and the availability of large areas on panes of glass make it a major contender for solar cells. A second application is in liquid-crystal displays for television monitors. These are termed *active matrix* displays and they use a primitive form of the MOSFET. Each small element of the picture (or pixel) is independently accessed by a thin-film device which supplies, when required, a suitable voltage to the cell. Each row and column is accessed via a shift register which is ideally produced on the plate of glass in order to reduce cost. The electrons in amorphous material are insuffiently mobile and polycrystal-line material is preferred for this application. A reactor, similar to that used for silane epitaxy, is employed. A capacity for operating at low pressures is important if the grain size and hence mobility are to be sufficiently high. An interesting, if more costly, method is to laser anneal amorphous silicon to produce polysilicon in those regions where higher performance circuitry is required. This large-area technology is also ideal for very big circuits where the speed limitations of the individual devices can be overcome by parallel, or neural, processing.

6.2 Thick- and thin-film microcircuits

Hybrid microcircuits use patterns of thick and thin conducting films on an inert substrate to replace individual resistors, and to connect components such as capacitors, transistors and ICs into a complete circuit, thus saving both space and weight compared with PCBs. The first application of microcircuit techniques was for the wartime production of proximity fuses, using printed carbon composition resistors on ceramic substrates. The basic methods are recognisable 60 years later, even though modern materials have greatly improved performance.

Hybrid technology is always extending in new directions, as new components are being developed, which need novel packaging and interconnection techniques, for example;

MCM	multi-chip modules
LTCC	low temp. co-fired ceramic packages
DBC	direct bonded copper for power circuits

Metal substrates for heater applications

For any electronics application, an appropriate construction can be selected which will meet the specification on performance, size and price.

Thick-film hybrids are most suitable for volume replication of relatively simple circuits, for example, the duplicated circuits in telephone exchanges. Screen printing is used to deposit patterns of conducting pastes on ceramic substrates, often at rates of thousands per machine per day. High-temperature firing fuses the pastes to the substrates, forming circuit elements of less than one-thousandth of an inch in thickness. Even so, these are still ‘thick’ films, compared with the film thickness of ‘thin’ film circuits. The process here is completely different, using photolithography and selective etching on purely metallic films to generate the circuits.

Microcircuits are discussed here in terms of materials and processes, and how they are applied to a range of circuit applications.

6.2.1 Thick-film materials

The materials used for hybrids are being continually up-graded and extended, to improve the final product. The basic properties of the materials are as follows.

The substrate material must be insulating, flat, non-reactive and thermally stable to the firing temperature (about 850°). The most widely used material is high purity (96%) alumina, used in plates from 50 × 50 mm (2 × 2 in.) to

Table 6.1 Common alloys and their properties

Material	*Resistivity* (mΩ/sq)	
Ag	2–5	Inexpensive. Poor leach resistance in molten solder.
Ag/Pd 30:1	3–6	Budget conductor. Adequate leach resistance.
Ag/Pd 3:1	20–35	Less used because of price. Good adhesion, leach resistance. Wire bondable.
Pd/Au	50–90	Expensive. Better adhesion, leach resistance. Wire bonding fair.
Pt/Au	70–100	Very expensive. Excellent adhesion, leach resistance. Wire bonding poor.
Au	3–4	Most widely used multilayer conductor. Excellent conductivity, good wire bonding, not solderable.
Au/2% Pd	5–7	Modified gold for aluminium wire bonding.
Cu	2–4	Good solderability but needs non-oxidising furnace atmosphere (nitrogen).

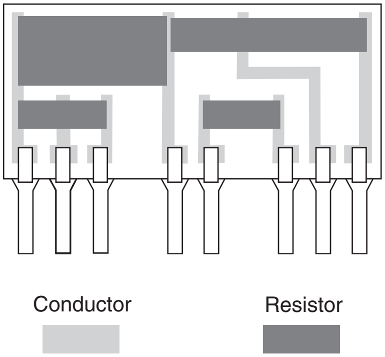


Figure 6.1 Typical resistor network in single-in-line form

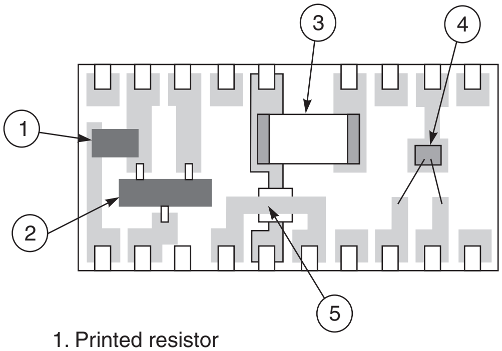
175 × 125 mm (7 × 5 in.), and usually 0.63 mm (0.025 in.) in thickness, although any thickness between 0.25 and 2.5 mm (0.010–0.100 in.) can be obtained.

Exotic materials such as beryllia and aluminium nitride are occasionally used because of their better thermal conductivity in high power dissipation situations. Stainless steel, suitably insulated with glaze, is also used in power applications.

For any particular application, a conductor is chosen which will give adequate performance in terms of adhesion, solderability, etc., at an economic cost. The air-firing conductors are based on the precious metals silver, gold, palladium and platinum. The industry standard has been 3:1 Silver/Palladium, but demand for palladium for automobile catalytic converters has increased the cost greatly. Manufacturers now use conductors with ratios as low as 30:1, as well as replacing part of the palladium with platinum in a ternary alloy.

Copper and nickel systems have been developed because of the lower cost of the metal, but these conductors must be fired in a non-oxidising atmosphere (nitrogen, with oxygen <20 ppm).

Modern resistor pastes are based upon ruthenium, either as the dioxide or as bismuth ruthenate. Resistivities between 1 Ω/sq and 10 MΩ/sq can be obtained, and intermediate values can be made by blending. The temperature coefficient of resistance (TCR) is ±100 ppm or better, and load stability is good. Fired resistors are often overglazed with low-temperature glaze (550–600°C) which contributes to the mechanical and environmental protection of the resistors.



1. Printed resistor
2. Plastic packaged transistor
3. Chip capacitor
4. Wire bonded transistor chip
5. Printed crossover

Figure 6.2 Typical hybrid, with add-on components, in dual-in-line form

The circuit need not be limited to a single level on the substrate. Crossovers, using areas of dielectric glaze, enable conductors to cross one another, and if this is extended to cover the majority of the substrate, these circuits are known as ‘multilayers’. Multilayer hybrids are built up from layers of conductor tracks, separated by layers of dielectric glaze and interconnecting through windows (called v*ias*) in the glaze. The glaze must tolerate multiple re-fires as each layer is added, and have a low dielectric constant, to minimise stray capacitance between tracks which cross. Resistors can be fired on top of glazes loaded with ceramic, giving results similar to resistors on alumina substrates.

Screen-printable protections are printed over trimmed resistors and cured at 200°C or less. Complete circuits can be dip-coated in liquid resin suspensions, powder coated, moulded or potted. The usual resin types which are used are silicones, epoxides or phenolics. Resins can be cured very quickly by U.V. lamps.

The components which can be attached to hybrids are of all shapes and sizes. The name ‘hybrid circuit’ is derived from the ability of this technology to mix all kinds of component to achieve the desired circuit characteristics.

Hybrids with only resistive, capacitive or inductive elements are called ‘passive’ hybrids. If semiconductor elements are included, they become ‘active’ hybrids. Miniature plastic-packaged semiconductor components are designed to be attached with solder, to pads on the hybrid. In this technique, known as ‘Surface Mount’, solder cream is printed over conductor pads on the substrate, into which the component feet are placed, prior to reflow. The solder cream consists of small solder balls, flux and solvent, and it is printed through a metal-foil stencil screen to give a suitable thickness of deposit.

Semiconductor chips are used in many packaging styles. Bare chips without any protective coating are wire bonded into circuits, and chips with moulded plastic protection are soldered. Complex chips are packaged in chip carriers, which are small hermetic packages, and permit the electrical parameters to be fully tested before soldering into a circuit.

The availability of active components for surface mounting led to the development of chip resistors which could be attached in the same way. Resistors are printed in a matrix on large ceramic substrates, which are divided into individual resistors. Solderable terminations at the ends are added by sputtering and plating. The immense market for mobile phones and the pressure to make them smaller has led to resistor chips as small as 0.5 × 0.25 mm (0.020 × 0.010 in.) which can be robotically placed. These tiny sizes are also popular because they use less ceramic and materials, hence they are more economical.

Chip capacitors consist of alternate layers of dielectric and conductor, sintered into a solid block, and with metallised terminations. The sizes range from 0.5 × 0.25 mm to 6 × 5.5 mm (0.020 × 0.010 in. to 0.240 × 0.220 in.) and in value from 1 pF to 1 μF. High-value capacitors are miniaturised tantalum capacitors with metal end caps, or moulded with metal tabs.

Small wire-wound components such as inductors and transformers have also become available for hybrid use. A fibre-optic cable can be fed through the side wall of a hybrid package, for conversion of light information into an electrical output.

6.2.1.1 Thick-film processes

There are many stages between the customer's original drawing and the final circuit.

The circuit elements are laid out using computer-aided design (CAD), from which photopositives are made with a photoplotter for each different ink to be printed on the substrate. The CAD will also prepare all the necessary drawings and instructions for production personnel. Typical conductor line width is 0.25–0.75 mm (0.010–0.030 in.) and resistors are usually not less than 1.25 mm (0.050 in.) square.

The photopositive is exposed and developed on a screen coated with U.V. sensitive emulsion. The screen mesh is stainless steel or synthetic fabric, with between 2.4 and 16 meshes/mm (60–400 meshes/in.). A thick film paste placed on the screen is printed onto the substrate by a moving flexible squeegee blade.

Printing machines can be loaded by hand, or fitted with mechanical feed systems, which take substrates from magazines, and transfer printed substrates to belt driers. The dried pastes are fired in multizone belt furnaces through a controlled temperature-time profile. A typical profile for conductors/resistors is 60 min through-time, with 10 min at a peak temperature of 850°C.

Resistors are trimmed by laser energy, which has the advantages of speed, precision and cleanliness. Substrates, printed with multiple circuits, are scribed using a more powerful laser, so the individual circuits can be snapped out.

Components attached by solder joints are placed on printed solder cream pads. To achieve the volume of work necessary, this is usually done by computer-controlled pick-and-place machines, which take each component from a separate reel, and place them in the correct position. There are several methods used to reflow the solder cream, which use heated belts, tunnel furnaces under nitrogen, and IR lamps, for example.

After soldering processes involving fluxes, the units may be cleaned, but many solder systems now are 'non-clean', because the flux is not a hazard to the operation of the circuit.

It is planned to cease using solders containing lead in electronics, for environmental reasons. There are several systems that could be used as alternatives, using tin alloyed with a variety of other metals. The greatest change affecting their use in hybrids is that the melting point is usually 30–40°C higher than the tin/lead eutectic, and the components have to withstand being heated to a higher temperature during assembly.

Semiconductor chips are usually glued into position with epoxy. This can be made conductive with silver or gold powder. Conductive epoxy is also used to attach capacitors to gold conductors, because solder dissolves gold very quickly. Bare chips are bonded with aluminium or gold wires of about 0.025 mm (0.001 in.) diameter. Aluminium wire is not used with pure gold conductors, because the bond strength deteriorates rapidly.

Hybrids may be mounted on their own terminals, or mounted inside a metal or ceramic package, which has its own terminals. The terminals are supplied on reels with perhaps 50 000 on a continuous strip. Insertion machines crop unwanted terminals and insert substrates into the strip, reeling up the strip afterwards. The strip is then fed through a continuous wave-soldering machine, then the soldered strip may be processed by a finishing machine, which crops each unit from the strip, tests it and packs it.

Metal packages are closed by soldering or welding a lid or cover to the header on which the substrate is mounted. A dry nitrogen atmosphere is maintained inside the package, to preserve the circuit from corrosion. The packages themselves are protected by plated films of tin, solder, nickel or gold.

6.2.2 Thin-film materials

The preferred substrate materials are glass or 99.6% alumina, because the surface finish has to be very smooth to allow the deposition of a uniform metal film. The film itself has a sheet resistivity between 50 and 500 Ω/sq with TCR ±40 ppm, and many materials are in use, such as nichrome or tantalum nitride. The resistive film is overlaid by a conductive metal film. Conductors for wire bonding use gold, whereas if solderable conductors are required, nickel covered with a gold flash is used. To optimise conductor properties, there is a complex layered structure, with titanium and palladium layers under the gold to minimise diffusion effects, especially of chromium to the surface of the conductor.

Substrates can be metallised in-house, or bought-in with a film of known resistivity from an outside supplier. The add-on components are the same as for thick film.

6.2.2.1 Thin-film processes

Typically, the smallest resistor line width is 10 μm and conductor lines can be 50 μm. Designs can avoid cross-overs by wire-bonding over several lines, using the fine-line capability to crowd the conductor tracks. Designs cannot be photoplotted at this line width, so must be produced at a magnified scale, and reduced photographically. The photopositives are produced on glass because of the dimensional stability required.

A thin film of liquid photoresist is spun onto the surface of a substrate, and dried. The pattern is exposed and developed. Selective etchants remove the metal layers as required to produce the circuits.

Thin-film circuits produced in a matrix on glass substrates are separated by dicing with a diamond cutting wheel. Thereafter, individual circuits are assembled using the same methods as for thick film, although generally they are used in hermetic packages, which protect the glass substrates.

Thin film is less widely used than thick film for a number of reasons. The low resistor film resistivity means high value resistors take up too much area, the photomechanical process is not amenable to volume production, and the final package cost is high.

6.2.3 Types of hybrids, and their applications

With the wide range of materials and components available, the hybrid process can be used to produce any kind of circuit, from simple resistor networks to complex multilayers which can cost 1000 times the price of the simple unit. The following examples illustrate the capability of hybrids to produce circuits of increasing complexity.

6.2.3.1 Resistor networks

These networks are usually fabricated on ceramic substrates using thick film techniques, where the wide range of paste resistivities can be used to achieve any mix of resistor values. The package may be terminated by a single row of terminals along one edge (single-in-line SIL) or with two rows of terminals along 2 edges (dual-in-line DIL) with the substrate horizontal against the PCB. With a substrate thickness of 0.063 mm (0.025 in.), SIL circuits can be easily fitted to printed circuit boards on 2.5 mm (0.1 in.) pitches, and the height above the board can be between 3.75 mm (0.150 in.) and the maximum the customer will allow.

Design of a hybrid is very flexible, both in the size and shape of the substrate, and in the number and position of the terminals (the ‘pin-out’). Both sides of the substrate may be printed with resistors, provided the total dissipation does not exceed the substrate capability. If the network has to withstand power pulses from lightning surges, a thicker substrate may be used, which is stronger, and withstands the sudden heating effect. Many millions of these simple resistor networks are used in Telecoms, for example, as line feed resistors.

Divider networks made with the same resistor paste have good tracking performance, but to obtain the ultimate performance in this application, thin film is best.

6.2.3.2 Units with surface-mounted devices

It is a small step to convert a passive resistor network into an active circuit by incorporating semiconductor devices and capacitors. Every kind of component is available suitable for surface mounting. The components are assembled using pick-and-place machines, and reflow-soldered to the hybrid. The circuit may be protected by printed or dipped organic resin, or the circuit may be left in its overglazed state, since the add-on components have their own protection.

This is the most widely used type of circuit. It is used in Automotive applications for engine management systems and vehicle safety systems, and for general industrial applications.

6.2.3.3 Networks with wire-bonded chips

A substantial improvement in packing density can be realised by using bare chips, but the wire bonds must be protected, inside hermetic packages, or inside a globule of hard resin (‘globtop’). Thick or thin film can be used, depending on the resistor values and tolerances required. The semiconductor dice are attached by conductive epoxy or eutectic bonding, then they are wire-bonded to the conductor pads. Manually, this can be a relatively slow operation, but automatic bonders are available with pattern recognition cameras, which can identify the bonding pads and make the bonds.

These circuits are used for military and aerospace applications—for radio and radar, sonar, ordinance fuses, electronic gauges for Head-up-displays. Other applications have been for proximity detectors, and control circuits for laser displays.

6.2.3.4 Power microcircuits

The hybrid process has been adapted to make motor drive circuits. These need very low conductor resistance, and the ability to bond 250 µm (0.010 in.) Al wire to the circuit and the power devices. These circuits use Direct-bonded Copper (DBC) substrates. Copper foil 300 µm (0.012 in.) in thickness is bonded directly to the alumina substrate, and the pattern is etched. The copper is given a nickel flash plate, then the components are attached and wire-bonded. The ceramic is then bonded to a heatsink, to prevent the semiconductors overheating.

6.2.3.5 Multilayers and multichip modules

The interconnect density of a hybrid can be multiplied several times by using multilayered conductors. It is possible to fit into a 30-pin package 40 × ~~25~~ mm (1.6 × ~~1~~ in.) about 15–20 integrated circuit (IC) chips, which would normally need a PCB at least 400 mm (4 in.) square. A high-conductivity conductor is used, which may be gold or silver.

A circuit can often be assembled from standard ‘building blocks’, which are already available as separate chips. Much time and expense may be saved, by making a hybrid to interconnect the separate chips, rather than designing a custom IC, with the necessary design and mask costs. This type of hybrid is known as a multichip module. One example made this way was an aircraft fuel status indicator.

6.2.3.6 LTCC modules (low temperature co-fired ceramic)

This is a multilayer, without the ceramic substrate. Dielectric is produced as a 250 µm (0.010 in.) ‘green tape’™*, which is punched with vias, and printed with conductors. Circuit levels are made on separate sheets, which are stacked like a book, and fired at 850°C. The stack sinters into a solid block, containing all the interconnections, with excellent robustness and repeatable electrical characteristics. This technique is only viable for very large volume applications which can justify the tooling costs for punching the ceramic tape.

This manufacturing technique has been used in the ‘Bluetooth’ project. This is a system to allow any kind of computer-controlled device to link up with any other.

6.2.3.7 Microwave circuits

These circuits are made chiefly in thin film, because of the better control of line sizes and edge definition with this process, and the substrate material can be varied to exploit different dielectric constants.

6.2.3.8 Metal substrates

Although ceramic has admirable properties as a substrate, it is brittle, and the thermal conductivity is not as good as metals. Applications have arisen for heater elements, for example, in domestic kettles, which have been met by printing conductive elements on an insulating layer on a steel substrate. By careful matching of expansion coefficients, the element can withstand the thermal cycling, and in the case of the kettle elements, the rate of heat transfer is excellent.

The process is versatile; elements have been printed on tubes for electric showers, and on panels to prevent condensation in mobile phone base-stations, for example.

6.2.3.9 Standards

The appropriate UK Standard for hybrids is BS 9450, but this is now being replaced by a European standard in the form of EN 165000, which covers Capability and Qualification Approval. This incorporates best practice from UK, European and US standards. An additional standard EN 265000 is being produced for Technology Approval. The EN 165000 has been submitted for approval for conversion to an IEC standard.

[*™] DuPont trademark.

References

Physical constants

Handbook of Chemistry and Physics, The Chemical Rubber Company, Cleveland, OH

Physical principles and materials constants

HERBERT, J. M., *Ceramic Dielectrics and Capacitors (Electrocomponent Science Monograph)*, Gordon and Breach, London (1982)
SIZE, S. H., *Semiconductor Devices Physics and Technology*, Wiley, New York (1985)

Current developments in silicon devices and technology

ECCLESTON, W. and UREN, M. (Eds), *Proceedings of European Solid State Device Research Conference 1990*, Adam-Hilger, Nottingham

Thick- and thin-film microcircuits

COLCLASER, R. A., *Microelectronics Processing and Device Design*, Wiley, New York (1980)
KITTEL, C., *Introduction to Solid State Physics*, 5th edition, Wiley, New York (1986)
MAZDA, F. (Ed.), *Electronic Engineers Reference Book* 6th edition, Butterworths, London, Chaps 28 to 31 (1989)
MCKELVEY, J. P., *Solid State and Semiconductor Physics*, Harper & Row, New York (1966)
MORGAN, D. V. and BOARD, K., *Introduction to Semiconductor Microtechnology*, Wiley, New York (1983)

7

Insulation

A J Pearmain BSc(Eng), PhD, MIEE, CEng
Queen Mary and Westfield College, University of London
(Sections 7.1 to 7.9)

A Haddad Ing.d'Etat, PhD
Cardiff University
(Sections 7.10 and 7.11)

Contents

7.1 Insulating materials 7/3
  7.1.1 Classification 7/3
  7.1.2 Temperature index 7/3
  7.1.3 Effect of frequency 7/3
  7.1.4 Fire behaviour 7/3

7.2 Properties and testing 7/4
  7.2.1 Physical properties 7/5
  7.2.2 Mechanical properties 7/5
  7.2.3 Electrical properties 7/6
  7.2.4 Chemical properties 7/7

7.3 Gaseous dielectrics 7/8
  7.3.1 Breakdown mechanisms in gases 7/8
  7.3.2 Air 7/8
  7.3.3 Nitrogen 7/10
  7.3.4 Sulphur hexafluoride 7/10
  7.3.5 Hydrogen 7/11
  7.3.6 Vacuum 7/11

7.4 Liquid dielectrics 7/11
  7.4.1 Breakdown mechanisms in liquids 7/11
  7.4.2 Insulating oils 7/11
  7.4.3 Inhibited transformer oil 7/13
  7.4.4 Synthetic insulating liquids 7/13

7.5 Semi-fluid and fusible materials 7/13
  7.5.1 Bitumens 7/14
  7.5.2 Mineral waxes and blends 7/14
  7.5.3 Synthetic waxes 7/14
  7.5.4 Natural resins or gums 7/14
  7.5.5 Miscellaneous fusible compounds 7/15
  7.5.6 Treatments using fusible materials 7/15
  7.5.7 Synthetic resins 7/15
  7.5.8 Thermoplastic synthetic resins 7/17
  7.5.9 Thermosetting synthetic resins 7/18
  7.5.10 Encapsulation 7/19

7.6 Varnishes, enamels, paints and lacquers 7/19
  7.6.1 Air-drying varnishes, paints, etc. 7/20
  7.6.2 Baking varnishes and enamels 7/20
  7.6.3 Solventless varnishes 7/20
  7.6.4 Silicone varnishes 7/20
  7.6.5 Properties of varnishes, etc. 7/20

7.7 Solid dielectrics 7/20
  7.7.1 Breakdown in solids 7/20
  7.7.2 Rigid boards and sheets 7/21
  7.7.3 Tubes and cylinders 7/23
  7.7.4 Flexible sheets, strips and tapes 7/24
  7.7.5 Sleevings, flexible tubings and cords 7/27
  7.7.6 Wire coverings 7/28
  7.7.7 Moulded and formed compositions, plastics, ceramics, etc. 7/28
  7.7.8 Methods of moulding and forming materials 7/28

7.8 Composite solid/liquid dielectrics 7/30
  7.8.1 Breakdown mechanisms in composite dielectrics 7/30
  7.8.2 Oil/paper systems 7/30

7.9 Irradiation effects 7/30
  7.9.1 Type of radiation 7/30
  7.9.2 Irradiation effects 7/31

7.10 Fundamentals of dielectric theory 7/32
  7.10.1 Basic definitions 7/32
  7.10.2 Types of dielectrics 7/32
  7.10.3 Polarisation in dielectrics 7/33
  7.10.4 Quantification of dielectric polarisation 7/33

7.10.5 Properties of dielectric materials 7/33
7.10.6 Example of ferroelectric material: Barium titanate and its applications 7/34
7.10.7 Frequency response of dielectrics 7/35
7.11 Polymeric insulation for high voltage outdoor applications 7/35
  7.11.1 Materials 7/35
  7.11.2 Hydrophobicity loss and recovery 7/35
  7.11.3 Degradation ageing factors of polymeric insulator surfaces 7/35

7.1 Insulating materials

Electrical insulating materials can be solid, liquid or gaseous, often in combination such as in oil-impregnated paper. The materials may be organic or inorganic and natural or synthetic. Both the electrical and mechanical properties of the materials are important, the variation of these properties with temperature being particularly important. In some applications the variation in electrical properties with frequency is significant and the chemical properties of the materials can often be vital because of the problems of compatibility between materials and of the behaviour of a material under fire conditions.

7.1.1 Classification

Insulating materials, especially those used in generators, motors, transformers and switchgear, are often classified on the basis of their thermal stability according to the scheme described in BS 2757:1986 and IEC 60085:1984. This scheme uses nine temperature classes, allocating materials to a class with 'temperature limits that will give acceptable life under usual industrial conditions of service'. These standards recommend that temperatures above 250°C should increase in steps of 25°C and the class is designated accordingly.

Temperature limits

Class	Y	A	E	B	F	H	200	220	250
Temperature (°C)	90	105	120	130	155	180	200	220	250

Examples of materials in each class are given below.

Class Y: Unimpregnated paper, cotton or silk, vulcanised natural rubber, various thermoplastics that have softening points that would only permit their use up to 90°C. Aniline and urea formaldehydes.

Class A: Paper, cotton or silk impregnated with oil or varnish, or laminated with natural drying oils and resins or phenol formaldehyde. Polyamides. A variety of organic varnishes and enamels used for wire coating and bonding.

Class E: Polyvinyl formal, polyurethane, epoxy resins and varnishes, cellulose triacetate, polyethylene terephthalate, phenol formaldehyde and melamine formaldehyde mouldings and laminates with cellulosic materials.

Class B: Mica, glass and asbestos fibres and fabrics bonded and impregnated with suitable organic resins such as shellac bitumen, alkyd, epoxy, phenol formaldehyde or melamine formaldehyde.

Class F: As class B but with resins that are approved for class F operation such as alkyd, epoxy alkyd and silicone alkyd.

Class H: As class B but with silicone resins or other resins suitable for class H operation. Silicone rubber.

Classes 200, 220, 250: Mica, asbestos, ceramics and glass alone or with inorganic binders or certain silicone resins. Polytetrafluoro-ethylene.

The allocation of materials to classes such that their life will be adequate under usual industrial conditions means that the materials may not give adequate life if the service is unusually severe, e.g. equipment normally operated very near to full load. Conversely, it may be economical to use materials of a lower temperature class for equipment operated infrequently or normally operated at very low load.

7.1.2 Temperature index

Various suggestions have been made for alternative temperature classification systems, especially as techniques such as cross-linking enable the thermal stability of materials to be significantly improved and new high-temperature materials are constantly being developed. IEEE 98:1984 has withdrawn the term 'Temperature Classification' but the institution has suggested that a 'temperature index' related to the temperature capability of the material should be assigned to a material on the basis of experience or comparison with materials that have established indices. The index would be based on the life of the material in particular environmental conditions and would preferably be a number chosen from the series 90, 105, 130, 155, 180, 200, 220. For temperatures above 250°C, no index has been established yet.

7.1.3 Effect of frequency

Insulating materials can have a very large variation in the dielectric loss and permittivity of the material with the frequency of the applied signal. Whilst this is important for insulation used in high-frequency electronics, it is not normally important in conventional power equipment, except for some condition of very fast transients. *Table 7.1* shows the loss tangents (tan δ) at 50 Hz, 1 kHz and 1 MHz. Fortunately, the losses generally decrease with frequency for the materials that are used in electrical power insulation. Occasionally there can be significant changes at low frequencies and polymethylmethacrylate has a peak in dielectric loss at around 2 Hz, depending on temperature, where it increases by a factor of 3.

7.1.4 Fire behaviour

Despite the extensive precautions that are normally taken to prevent fire in electrical installations, fires do occasionally occur. There may be hundreds of wires bunched together in ducting and this presents serious problems in the event of a fire. One problem is that if fire propagates along the insulation, electrical systems will fail completely and these may be essential to the orderly shut-down of the equipment, or they may be responsible for important telecommunication systems. Another problem is that the fumes that are produced by combustion of the insulating material may be toxic and may also cause corrosive damage to neighbouring materials. It is therefore important that electrical insulation is assessed for behaviour in the event of fire. Some materials that have been used in the past are liable to produce toxic fumes. The former use of polychlorinated biphenyls as transformer insulation is a particular cause for concern.

Ease of ignition, or flammability, has traditionally been considered the most important property of a material when assessing fire hazards, but it has now been realised that this is only one of several factors that must be considered. The amount of heat, smoke and toxic gases released by the material and the way that the flame spreads in the material are all important. The traditional tests use very small quantities of the material which means that the information obtained is of limited value. More realistic tests require more specialised facilities and tend to be expensive to perform.

Gaseous insulation does not normally present any fire problem, although sulphur hexafluoride can decompose into toxic fractions under certain conditions. Commonly used insulating liquids, such as transformer oil, are flammable and so present a fire hazard. This is often unacceptable for transformers or switchgear situated in substations in blocks of offices or flats and alternative designs using air insulation or non-flammable liquids must be used. The same problems occur for equipment to be used in hazardous environments such as mines. Polychlorinated biphenyls

Table 7.1 Representative properties of typical insulating materials*

Insulant	*n*	ϵ_r	*tan δς* *50* Hz	 *1* kHz	 *1* MHz
Vacuum	∞⇐	1.0	0	0	0
Gases					
Air	∞⇐	1.0006	0	0	0
Sulphur hexafluoride	∞⇐	1.002	0	0	0
Liquids					
Mineral insulating oil	11–13	2–2.5	0.0002	0.0001	—
Dodecylbenzene	12–13	2.1–2.5	0.0002		
Organic esters	10–12	2.9–4.3	0.001		
Polybutenes	12–14	2.1–2.2	0.0005		
Silicone fluids	12	2.7	0.0001		
Solids					
Paraffin wax	14	2.2		0.0003	0.0001
Bitumen	12	2.6	0.008		
Pressboard	8	3.1	0.013		
Bitumen-asbestos	10–11				0.08
Paper: dry	10	1.9–2.9	0.005	0.007	
oil-impregnated	14–16	3.2–4.7	0.002		
Cloth: varnished cotton	13	5	0.2	0.15	
Ethyl cellulose	11	2.5–3.7	0.02	0.03	0.02
Cellulose acetate film	13	4–5.5	0.023	0.04	
Cellulose acetate moulding	10	4–6.5	0.016	0.03	0.06
Synthetic-resin (phenol) bonded paper	11–12	4–6	0.02	0.03	0.04
Mica	10	5.5–7	0.0005	0.0005	0.0005
Nylon	11	3.8			0.03
Phenol–formaldehyde	10	4–7	0.05	0.03	0.02
cast	9	7–11	0.1	0.2	0.25
Polystyrene	15	2.6	0.0002	0.0002	0.0002
Polyethylene	15	2.3	0.0001	0.0001	0.0001
Polypropylene	15	2.3	0.003		0.0003
Polytetrafluoroethylene	15	2.1	0.0002		0.0002
Methylmethacrylate	13	2.8	0.06	0.03	0.02
Synthetic-resin compounds					
Phenol formaldehyde mineral filled	10–12	5	0.015	0.015	0.01
Urea formaldehyde mineral filled	10	5–8	0.1	0.1	0.038
Polyvinyl chloride	11	5–7	0.1	0.1	

*Volume resistivity $\rho = 10^n$ Ω-m; the value of *n* is tabulated. Relative permittivity ϵ_r. Loss tangent tan δ.

were once used as non-flammable insulation in these applications, but the discovery of their toxic nature has required a search for alternative non-flammable liquids. Silicone oils are used but, although these are less flammable than transformer oil, they are still flammable.

Fire tests that are used to try to give more relevant fire hazard information for liquids use pools of the liquid from about 15 cm diameter to 9 m^2 of burning liquid area.

Much of the work on fire behaviour is concerned with insulation for cables. Here the standard tests such as BS 6387:1994 only require a flame of between 650 and 950°C, depending on the fire resistance category, to be applied to about 600 mm of the cable for 3 h (only 20 min for the highest temperature) without the insulation failing. However, more stringent tests involving longer lengths of cable on cable trays in vertical ducts are often required so that the conditions met in situations such as power stations can be reproduced more accurately. A typical test involves 8 m of cable on a cable rack ignited by an 88 kW methane flame with a forced air draft to propagate the flame. Measurements are made of the distance that the flame spreads and the optical density of the smoke produced.

Tests are now being made of the heat evolution of materials in a fire. There is still a lack of agreement about exactly what tests are most appropriate and how reliably small-scale tests can be scaled to possible full-scale fire scenarios. There is particular lack of agreement about toxicity testing.

7.2 Properties and testing

The properties of insulating materials fall into the following categories:

(1) physical,
(2) mechanical,
(3) electrical, and
(4) chemical.

Insulating materials may have to operate in the vicinity of apparatus producing high intensity radiation such as nuclear reactors, isotopes, microwave and electron generators, and considerable work has been done on the properties of materials under these conditions.

7.2.1 Physical properties

7.2.1.1 Density

This is of importance for varnishes and oils. The density of solid insulants varies widely; in a few cases it is the measure of relative quality (as in pressboard).

7.2.1.2 Moisture absorption

This usually causes serious depreciation of electrical properties, particularly in oils and fibrous materials. Swelling, warping, corrosion and other effects often result. Under severe conditions of humidity, such as occur in mines and in tropical climates, moisture sometimes causes serious deterioriation; products made from linseed-oil varnishes, for example, are prone to complete destruction of the varnish film in a damp atmosphere. Fungus growth and electrolysis are other examples of effects due to moisture.

It is usual to determine the absorbency of solid materials by ascertaining the weight of water absorbed by a standard specimen when immersed for a specified period: however, the quantity of water absorbed is not a reliable criterion of the *electrical* performance of a material if taken in isolation. Some British Standard methods require that electrical tests, especially those for insulation resistance and loss tangent, be carried out immediately after the samples have been removed from water following a period of immersion of 24 h.

7.2.1.3 Thermal effects

These often seriously influence the choice and application of insulating materials, important features being: freezing point (of gases and liquids); melting point (e.g. of waxes); softening or plastic yield temperatures; flash point of liquids; ignitability, flammability, ability to self-extinguish if ignited; resistance to electric arcs; liability to carbonise or track; specific heat; thermal resistivity or conductivity; and coefficient of expansion.

7.2.1.4 Ageing

Although ageing has been placed in the physical-properties section, ageing involves changes in the physical, mechanical, electrical and chemical properties of the material when subjected to prolonged thermal and electrical stresses. In some applications mechanical stress may also be important.

A major part of life testing is the determination of the maximum temperature that a material, or combination of materials, can withstand for a long period without serious degradation of important properties. It is necessary for all components of an insulation system to be present during ageing tests because of the possibility of *compatibility* problems. Testing of this type is generally carried out on models made to reproduce, as far as possible, the conditions met in service. Such model investigations, often called 'functional testing', are generally accelerated by using temperatures considerably above those envisaged for service; but, provided that agreed procedures are used, it is often possible to extrapolate long-term results from comparatively short tests. A statistical technique that is often used for extrapolation is called *Weibull* statistics.

Thermal analysis tests that are routinely used to evaluate insulating materials are shown in *Table 7.2*. There is increasing interest in conducting ageing tests with a combination of heat and electric stress in order better to predict lifetime since there is an interaction between the decomposition products of electrical discharge and products released due to purely thermal effects in the material. Both increased temperature beyond that encountered in service and increased electric stress beyond the service stress can be used to accelerate ageing, but it is difficult to verify the prediction of service lifetime from the effect of a combination of acceleration techniques.

7.2.1.5 Miscellaneous characteristics

These include viscosity (of liquids such as molten bitumen), moisture content (of wood, pressboard, etc.), uniformity of thickness and porosity (of papers, porcelain, etc.).

7.2.2 Mechanical properties

The usual mechanical properties of solid materials are of varying significance in the case of those required for insulating purposes, *tensile* strength, *cross-breaking* strength, *shearing* strength and *compressive* strength often being specified. Owing, however, to the relative degree of inelasticity of most solid insulations and the fact that many are quite brittle, it is frequently necessary to pay attention to *compressibility,*

Table 7.2 Thermal-analysis techniques

Technique	*Parameter measured*	*Applied stress*
Differential scanning calorimetry (DSC)	Energy necessary to establish zero temperature difference with a reference material	Environment heated or cooled at a controlled rate
Differential thermal analysis (DTA)	The difference in temperature between the material and a reference material	Environment heated or cooled at a controlled rate
Evolved-gas analysis (EGA)	Nature and quantity of volatile products formed	Heating
Thermally stimulated current measurement (TSC)	Polarisation or depolarisation current	Temperature change while electric field is applied
Thermogravimetry (TG)	Weight change with time or temperature	Heating or cooling
Thermomechanical analysis (TMA)	Mechanical strain	Mechanical vibration with heating or cooling

deformation under bending stresses, *impact* strength and *extensibility; tearing* strength and ability to fold without damage are important properties of thin-sheet insulations such as papers, pressboards and varnished cloths.

Methods of test for the above properties are given in British Standards.

Many other mechanical features of insulating materials have to be considered, for example: machinability (especially as regards drilling and punching) and resistance to splitting, the latter being of particular importance in the case of laminated materials, wood and pressboards.

7.2.3 Electrical properties

The essential property of a dielectric is, of course, that it shall insulate. But there are properties other than resistivity that determine the insulation value: these are the electric strength, permittivity and loss tangent.

7.2.3.1 Resistivity

This concerns volume resistivity (a bulk property) and surface resistivity (concerning leakage current across the insulator surface between electrodes having a potential difference). The former is specified in ohm-metres (or megohm-metres) and the latter in ohms per square: the surface resistance between opposite sides of a square surface is independent of the size of the square. The properties are affected by surface or bulk moisture, so that measurements of insulation resistance of pieces of material or of insulated systems are often used to assess the state of dryness. Values of volume resistivity are given in *Table 7.1.*

7.2.3.2 Electric strength

Electric strength (or dielectric strength) is the property of an insulating material which enables it to withstand a given electric field magnitude without failure. It is usually expressed in terms of the minimum electric field magnitude (i.e. potential difference per unit thickness) that will cause failure or 'breakdown' of the dielectric under specified conditions, e.g. shape of electrodes, temperature and method of application of voltage, as these and several other features all influence the liability of the material to fail under electric stress. It is, therefore, important to state most of these conditions when quoting values of electric strength, and they have been standardised accordingly by BSI and others. The standard method for testing oils for electric strength is given in BS 148:1998, and that for proof tests on bitumen-based filling compounds in BS 1858:1975. Details of the standard method for proof tests on solid insulations, such as moulded compounds and sheet materials, are given in BS 5734:1990, BS 6091:1995 and BS EN 60243:1998.

The electric strength of most materials falls with increasing temperature and it is usual to carry out tests for this property at suitably elevated temperatures.

Other features which vitally affect the apparent electric strength are: the sharpness or radius of edges of electrodes; the waveform of the voltage (as breakdown is dependent on the *peak* value); the rate of increase in voltage and the time any voltage stress is maintained; the moisture content of the material; the thickness of specimen tested and the medium (usually air or oil) in which the test is made. Comparisons of electric strength are made generally by determining the electric stress that will cause failure 1 min after its application. Specifications frequently call for a *proof test,* the material being required to withstand for, say, 1 min a specified electric stress under controlled conditions.

In view of all the features which affect the apparent electric strength of dielectrics it is preferable to obtain comparative values, say, at a range of temperatures, thicknesses and test durations. Tests may be made with alternating or direct voltages; and it is now becoming more usual to test with lightning or switching-impulse voltages if the material is liable to sustain transient voltages in operation such as occur with overhead-line insulators, switchgear, power transformers and some machine windings. The object is to determine the highest stress that a material or assembly will withstand indefinitely. An indication can be obtained from a voltage/time curve (*Figure 7.1*) plotted from the stresses that cause breakdown in measured periods. The safe operating stress is then settled by experience, the use of safety factors, and the data from comparative tests. *Figure 7.1* gives typical results of the variation of electric strength with thickness of specimen and with temperature.

7.2.3.3 Surface breakdown and flashover

When a high-voltage stress is applied to conductors separated only by air where they are closest together, and the stress is increased, breakdown of the intermediate air will take place when a certain stress is attained, being accompanied by the passage of a spark from one conductor to the other, i.e. the electric strength of the air has been exceeded. If the stress is sustained, this may also be followed by a continuous arc. The voltage at which this occurs is the *sparkover* or *flashover value.* Similar conditions can be obtained with oil as the insulant when a spark passes through the oil between the conductors.

In electrical assemblies where the live parts are separated by both solid insulation the ambient air, failure may take place either by breakdown of the solid material or by flashover through the air. Often the process involves surface leakage, deterioration and surface flashover. This phenomenon is generally due to the nature and design of the metal parts, as sharp edges of nuts and washers (for example) give local concentrations of stress. In addition, the onset of surface discharges at metal edges (which can initiate breakdown) is influenced by the permittivity of the dielectric material. The higher the permittivity, the lower the voltage at which flashover is likely to occur. Pollution on the insulation surface can reduce flashover voltage by a factor of 100. Insulating materials are sometimes tested for surface breakdown or flashover between two electrodes on a typical surface but, unless the material itself or its surface is poor electrically, flashover in air takes place in preference, usually at values of about 20 kV r.m.s. for 25 mm distance between two 38 mm diameter electrodes with fairly sharp edges.

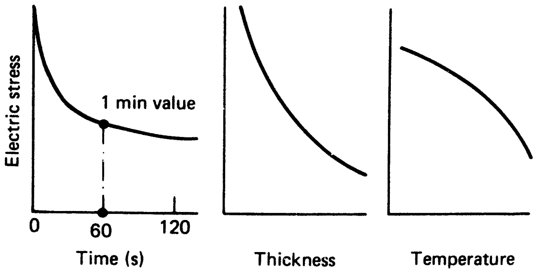


Figure 7.1 Effect of time, thickness and temperature on electric strength

7.2.3.4 Tracking

Leakage along the surface of a solid insulating material, often a result of surface contamination and moisture or of discharges on or close to the surface, may result in carbonisation of organic materials and conduction along the carbonised path. This is known as 'tracking'. It is usually progressive, eventually linking one electrode to another and causing complete breakdown along the carbonised track. The methods for evaluating the resistance to tracking and erosion of electrical insulating materials are given in BS 5604:1986 and its IEC equivalent IEC 60587:1984.

7.2.3.5 Permittivity

This property is specific to a material under given conditions of temperature, frequency, moisture content, etc. When two or more dielectrics are in series and an electric stress is applied across them, the voltage gradient across each individual dielectric is inversely proportional to its permittivity. This is particularly important when air spaces exist in solid and liquid dielectrics, as the permittivities of these are always higher than that of air, hence the air is liable to have the higher stress and may fail and cause spark-over through the air space in consequence. The permittivity of dielectric materials is strongly dependent upon frequency with a tendency to fall to low values at higher frequencies. In the case of ferroelectric materials, increasing the temperature will lead to an increase in permittivity up to the 'Curie Point' after which permittivity falls rapidly with temperature.

Values of permittivity for some insulating materials are given in *Table 7.1*.

7.2.3.6 Dielectric loss

A capacitor with a perfect dielectric material between its electrodes and with a sinusoidal alternating voltage applied takes a pure capacitive current $I = \omega CV$ with a leading phase angle of 90°. In a practical case, conduction and hysteresis effects are present, the phase angle is less than 90° by a (normally) small angle δ. The power factor, no longer zero, is given by $\cos(90° - \delta) = \sin\delta \simeq \tan\delta$: the latter is called the *loss tangent*. The power loss is, to a close approximation, $P = V^2\omega C \tan\delta$ where $\omega = 2\pi f$: it is proportional to the square of the voltage and to the product $\epsilon\tan\delta$, because the absolute permittivity ϵ determines the capacitance of a system of given dimensions and configuration.

The loss tangent varies, sometimes considerably, with frequency and also with temperature; values of $\tan\delta$ usually increase with rise of temperature, particularly when moisture is present, in which case the permittivity also rises with the temperature, so that total dielectric losses are often liable to a considerable increase as the temperature rises. This is very often the basic cause of electric breakdown in insulation under a.c. stress, especially if it is thick, as the losses cause an internal temperature rise with consequent increase in the dielectric loss, this becoming cumulative and resulting in thermal instability and, finally, breakdown.

Permittivity and loss tangent are usually determined by means of a Schering bridge (BS 7663:1993). For power devices such as cables and bushings, the test is made at 50 Hz; but for high-frequency equipment it is necessary to determine loss tangent and permittivity at much higher frequencies. BS 2067:1953 and BS 4542:1970 cover such measurements by the Hartshorn and Ward method at frequencies between 1 kHz and 100 MHz. Other methods are available for other frequencies (see IEC 60250:1969). Typical values of loss tangent and permittivity for some of the principal insulating materials used for high voltages and for high frequencies are given in *Table 7.1*.

7.2.4 Chemical properties

The chemical and related properties of insulating materials of importance may be grouped as follows:

(1) resistance to external chemical effects,
(2) effects on other materials, and
(3) chemical changes of the insulating material itself.

Under (1) there are such properties as resistance to:

(a) the effect of oil on materials liable to be used in oil (in transformers and switchgear), or to be splashed with lubricating oil;
(b) effects of solvents used with varnishes employed for impregnating, bonding and finishing;
(c) attack by acids and alkalis, e.g. nitric acid resulting from electrical discharge, acid and alkali vapours and sprays in chemical works, and deposits of salts from sea spray;
(d) oxidation, hydrolysis and other influences of atmospheric conditions, especially under damp conditions and in direct sunlight; and
(e) effects of irradiation by high-energy nuclear radiation sources, e.g. neutrons, β particles and γ rays.

In group (2), typical effects of the insulating materials on other substances with which they may be used are:

(a) direct solvent action, e.g. of oils and of spirits contained in varnishes, on bitumen and rubber; corrosion of metals in contact with the insulation; and attack on other materials by acids and alkalis contained in the insulating materials in a free state;
(b) effects of impurities contained in the insulation; and
(c) effects resulting from changes in the material, for example acids and other products of decomposition and oxidation affecting adjacent materials.

These effects are generally referred to under the heading 'compatibility'. If meaningful test results are to be obtained, all components of an insulation system must be present and they must have been treated in the same way as will be used in manufacture.

Group (3) includes such features as:

(a) oxidation resulting from driers included in varnishes;
(b) deterioration due to acidity (e.g. in oils, papers and cotton products);
(c) chemical instability of synthetic resins;
(d) self-polymerisation of synthetic compounds; and
(d) vulcanisation of rubber–sulphur mixtures.

Most of these chemical properties are determined by well-known methods of chemical analysis and test. The principal tests are for acidity and alkalinity, pH value, chloride content in vulcanised fibre, and conductivity of aqueous extract (for presence of electrolytes). Some of these are dealt with in BS 5591:1978, BS 2782:1991, BS 5626–2:1979 and BS EN 1413:1998.

Increasing attention is being paid to chemical features of the raw materials and processes used in the manufacture of insulating materials—particularly varnishes, synthetic resins and all manner of plastics—and much research work is being carried out on these features and on the correlation of the chemical structure of dielectrics with their physical, electrical and mechanical properties.

7.3 Gaseous dielectrics

7.3.1 Breakdown mechanisms in gases

As a gas is a highly compressible medium, breakdown processes in gases depend on the density of the gas and values quoted for air in Section 7.3.2 must be corrected for the air density (*d*) relative to normal temperature and pressure (20°C and 1013 mbar, respectively):

$$d = \frac{p}{1013}\,\frac{273+20}{273+t}$$

where p is the pressure in millibars and t is the temperature in degrees Celsius.

A discharge in a gas and subsequent development into a visible spark or flashover starts with the production of electrons in the gas by emission from one of the electrodes, or even from cosmic rays. The initial eletrons are then multiplied by various processes of ionisation that give a growth in the current and lead, ultimately, to breakdown. In most gases, when an electron collides with a neutral molecule an extra electron and a positive ion will be produced, provided the energy of the original electron is higher than the ionisation energy of the molecule (*Figure 7.2*(*a*)) but in *electronegative* gases such as sulphur hexafluoride an electron can be captured by the molecule to give a negative ion (*Figure 7.2*(*b*)). This process is the opposite of electron multiplication so these gases have a high breakdown strength.

The breakdown strength of a gas is affected by the uniformity of the electric field, the waveform of the applied voltage, the support insulators and solid particle contaminants, in addition to the gas density. The effect of non-uniform field is particularly significant for d.c. insulation as the breakdown voltage is quite different depending on whether the point in a point-plane system is positive or negative. For large gaps the breakdown strength for the negative point is more than twice the voltage obtained when the point is positive.

In practical apparatus the breakdown strength of compressed gas can be reduced severely by the presence of dust and solid particles. Particles near an electrode can induce a spark at a substantially lower voltage than for a particle-free situation. In one study on sulphur hexafluoride, breakdown voltages were reduced to between 20% and 90% of the particle-free values, depending on the number and size of the particles present.

7.3.2 Air

Air is the most important gas used for insulating purposes, having the unique feature of being universally and immediately available at no cost. The resistivity of air can be considered as infinite under normal conditions when there is no ionisation. There is, therefore, no measurable dielectric loss, negligible tan δ, and a relative permittivity (for all practical purposes) of unity. The electric strength under normal atmospheric conditions is 30 kV/cm (peak) for a uniform field. In a practical airgap the voltage gradient is a maximum at the electrode surfaces. The sparkover (breakdown) voltage of an air gap is therefore a non-linear function of its length. The gap geometry and configuration affect the breakdown voltage, and empirical expressions have been derived to account for the 'gap factor'.

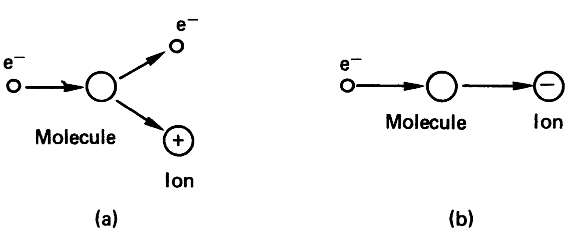


Figure 7.2 Electron–molecule collisions in gases: (a) normal gases; (b) electronegative gases

Table 7.3 Typical breakdown voltages in air under normal atmospheric conditions (kilovolt peak at 50 Hz)

Two-electrode system	*Spacing or gap* (mm)				
	10	*50*	*100*	*200*	*300*
Spheres, diameter 1.0 m	—	137	266	503	709
0.25 m	31	137	243	363	—
Needle points	13	50	78	127	178
Parallel wires, diameter 8.25 mm	—	38	57	83	117

Concentric cylinders: outer/inner radius (mm)	38/1.3	38/11	67/17	67/2
Breakdown voltage (kV)	26	55	88	103

Partial breakdown of air, locally, often occurs when the voltage gradient in a particular region exceeds the critical value for air. This happens readily at points of electric flux concentration, e.g. sharp edges of metal parts. If this local breakdown becomes unstable—as it will when the voltage between conductors is increased sufficiently—sparkover will occur. This may be an isolated spark from one conductor to the other, and the intervening air then re-heals itself. If the voltage is maintained (or increased), the spark may be followed by a continuous stream of sparks.

Typical values of breakdown voltages for gaps of different forms and sizes of electrodes (under normal atmospheric conditions) are given in *Table 7.3*.

Partial or complete breakdown of air in gaps can be influenced by suspending sheets of material at particular places in the electric field. In some cases, the sheets may be of metal, and in others of insulating materials; even woven fabrics can have an effect. The effect of this barrier is generally greater for more divergent fields. This solution can be useful where clearances are limited inside equipment, or particularly in high-voltage test areas.

For plane gaps, all gases exhibit a minimum breakdown voltage known as the Paschen minimum; this occurs at a given value of the product *Pd* of absolute gas pressure and gap length. For air this minimum occurs at $Pd \simeq 6$ (in torr-millimetre). Thus as the value of *Pd* is reduced from a higher region the breakdown voltage falls to the minimum value quoted, and further decrease in either *P* or *d* results in an increase in the voltage required to break down the gap. This explains why quite small gaps under conditions of high vacuum can sustain very high voltages. *Table 7.4* gives the values of *Pd* and the minimum voltages for several gases.

7.3.2.1 Sphere gaps

As the electric strength of air is dependable, standard sphere gaps can be used as reliable and accurate means for measuring high voltages (*Table 7.5*), particularly where peak voltages are to be measured, as it is, of course, the *peak* value

Table 7.4 Minimum breakdown voltage for gases: 1 mmHg, 1/760 atm

	CO_2	*Air*	O_2	N_2	H_2	*Ar*	*He*	*Ne*
Pd (T-mm)	5	6	7	7.5	12.5	15	25	30
Direct voltage (V)	420	330	450	275	295	265	150	244

Table 7.5 Sphere-gap breakdown voltages (kilovolts at peak)*; BS 358:1960

Gap (mm)	*Sphere diameter* (m)							
	0.02	*0.0625*	*0.125*	*0.25*	*0.5*	*0.75*	*1.0*	*1.5*
0.5	2.8	—	—	—	—	—	—	—
1	4.7	—	—	—	—	—	—	—
1.5	6.4	—	—	—	—	—	—	—
2	8.0	—	—	—	—	—	—	—
4	14.4	14.2	—	—	—	—	—	—
5	17.4	17.2	16.8	—	—	—	—	—
6	20.4	20.2	19.9	—	—	—	—	—
8	25.8	26.2	26.0	—	—	—	—	—
10	30.7	31.9	31.7	31.7	—	—	—	—
15	(40)	45.5	45.5	45.5	—	—	—	—
20	—	58.5	59.0	59.0	—	—	—	—
30	—	79.5	85.0	86.0	86	86	86	—
40	—	(95)	108	112	112	112	112	—
50	—	(107)	129	137	138	138	138	138
100	—	—	(195)	244	263	265	266	266
150	—	—	—	(314)	373	387	390	390
200	—	—	—	(366)	460	492	510	510
300	—	—	—	—	(585)	665	710	745
400	—	—	—	—	(670)	(800)	875	955
500	—	—	—	—	—	(895)	1010	1130

*In air at 20°C, 1013 mbar. One sphere earthed. For alternating voltages of either polarity; and for standard negative impulse voltages (50% breakdown value). Figures in brackets not reliable.

which determines the breakdown. Standard sizes of spheres are generally used as electrodes, as, provided the size is appropriate and proper precautions are taken (e.g. to avoid effects such as those due to the proximity of other objects and uncontrolled irradiation of the gap by other discharges), clean, smooth, metal spheres are most reliable as a means of determining high voltages; this is largely due to the absence of corona prior to flashover if the spacing does not exceed the radius of the spheres. Sphere gaps are suitable also for the measurement of impulse voltages. Voltages of about 2 kV and upwards can be measured reliably. BS 358:1960 gives detailed information on the effects of humidity, air density (or barometric pressure), etc. The effect of density is pronounced in the case of equipment used at high levels above 1000 m, in aircraft where altitudes up to 15 km may be met, or in spacecraft where outer space is an almost perfect vacuum.

7.3.2.2 Needle gaps

Humidity has here a strong influence on breakdown voltage where the electrode shape leads to field concentration. For this reason, as well as that of a degree of frequency dependence, needle gaps are unreliable for high-voltage measurements. Rod gaps (e.g. 12 or 16 mm square-section rods with sharp corners) are used for chopping impulse voltages, but with these too a humidity correction is necessary.

7.3.2.3 Corona

This term is used to describe the glow or 'brush' discharge around conductors when the air is stressed beyond the ionisation point without flashover developing. It is of more or less serious consequence according to the application concerned. It causes a certain amount of energy loss with alternating current, which may become appreciable on high-voltage transmission lines. It produces radio interference and may initiate surface deterioration and breakdown on solid insulation surfaces. Corona is also known to produce secondary chemical effects.

In thin films, particularly in spaces between layers of sheet insulation, air can readily become ionised due to the electric stress across such spaces exceeding the critical value. This is often due to the fact that, with dielectrics in series, the stress in each section is inversely proportional to its permittivity. When the critical stress in the air or gas is exceeded, discharges occur (often called corona, ionisation, glow or brush discharges) and this causes splitting up of the gas molecules. In air this leads to the formation of ozone and nitrogen oxides which in the presence of moisture produce nitric acid. The ozone has, of course, a strong oxidising effect, but the more serious chemical effects of ionisation are those due to the nitrogen products, as the nitric acid attacks most of the organic insulating materials and causes corrosion of metal parts. The action of either or both the ozone and the nitrogen oxides on many materials is to cause decomposition and often the formation of acids; for example, oxalic acid, known for causing brittle fracture in polymeric insulators, by the oxidation of cellulose materials, and acetic acid from the decomposition of cellulose acetate.

In addition to the chemical effects, discharges in spaces, films or cavities within dielectrics can have serious consequences mainly due to the high energy in some of the individual discharges. Mechanical electrical and thermal damage can occur and breakdown in service may result after long periods. There has been considerable advance in the methods for detecting the presence of such partial discharges in various types of equipment especially where oil-impregnated paper dielectrics are used. Discharges within air or gas films in such material can cause severe damage often followed by complete breakdown.

7.3.2.4 Compressed air

This is used as the arc-extinguishing medium and dielectric insulation in air-blast circuit-breakers.

7.3.3 Nitrogen

Instead of air, which is a mixture of approximately 21% oxygen and 79% nitrogen, nitrogen alone is sometimes used when there is a risk of oxidation of another material such as insulating oil. Nitrogen is often used in gas-filled high-voltage cables, as an inert medium to replace air in the space above the oil in some transformers, in low-loss capacitors for high-voltage testing, etc. There is no appreciable difference between the electric strength of nitrogen and that of air. Some results relating to the electric strength of nitrogen for uniform fields at pressures above atmospheric up to 20 atm are given in *Table 7.6*. Included are some similar results for carbon dioxide.

Table 7.6 Breakdown voltages of nitrogen and carbon dioxide (kilovolts at peak) under uniform field conditions

Gas	*Gap* (mm)	*Pressure (abs)* (atm)				
		3	*5*	*10*	*15*	*20*
Nitrogen	1	10	15	27	35	45
	8	90	123	180	220	—
Carbon dioxide	1	13	17	27	38	52
	8	85	115	200	260	—

7.3.4 Sulphur hexafluoride

Sulphur hexafluoride is an electronegative gas which has come into wide use as a dielectric (in X-ray equipment, in waveguides, coaxial cables, transformers, etc.) and as an arc-quenching medium in circuit-breakers. Its electric strength is of the order of 2.3 times that of air or nitrogen, and at a pressure of 3–4 atm it has an electric strength similar to that of transformer oil at atmospheric pressure. The gas sublimes at about – 64°C and it may be used at temperatures up to about 150°C. Although the gas is considered to be non-toxic, non-flammable and chemically inert, under the influence of arcs or high-voltage discharges, there may be some decomposition with consequent attack on certain insulating materials and metals, and more importantly some recent environmental concerns. In circuit-breakers this problem is overcome by careful selection of materials (e.g. polytetrafluoroethylene for interrupter nozzles) and by the use of filters and absorbents to remove the products of decomposition after circuit interruption. Some figures relating to the electric strength of sulphur hexafluoride and mixtures of this gas with nitrogen are given in *Table 7.7*.

Numerous other electronegative gases such as perfluoropropane (C_3F_8), octafluorocyclobutane (C_4F_8) and perfluorobutane (C_4F_{10}) have been developed, but few have found such widespread use as sulphur hexafluoride. The main interest for these gases is as dielectrics in transformers, waveguides, capacitors, etc., but one difficulty is that the temperature at which condensation occurs may not be sufficiently low for safety in outdoor equipment likely to remain un-energised for long periods. This problem can be overcome partly by fitting heaters or by using admixtures with a more volatile gas (such as nitrogen). Addition of nitrogen often improves some of the characteristics, while at the same time reducing the overall cost. Some of these gases can be used at temperatures well above 200°C.

Most of the fluorinated gases have an electric strength between two and five times that of air or nitrogen under the same conditions but, as with sulphur hexafluoride, care must be taken to prevent high-voltage discharges or arcs in the gas because of the dangers of producing decomposition products. Recent research efforts in the electrical power industry focussed on improving the use of gas mixtures (essentially 90% nitrogen and 10% SF_6), which are more environmentally acceptable. Gas insulated high voltage lines are now in operation using these gas mixtures.

Some electric machines and special devices have to operate in a gas other than air—for example, most refrigerator compressor motors operate in gaseous refrigerants mostly

Table 7.7 Breakdown voltage of nitrogen and sulphur hexafluoride (kilovolts at peak): direct voltage and uniform field

Gas	*Pressure* (atm)	*Positive polarity gap* (mm)					*Negative polarity gap* (mm)				
		5	*10*	*15*	*20*	*25*	*5*	*10*	*15*	*20*	*25*
Nitrogen	1	—	30	—	56	—	—	30	—	56	—
	2	30	55	—	100	—	30	54	—	100	—
	3	41	76	114	147	180	42	77	113	147	178
80% Nitrogen	1	38	74	111	145	178	38	74	111	146	178
20% Sulphur hexafluoride	2	72	143	212	—	—	72	142	210	—	—
	3	111	220	—	—	—	111	221	—	—	—
Sulphur hexafluoride	1	44	88	133	175	213	44	88	134	176	208
	2	85	171	252	—	—	84	171	251	—	—
	3	132	260	—	—	—	131	258	—	—	—

based on chlorofluorohydrocarbons (such as Arcton, Freon, etc.). These materials can act as solvents for some of the components used in insulating materials with consequent failure of the equipment due to blocked tubes and valves in the refrigerator circuit. Careful selection of materials for resistance to these fluids is essential.

7.3.5 Hydrogen

This gas is used as a cooling medium in some large turbo generators and synchronous motors; the main advantages are the efficient removal of heat and reductions in windage loss. Although there is a fire and explosion risk, troubles of this kind have been few during the many years that the gas has been used commercially for electrical machines. The electric strength of hydrogen at atmospheric pressure is about 65% that of air but most machines operate at pressures of 2–5 atm, and over this range the electric strength is higher than for air at atmospheric pressure. High-voltage discharges are thus not likely to be any more severe, and as discharges in hydrogen do not produce ozone or oxides of nitrogen, injurious effects are considered to be negligibly small.

7.3.6 Vacuum

Considerable investigation has been made into the utilisation of high vacua both for the insulation of equipment and as the interrupting medium in vacuum circuit breakers and contactors. The major advantage is due to the fact that very high electric field strengths can be achieved with a maximum operating pressure of 1 atm (negative), whereas with gases, very high operating pressures are generally essential and this complicates the mechanical design of the tank or other containing structure. Vast improvements in high-vacuum technology together with the need to replace oil-insulated switchgear has led to compact vacuum bottles used to retrofit old oil-filled circuit breakers.

7.4 Liquid dielectrics

The liquids which are most commonly used for electrical insulation are petroleum oils. For some applications these are being replaced by synthetic hydrocarbon oils, particularly as impregnant for oil-impregnated paper insulated power cables. Polychlorinated biphenyls (askarels) were widely used where non-flammable insulation was required for transformers, and for capacitor dielectrics. However, these have now been withdrawn for most applications because of environmental pollution effects and health risks. Silicone oils are now used for non-flammable transformer insulation. Capacitors often use silicone liquids or synthetic hydrocarbons as dielectrics, but various esters are now being introduced that offer a higher permittivity and hence a higher capacitance value for the same dimensions. An insulating liquid that is sometimes used is castor oil.

The principal uses of liquid dielectrics are:

(1) as a filling and cooling medium for transformers and some electronic equipment, and as a filling medium for capacitors, bushings, etc.;
(2) as an insulating and arc-quenching medium in switchgear;
(3) as an impregnant of absorbent insulation, e.g. paper, porous polymers and pressboard—these are used in transformers, switchgear, capacitors and cables; and
(4) as a heat transfer medium in addition to its insulation rôle in power cables, especially in force-cooled high-pressure oil-filled cables.

The important properties of the liquid used vary with the application, but they include electric strength, permittivity, chemical and thermal stability, gassing characteristics, fire resistance and viscosity.

7.4.1 Breakdown mechanisms in liquids

Breakdown strengths of liquids are very dependent on liquid purity and all the breakdown mechanisms that are met in the practical use of liquids as electrical insulation are contamination mechanisms. Breakdown is caused by one of three contaminants: particles, water, or gas bubbles.

Particle-induced breakdown requires that there are dust particles, cellulose fibres from adjacent solid insulation or similar particles present in the liquid. If the particle has a higher relative permittivity than the liquid, electrostatic theory tells us that there will be a force acting on the particle moving it towards the region of greatest stress between the electrodes. If the particle contains moisture the force will be larger because of the high ϵ_r value for water. Other particles will be attracted to the same region and they align end-to-end, eventually forming a bridge between the electrodes. Current flows along this bridge giving localised heating and breakdown.

Water itself is inevitably likely to be present in practical liquids. Careful procedures for filling equipment and maintaining desiccants at breathing points in the apparatus can normally keep moisture levels to less than 20 ppm. Any globule of water present in the liquid will become elongated in the field direction by the action of an applied field. Breakdown channels will propagate from the ends of the globule and produce total breakdown. The electric strength of an oil can be halved by the presence of 50 ppm of water. The presence of water will significantly increase the dielectric loss in the oil and reduce the breakdown strength.

Bubbles can be formed by gas pockets in pits or cracks on surfaces containing the liquid, or they can arise from dissociation of liquid molecules, or local liquid vaporization through electron emission from sharp points on an electrode. Such bubbles will become elongated by the field in a similar way to water globules. As the breakdown strength of the gas in the bubble will be much lower than that of the liquid, the field inside the bubble may exceed the breakdown strength of the vapour. This will give a spark in the bubble which may cause dissociation of some of the surrounding liquid to generate more gas. Eventually the bubble will become so large that a complete breakdown between the electrodes will ensue.

7.4.2 Insulating oils

The insulating oils used extensively are highly refined hydrocarbon mineral oils obtained from selected crude petroleum, and have densities in the range 860–890 kg/m^3 at 15°C. Oil for transformers and switchgear is dealt with in BS 148:1998. A number of special mineral oils are employed for impregnated paper capacitors and cables and others—usually of higher viscosity and flash point—for rheostats and for filling bus-bar chambers in switchgear. Typical properties are given in *Tables 7.3* and *7.9*.

7.4.2.1 Electric strength

This is a property involving similar phenomena to spark-over in gases. On raising the voltage between two electrodes

Table 7.8 Effect of contaminants on electric strength of mineral insulating oil: breakdown voltage in kilovolts r.m.s. (between 13 mm diameter spheres, 4 mm gap)

Contaminant	(g/m^3)	*Water present (parts/10 000)*			
		0.2	*1.0*	*2.5*	*5*
Clean oil	0	86	80	80	80
Cotton	0.02	68	36	30	28
	0.28	33	11	10	10
Pressboard fibres	0.08	82	64	56	54
	0.37	56	30	29	28
	1.4	26	12	11	11
Carbon	1.9	83	80	80	79
	35.0	73	70	70	69

in oil, electrical discharges may first appear in the space surrounding the electrodes—particularly at sharp corners—and at a higher voltage, sparks pass across the intervening space between the conductors: these are often intermittent ('pilot') sparks, and, on raising the voltage further, a continuous stream of sparks usually occurs and may develop into an arc, with complete breakdown of the oil.

The electric strength is generally tested with electrodes consisting of two metal spheres of about 13 mm diameter separated by a gap of 4 mm. For clean, dry oil the breakdown voltage should be in the region of 100 kV r.m.s. or more, but careful treatment, storage and handling are needed to maintain this level. For the oil to comply with BS 148:1998, it should withstand for 1 min without breakdown 40 kV r.m.s. applied between the spherical electrodes under conditions laid down in the Specification.

The electric strength of insulating oil is strongly affected by impurities, especially water and particles of fibrous material. The latter are attracted to the testing gap by the electric field and readily align themselves across the shortest space. The presence of moisture in oils is shown by electric strength tests when particles of such solid impurities—particularly organic fibres—are present, the breakdown voltage being reduced considerably by even small quantities.

Typical values of electric strength in different conditions are given in *Table 7.8*. Water and other impurities can be removed from oil by means of filter presses, centrifuges or (where high voltages are concerned) the application of vacuum. In addition to removing moisture, the vacuum will remove dissolved gases, but it is necessary to heat the oil and to spread it out over a very large surface area to facilitate the process. Once oil has been treated in this way, it must be stored out of contact with air and for preference at a temperature higher than the ambient.

7.4.2.2 Viscosity

This property, particularly at low temperature, is of great importance in oils used primarily for cooling in transformers and rheostats, it being necessary for the viscosity to be sufficiently low to ensure the necessary convection at the operating temperatures. This property is usually determined by methods such as those described in BS 188:1997, and the viscosity is expressed in centistokes (cSt). Oil to BS 148:1998 has a maximum viscosity (kinematic) of 37 cSt at 21.1°C (70°F). This is approximately equivalent to 151 s at 21.1°C and 200 s at 15.5°C (60°F) obtained with a Redwood No. 1 viscometer. *Figure 7.3* gives typical viscosity–temperature characteristics.

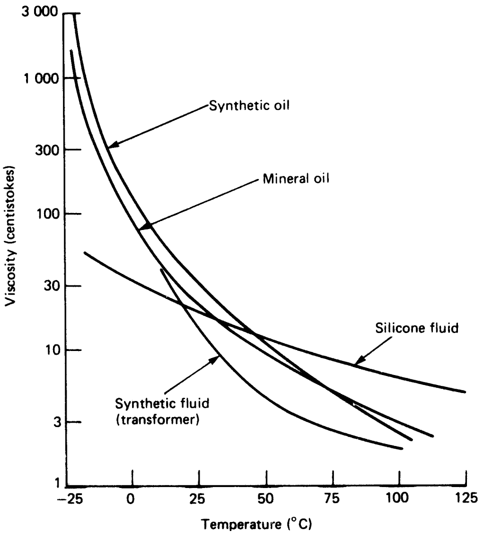


Figure 7.3 Viscosity–temperature characteristics of insulating oils

7.4.2.3 Flash point

For standard oil this is not less than 146°C, and may be as high as 240°C for rheostats. These values refer to a *closed* flash-point tester.

7.4.2.4 Thermal properties

The specific heat is about 1900 J/(kg K) at 15°C and 2200 at 80°C. The thermal conductivity is of the order of 0.15 W/(m K).

7.4.2.5 Chemical stability

Insulating oils should be stable and not liable to deteriorate other materials or cause corrosion. The *acidity* is therefore closely controlled and oils are tested to ensure that they do not cause discoloration of copper. The worst feature of oils in this connection is the formation of sludge. This is mainly due to the oxidation of unsaturated hydrocarbons, particularly at high temperatures, and is accelerated by exposure of the oil to air and light, and (due to catalytic action) to copper. BS 148:1998 includes tests for acidity, discoloration of copper, tendency to sludge formation and development of acidity.

Useful guidance on means of maintaining insulating oils in service is given in the British Standard BS 5730:1979, for the *Code of Practice for Maintenance of Insulating Oil with Special Reference to Transformers and Switchgear*. This refers to oil supplied to BS 148:1998 and describes the nature of deterioration or contamination likely to occur in storage, or in the course of handling or in service. It also gives recommendations for routine methods of sampling and testing to enable the suitability of oil for further service to be determined.

BS EN 50195:1997 and BS EN 50225:1995 give guidance on the safe use of oil-filled equipment containing Askarel and PCB contaminations respectively.

Table 7.9 Properties of typical oils and fluids

Property	*Unit*	*M**	*S†⇐*
Density at 15°C	kg/m^3	880	970
Viscosity at 21°C	cSt	35	21
Viscosity at 60°C	cSt	6	11
Boiling range	°C	170–200	>200
Evaporation loss at 110°C	%	0.7	0
Flash point (closed)	°C	149	271
Pour point (max.)	°C	−31	−50
Sludge value	%	0.8	0
Acidity	gKOH/kg	0.01	—
Breakdown voltage, r.m.s.‡⇐	kV	45–70	40–60
Relative permittivity at 20°C, 50 Hz	—	2.1	2.7
Loss tangent at 20°C, 50 Hz	—	0.0002	0.0002
Coefficient of cubic expansion per °C	—	0.0008	0.001

*M, mineral oil to BS 148:1998.
† S, liquid methyl silicone.
‡In standard test cell.

The properties of a typical mineral oil, complying with BS 148:1998, are shown in *Table 7.9*.

7.4.3 Inhibited transformer oil

Oils operating at comparatively high temperatures, in the presence of oxygen and various catalytic materials, develop sludges and high acidity. These effects can be alleviated by adding various inhibiting substances to the oil, the most widely known being di-tertiary-butylparacresol used in quantities generally less than about 0.5% of the oil by weight. Materials of this type delay the point at which sludge and acid formation begin; but once the inhibitor has been used up, deterioration will proceed at the same rate as if no inhibitor had been used. For large power transformers, it has not been found necessary to use these inhibiting substances because of improvements in the construction which have reduced access of oxygen by conservators, hermetic sealing or the use of a nitrogen blanket above the oil surface. Another improvement has been the covering of copper surfaces so reducing the catalytic effect. For transformers operating under more adverse conditions of temperature such as distribution or pole-mounted units, a better case can be made for using inhibited oils.

7.4.4 Synthetic insulating liquids

Synthetic hydrocarbons are fairly widely used for power-cable insulation and as capacitor dielectrics. These are more expensive than petroleum oils but they generally have better electrical properties because of their lower contamination levels and they can have better gas-absorbing properties. A commonly used synthetic oil is poly-iso-butylene, commonly known as polybutene. Different polymer chain lengths can be produced giving a wide range of viscosities from low viscosity liquids to sticky semi-solids. The high molecular weight tacky and rubbery materials can be used mixed with oil, resin, bitumens, polyethylene and inorganic fillers to produce non-draining and potting compounds. Another synthetic oil that is used for cable insulation is dodecylbenzene, an aromatic compound. The physical properties are similar to mineral oils, but the viscosity–temperature characteristics show much higher low-temperature viscosity than comparable polybutenes. However, the gas-absorbing characteristics are good. Electrical properties are generally similar to mineral oils but the permittivity is somewhat higher for dodecylbenzene.

Polychlorinated biphenyls (also called askarels) have been used as high permittivity (3–6) fire-resistant insulating liquids since the 1930s but their effect as an ecological poison has limited their use to sealed equipment in recent years and all use of these liquids is being discouraged.

Silicone fluids (poly-dimethylsiloxanes) have been used as alternative fire-resistant insulating liquids, but their fire resistance is inferior to the askarels. They are generally gas evolving and their arc products can cause problems. However, they are very stable and have good electrical properties and are used in transformers and capacitors.

Several liquids have been developed as alternative high permittivity insulating liquids to replace the askarels for capacitor dielectrics. One possible group of materials is the organic esters. They have good viscosity temperature characteristics and are less flammable than mineral oil and can be either gas producing or gas absorbing, depending on their composition. When carefully purified, their electric strengths are about 20 kV/mm and dissipation factors average 0.001 at 20°C. Diesters have relatively high permittivities (4.3 for di-2 ethylhexylphthalate). Other liquids that may be suitable for electrical insulation are phosphate esters, halogenated hydrocarbons, fuoroesters and silicate esters. Castor oil is a good insulation material for d.c. stress with a permittivity of 4.7, but it has a high dissipation factor of 0.002 that makes it unsuitable for most a.c. applications.

BS EN 60867:1994 and BS 61099:1992 give specifications for unused liquids based on synthetic aromatic hydrocarbons and for unused synthetic organic esters respectively.

7.5 Semi-fluid and fusible materials

A few semi-fluid or semi-plastic compounds, and various fusible materials which are solids at normal temperature and melt to liquids of low viscosity or soften considerably with heat, are used principally in the following ways:

(1) for filling small cavities and large spaces, e.g. in metal-clad switchgear, transformers, cable-boxes and capacitors;
(2) for impregnating absorbent materials and windings;

(3) as the bond in laminated materials;
(4) as the basic material in moulding compounds; and
(5) for external coverings of parts and apparatus (i.e. envelopment and encapsulation).

The materials most commonly used for these purposes are bitumen, natural waxes, shellac, synthetic waxes and synthetic resins; with the exception of many of the latter, and shellac, these are all *thermoplastic* materials, i.e. they soften and melt on heating and solidify again on cooling without any substantial chemical change, and they can be re-softened or re-melted. In the case of some of the synthetic resins, especially those of the phenolic type, gradual hardening takes place as they are heated, and the melting point rises, so that, after being melted, solidification takes place on further heating, the material then becoming infusible: i.e. the process of melting and solidification on cooling is not repeatable; they are therefore known as *thermosetting* materials. Shellac also has thermosetting properties, but it requires longer heating to effect marked rise of melting point than in the case of many synthetic resins.

The properties of chief importance in such materials are: mechanical strength; electric strength; freedom from impurities; softening and melting temperatures; viscosity at pouring or impregnating temperatures; coefficient of expansion; and chemical effects on other materials.

Several materials which are *semi-fluid* at normal temperatures are used for filling and sealing purposes. For example: good grades of petroleum jelly of the Vaseline type are preferred to oil for filling apparatus and components where a liquid is undesirable, or where molten compounds cannot readily be poured or may affect other materials which are present (e.g. rubber and thermoplastic materials).

Various '*semi-plastic*' compounds or cements, of a putty-like consistency, are also used for plugging and filling purposes, where semi-fluid and molten compounds cannot readily be applied. Some of these are almost permanently plastic and are therefore preferred where, for example, a certain amount of flexibility is required (e.g. where leads of coils may be moved slightly in assembly or service). Others may harden gradually in course of time (as in the case of ordinary putty), or they may harden quickly by chemical action (e.g. litharge and glycerine cement) or by heating—the latter usually being necessary with synthetic-resin compounds.

7.5.1 Bitumens

Highly refined bitumens, which are usually steam distilled, and of numerous grades, varying from semi-liquids to hard bitumens of melting point over 120°C, are used extensively for filling cable boxes, transformers and switchgear. These have high electric strength and are very inert and stable. As the coefficient of expansion is high, care has to be taken in filling large spaces to prevent voids and cracks on cooling. Some of the bitumens, especially those of high melting point, are rather brittle; all are soluble in oil, but they have excellent resistance to moisture. BS 1858:1973 deals with bitumen-base filling compounds for electrical purposes. Properties of typical bituminous compounds are given in *Table 7.10*. Some bitumens are used as ingredients in varnishes and paints, rendering these very resistant to moisture and chemical attack. A few impregnating compounds contain bitumens, especially those used for treating high-voltage machine bars and coils.

7.5.2 Mineral waxes and blends

Various mineral waxes such as paraffin, ceresine, montan and ozokerite—including microcrystalline waxes—also blends and gels of these, having melting points in the range 35–130°C, are used for impregnating capacitors, radio coils and transformers, also for other purposes such as cable manufacture. Properties of mineral waxes are given in *Table 7.11*.

7.5.3 Synthetic waxes

A few synthetic waxes—principally chlorinated naphthalene—with melting points up to 130°C have certain advantages over natural waxes, particularly higher permittivity which enables smaller paper-insulated capacitors to be made. Properties of a typical synthetic wax of this type are given in *Table 7.11*.

7.5.4 Natural resins or gums

These materials, which may be classified broadly as shellac, rosin (colophony), copals and gum arabic, are used principally as ingredients in varnishes or liquid adhesives. In some cases they are used direct, e.g. as powders, for a bonding medium between layers of mica which are hot pressed, but they are usually dissolved in spirit solvents, e.g. methylated spirits (or water in the case of gum arabic), and applied as a solution to mica, paper, etc., for subsequent laminating and hot rolling, pressing or moulding (see *Table 7.11*).

Table 7.10 Properties of fusible bituminous compounds

Property		*Class (BS 1858:1973)*				
		I	*II*	*III*	*IV*	*V*
Density	kg/m^3	960	1030	1040	1050	1027
Softening point (R & B)	°C	—	55–60	85	118	143
Pouring temperature	°C	—	175	171	193	204
Flash point	°C	>200	260	308	260	312
Viscosity at 100°C	Rdwd.-s	750	—	—	—	—
Solubility in CS$_2$	%	99.5	99.8	99.5	>99	99.8
Acidity	g KOH/kg	1–2	2	1.5–4	4	4
Cubic expansion	per °C	0.000 65 for all classes				
Electric strength at 60°C	kV r.m.s.	15–25	30–40	25	25–30	25–40

Table 7.11 Properties of fusible waxes, resins and gums

Property		*Natural shellac*	*Natural copal gum (kauri)*	*Non-bituminous filling compound*	*Paraffin wax*	*Hydrocarbon wax*	*Synthetic chloronaphthalene wax*
Density	kg/m^3	1000–1100	1040	1100	900	800–1000	1550
Softening point	°C	50–70	60–90	70	45–50	—	90
Melting point	°C	80–120	120–180	80	50–60	40–130	123
Flash point	°C	—	—	230	200	275	—
Mineral ash	%	0.5–1	3	5	0	0	—
Acid value	g KOH/kg	60–65	70–85	—	0	<0.1	—
Saponification value	—	200–225	80	—	0	<0.5	—
Iodine value	%	9	90	—	0	—	—
Relative permittivity (20°C)	—	2.3–3.8	—	—	2.2	2–2.5	5
Resistivity (20°C)	Ω-m	10^{14}	—	—	10^{13}–10^{17}	10^{14}–10^{16}	$>5\times 10^{11}$
Electric strength* (20°C)	kV/mm	16–23	14–18	>30†⇐	12	>50‡⇐	6
Resistance to mineral oils		Fair	Good	Good	Poor**	Fair	—

*R.m.s. for 3 mm thickness, except for † 1.2 mm gap and ‡ 4 mm gap between 13 mm electrodes.
**Dissolves.

7.5.5 Miscellaneous fusible compounds

Numerous compounds of bituminous and other types are used for all manner of cavity-filling purposes, some (mainly rosin) are oil resisting and are employed where bituminous compounds cannot be used owing to the presence of oil, e.g. for bushings of oil circuit-breakers and oil-immersed transformers. Others are used for sealing over the tops of primary batteries and accumulators. Compounds containing beeswax are useful impregnants for small coils not exposed to heat, e.g. on telephone apparatus; and sulphur, 'sealing wax' and 'Chatterton's compound' are examples of materials finding uses for miscellaneous applications where, say, the heads of screws in insulating panels and mouldings require to be sealed over.

7.5.6 Treatments using fusible materials

Treatments with bitumen, waxes, etc., usually consist of thorough vacuum drying of the coils, capacitors or other parts to be treated, followed by complete immersion in the compound while in a molten condition and at a temperature such that the viscosity is low enough to facilitate penetration; the molten compound generally being admitted to the impregnating vessel under vacuum. Pressure (up to 10 atm) is often applied during the immersion period to assist penetration. Such treatments enable spaces in windings to be filled thoroughly, and absorbent materials such as papers and fabrics are often well saturated with the impregnants, especially in the case of waxes. These treatments provide good resistance to moisture absorption and improve transference of heat from the interior of coils, also eliminating discharges in high-voltage windings and capacitors by the filling of air spaces.

7.5.7 Synthetic resins

An increasing number of the well-known synthetic resins, which form the basis of the principal 'plastics', are of great use to electrical engineers on account of their fusibility or softening characteristics at elevated temperatures, which enables them to be converted to desired shapes. The synthetic resins can be divided into two groups: thermoplastic and thermosetting (see *Table 7.12*).

Table 7.12 Thermoplastic and thermosetting synthetic resins

Thermoplastic	*Thermosetting*
Polyethylene	Phenol formaldehyde
Polystyrene	Phenol furfural
Polyvinyl acetate	Urea formaldehyde
Polyvinyl chloride	Melamine formaldehyde
Acrylates	Silicones
Polyesters, alkyds, etc. (non-hardening)	Polyesters, alkyds, etc. (thermohardening)
Polyamides	Epoxy (epoxide)
Polyacetal	Polyurethanes
Polypropylene	Polyimide
Polycarbonate	Polyimide-amide
Polyphenylene oxide	Polyaralkyl ether/phenols
4-Methylpentene-1	
Acrylonitrile–butadiene–styrene	

Thermoplastic synthetic resins In the case of most of the thermoplastic materials of this type, heating to temperatures within a certain range causes considerable softening and sometimes melting of the material to a viscous liquid. This enables them to be cast, formed, moulded or extruded into various required shapes by virtue of re-solidification on cooling again to normal temperatures. Some of the resins (e.g. acrylates and alkyd resins) have good adhesive properties and can therefore be used for bonding purposes either in the form of a solution or, more usually, by the application of heat. Layers of sheet materials, such as paper and fabric, can thus be bonded together into boards, simple mouldings, tubes, etc. The resins are often used alone, but more usually mixed with materials such as fillers and plasticisers, and in both varieties these synthetic materials are usually capable of being formed to all manner of shapes by the usual moulding processes (see *Table 7.13*).

Thermosetting synthetic resins These enable useful compositions to be made, and withstand temperatures in excess of

Table 7.13 Properties of thermoplastic and casting resins

Thermoplastic resins		*Polyethylene*	*Poly-styrene*	*Poly-methyl-meth-acrylate*	*Poly-amide (Nylon 6.6)*	*Polyacetal*	*Polypro-pylene*	*Poly-carbonate*	*Poly-phenylene-oxide*	*4-Methyl-pentene-1*	*Acrylo-nitrile–butadiene–styrene*
Density	kg/m^3	920	1050	1190	1140	1420	920	1200	1200	830	900–1000
Softening temperature	°C	95	70–95	80–85	180	158	145	135	100–150	178	85
Melting point	°C	110	—	—	260	163	164	230	—	240	—
Linear expansion $\times 10^{6}$	per °C	220	75	80	100	95	110	70	35	120	60–120
Water absorption	%	0	0.05	0.4	<8	0.2	0.03	0.35	0.07	0.01	0.1–1
Elastic modulus	GN/m^2	1–2	3	3.3	3	2.6	1–1.4	2.4	2.5–8.4	1.4	0.7–2.8
Tensile strength	MN/m^2	12	41	59	45	69	31–38	93	69–120	27	17–62
Flexural strength	MN/m^2	Low	77	100	87	96	—	—	90–138	—	27–84
Impact strength	kgf-m	5–30	0.05	0.05	0.15–0.3	0.32	0.08–0.8	—	0.2–0.3	0.05–0.1	0.4–1.7
Resistivity (20°C)	Ω-m	3×10^{15}	10^{15}–10^{16}	$>10^{13}$	10^{13}	5×10^{12}	$>10^{14}$	2×10^{14}	10^{15}	$>10^{14}$	10^{11}–10^{14}
Relative permittivity (20°C)	—	2.3	2.5–2.7	2.8	3.5–6	3.7	2.23	3.1	2.65	2.12	2.7–4
Loss tangent (20°C)	50 Hz	0.000 1	0.000 2	0.06	0.015	0.004	0.000 2	0.000 9	0.000 5	0.000 1	0.004–0.07
	1 kHz	0.000 1	0.000 2	0.03	0.020	—	0.000 2	—	—	0.0000 5	—
	1 MHz	0.000 1	0.000 2	0.02	0.02–0.06	0.004	0.000 5	0.01	0.001	0.000 2	0.007–0.02
Electric strength*	kV/mm	15	20	10	15–19	20	30–32	16	20	28	12–15

Casting resins		*Polyester*		*Epoxy (bisphenol type)*		*Epoxy (cycloaliphatic type)*		
							Mineral filled	
		Unfilled	*Mineral filled*	*Unfilled*	*Mineral filled*	*Unfilled*	*Anti-track*	*Mechanical strength*
Density	kg/m^3	1100–1400	1600–1800	1100–1200	1600–2000	1100–1200	1700–1800	1700–1800
Linear expansion $\times 10^{6}$	per °C	75–120	60–70	45–65	20–40	90–95	38–43	38–43
Water absorption	%	0.15–0.6	0.1–0.5	0.08–0.15	0.04–0.1	0.04–0.05	0.02–0.04	0.02–0.04
Elastic modulus	GN/m^2	2–4	2.5–3	2–2.5	2.5–3	3.4–3.6	17–18	18–20
Tensile strength	MN/m^2	42–70	20–35	60–80	50–75	40–50	30–40	50–60
Compressive strength	MN/m^2	90–250	120–200	95–140	100–270	120	130–150	180–200
Flexural strength	MN/m^2	56–120	50–100	90–140	56–100	80–100	60–70	80–100
Resistivity (20°C)	Ω-m	10^{11}	10^{10}–10^{11}	10^{10}–10^{15}	10^{11}–10^{14}	5×10^{14}	6×10^{12}	5×10^{13}
Relative permittivity (20°C)	1 kHz	3.2–4.3	3.8–4.5	3.5–4.5	3.2–4	3.5–3.6	4.6–4.7	4.3–4.5
	1 MHz	2.8–4.2	3.6–4.1	33.4	3–3.8	—	—	—
Loss tangent (20°C)	1 kHz	0.006–0.04	0.008–0.05	0.002–0.02	0.008–0.03	0.01	0.02	0.035–0.039
	1 MHz	0.015–0.03	0.015–0.03	0.03–0.05	0.02–0.04	—	—	—
Electric strength*	kV/mm	20	15–20	16–22	16–22	20–21	15–17	19–21

*R.m.s. for 3 mm thickness.

100°C. The most widely used are the phenol formaldehyde type. The materials pass through three stages of physical condition:

1. in which the resins are fusible at temperatures such as 80°C, and are soluble in suitable solvents;
2. results from heating the stage (1) resin until it becomes relatively infusible and insoluble; and
3. the infusible and insoluble state reached by continued heating after stage (2); no further change occurs and the materials are 'fully cured' or 'completely polymerised'.

These physical stages make thermosetting resins suitable for three main uses.

1. In spirit solutions; as ingredients in varnishes for impregnating purposes and the production of surface finishes; as enveloping, potting or encapsulating materials, and as ingredients in filling compounds.
2. As adhesives for bonding layers of wood, paper, fabrics, etc., together to form laminated sheets, wrappings, and other simple shapes.
3. As the basic material in moulding compositions for use in making articles by compression or injection moulding, extrusion or casting.

Properties of typical thermosetting resins of the phenol formaldehyde type, unfilled, are given in *Table 7.14*.

Many of the thermosetting resins, e.g. phenol formaldehyde and melamine formaldehyde, require heavy pressure during the heating and hardening processes (2) and (3) above. Several resins requiring little or no pressure (polyesters, epoxies and polyurethanes) have been developed as 'low-pressure', 'contact' or 'casting' resins, or as 'solventless varnishes'. These resins are initially in a low-viscosity liquid state, to which a 'hardener' or catalyst (e.g. a peroxide) is added. In some cases polymerisation sets in at normal room temperatures, or at temperatures of only 80–100°C, the hardening process taking place more rapidly as the temperature is increased. Thus the resins can be readily cast to required shapes in 'moulds', and can also be used for impregnating and coating windings as they readily fill interstices and do not leave voids on hardening owing to the fact that no volatile constituents evaporate—hence their use as 'solventless varnishes'. Mixed with suitable fillers (e.g. glass fibres, asbestos or other minerals) or applied to fabrics, papers and other sheet materials (usually of glass fibres), they are used extensively for producing castings, mouldings and laminates of varying degrees of mechanical and electrical strength, sometimes in very large pieces which could not readily be made by normal moulding methods; they are usually referred to as 'reinforced plastics'.

Table 7.14 Properties of unfilled phenol formaldehyde resins

Property		*Varnish type*	*Casting type*
Density	kg/m^3	1260	1300
Softening point	°C	60–80*	—
Plastic yield temperature	°C	120	85
Hardening time at			
105°C	min	45*	—
150°C	min	6*	—
Linear expansion $\times 10^6$	per °C	20	28
Water absorption	%	0.1	0.07
Elastic modulus	GN/m^2	5–7	3
Tensile strength	MN/m^2	35	28–56
Compressive strength	MN/m^2	170	100
Resistivity	Ω-m	10^{10}	10^9
Relative permittivity	—	4–7	7–11
Loss tangent			
50 Hz	—	0.05	0.10
1 kHz	—	0.03	0.20
1 MHz	—	0.02	0.25
Electric strength (90°C, 3 mm thick)	kV r.m.s./mm	8–20	12

* In stage (1); other properties are for stage (3) (see text).

A brief description of the principal synthetic resins in electrical use is given below. Some are suitable for moulding with or without fillers, some for the preparation of laminated materials. Rod, sheet and tube forms are available in certain cases.

BS 1133-16:1997 deals with packaging adhesives and gives information on their characteristics and end use including advice on storage and precautions in use.

7.5.8 Thermoplastic synthetic resins

Polyethylene is waxy, translucent, tough and flexible, with a sharp melting point at about 110°C, and is used for high-voltage and high-frequency applications. It is readily injection moulded, extruded as wire coverings, and in sheet, rod and film form.

Polytetrafluoroethylene a white powder that can be moulded or extruded. It is highly resistant to moisture and chemicals, and withstands temperatures up to 250°C. It is used in high-frequency application.

Polystyrene softens at 70°C. It can be compression or injection moulded and may be used with a mineral filler to improve heat resistance.

Polyvinyl acetate and copolymers Polyvinyl acetates are obtained from acetylene and acetic acid: they are used as adhesives and enamels.

Polyvinyl chlorides These are obtained from the combination of acetylene and hydrochloric acid as a white powder used with stabilisers, plasticisers, etc., to produce various rubber-like materials that can be extruded as tubes for wire protection. Sheet and moulded polyvinyl chloride have a loss tangent too high for high-frequency use. Copolymers of the acetate and chloride forms are tough, rigid and water resistant and can be injection moulded.

Acrylates The most important product is polymethylmethacrylate, a rigid glass-clear material with good optical, electrical and mechanical qualities. It can be obtained in sheet, rod and tube form, and as a moulding powder. Its low softening point (60°C) limits its application to moderate temperatures.

Polyesters Some non-hardening *alkyds* have limited use as adhesives.

Polyethylene terephthalate This has a sharp melting point at about 260°C and is formed into filaments for textile manufacture. Also extruded to form films. It has a high resistance to temperature and ageing and to water absorption. The textiles are suitable for class E insulation and, when suitably varnished, may withstand temperatures greater than class E.

Cellulose acetate and triacetate These are also esters. Produced as lacquers, textiles, sheet, rod, film and moulding powder, they are suitable for machine windings. The acetate softens at 60–80°C, the triacetate at 300°C. The materials are available as fibrous cotton or paper tapes.

Polyamides Super-polyamides are known as 'nylon': they produce monofilaments and yarns, with very good mechanical properties. The electrical properties are not outstanding, but nylon gives tough and flexible synthetic 'enamel' covering for wires. Films and mouldings can also be produced.

Polyacetal The material has good dimensional stability and is tough and rigid. It can be injection moulded and extruded, and is replacing metal parts in relays.

Polypropylene This material has a low density, dielectric loss and permittivity. Special stabilisers may be necessary when the material is extruded on to copper conductors.

4-Methylpentene-1 is similar to polyethylene and polypropylene and with similar resistance to chemicals and solvents. It is the lightest known thermoplastic. The high melting point (above 240°C) cannot be fully exploited because of softening and oxidation. Its permittivity and loss are low and remain fairly constant over a wide frequency and temperature range.

Polycarbonate approaches thermosetting materials in retention of stability up to 130°C. It is self-extinguishing and useful for structural parts, housings and containers for hand tools and domestic appliances.

Polyphenylene oxide is stiff and resistant to comparatively high temperatures. Its permittivity and loss tangent are fairly constant at frequencies up to 1 MHz.

Acrylonitrile butadiene styrene has good dimensional stability and mechanical strength from − 40 to 100°C.

7.5.9 Thermosetting synthetic resins

Phenol formaldehyde These versatile resins are available in varnishes, adhesives, finishes, filling and impregnating compounds, laminated materials (boards, tubes, wrappings and sheets), moulding powders, and cast-resin products. The principal resins (Bakelite) are made by reacting phenolic material with formaldehyde. The final polymerising ('curing') time, which vitally affects the use, varies at 150°C from a few seconds to an hour or more. The resins are normally solids of softening point between 60 and 100°C. They are readily soluble in methylated spirit for coating papers, fabrics, etc., in the manufacture of laminates. Varieties are suitable for pouring molten into moulds followed by polymerisation. The most extensive use is as moulding powders with fillers (wood flour, powdered mica, fibres and colourings) to give mechanical strength and suitable electrical properties.

Phenol furfural is produced by the reaction of phenol with furfural, an aldehyde obtained by acid treatment of bran and fibrous farm waste. It is suitable for injection moulding.

Urea formaldehyde The main use is as the binder of cellulose, wood flour or mineral powder in mouldings, made by compression at 115–160°C. Mouldings can be delicately coloured.

Melamine formaldehyde Its properties are superior to those of the urea formaldehydes. It is suitable for mouldings for ignition equipment and has good resistance to tracking.

Silicones These are organic compounds of silicon. By variation of the basic silicon–oxygen structure and of the attached organic groups, many different products can be made, including fluids, resins, elastomers and greases. Their main properties are water repellency (their hydrophobicity makes them a better choice for outdoor high voltage insulation), stability to heat, cold and oxidation, and good electrical properties maintained up to 200°C and higher. Some silicones can work continuously at 200°C and intermittently to 300°C: they can be applied to insulation in classes F, H and C. Silicone resins are used for bonding mica, asbestos and glass-fibre textiles and for producing compounds, varnishes, micanite, wire coverings, etc.

A number of silicone compounds can be used for filling and sealing where heat and moisture resistance are required. One such, of the consistency of petroleum jelly, is of use as a waterproof seal in high-voltage ignition systems: it protects cable insulation from moisture, oxidation and electrical discharges.

Polyesters Alkyd resins are more rubbery than phenolic resins, have good adhesion and do not readily track; they are therefore of use for finishes and for varnishing glass fibre and similar material to produce heat-resisting varnished cloths. *Unsaturated polyesters* are useful for casting and potting, as solventless varnishes, and in the manufacture of laminates with glass fabric or mineral fillers.

Epoxies The epoxy resins have become important as casting, potting, laminating, adhesive and solventless varnish agents. They have good electrical properties and resistance to heat, moisture and tracking, and adhere well to metal parts. They have been applied for high-voltage insulation in switchgear and for casting, in which case they are often mixed with mineral fillers. Earlier epoxy resins showed damage when subjected to severe weather and to high electric stress on creepage surfaces. New types have been based on cycloaliphatic resins which, because of the different molecular structure, produce less carbon during the passage of surface discharges and leakage currents under polluted conditions. Further improvements have been made in this application by using specially selected and treated mineral fillers which also reduce the effects of weathering and surface tracking. These products have many uses on high voltage outdoor equipment.

Polyurethanes, isocyanates These are used mainly for coating fabrics (such as glass- and polyethylene-terephthalate fibre) to produce heat-resisting flexible sheet insulation and for coating wires.

Polyimides These have been specially developed for use at high temperature as mouldings, films, wire enamels and laminate bondings. The materials can be used continuously at temperatures in the 200–240°C region; and for very short periods, they can withstand temperatures up to 500°C without apparent damage. Their mechanical and electrical properties are good and their resistance to most chemicals, solvents and nuclear radiation is excellent. *Polyamide imide* resins are similar to the polyimides and are available in the same forms. The performance at high temperatures is marginally lower but the resins are simpler to use in the manufacture of laminates, and have longer shelf-life.

Polyaralkyl ether/phenols These have a high-temperature performance not quite as good as that of the polyimides, but are cheaper. The resins can be used as bonds for glass and asbestos laminates and mineral filled moulding powders are available. A high proportion of room temperature strength is retained at temperatures up to 250–300°C and long-term operation at temperatures of 220–240°C is possible. Their resistance to most chemicals and solvents is excellent.

7.5.10 Encapsulation

When electronic or electrical components and circuits must resist the effects of climate, industrial atmospheres, shock or vibration they may be encapsulated. They are then generally known as 'potted circuits'.

Certain thermosetting synthetic resins are of the greatest use as they can be easily poured from low-viscosity liquids and made to set without the use of pressure and, in some instances, with very little heat. A suitable material must: (1) be a good insulator over a wide range of temperatures (volume resistivity say $10^6\,\Omega$-m); (2) polymerise or set without spitting off water or other products; (3) have low viscosity at pouring temperature, low vapour pressure, and freedom from deleterious side-effects on personnel who are using it; (4) show small shrinkage, especially when changing from the liquid to the solid state; and (5) must adhere to all materials commonly found in electrical equipments, e.g. to brass, solder, steel and insulating boards. Only the epoxy resins possess all the essential requirements for successful encapsulation and even these need an inorganic filler to obtain the best heat resistance, low shrinkage, good electrical properties and high thermal conductivity.

The epoxide resins of use in potted circuits are derived from a condensation reaction between epichlorhydrin and bisphenol A. They are cross-linked with aliphatic or aromatic amines, acid anhydrides and a few other chemical compounds to give thermally, electrically and mechanically stable resins. The addition of inorganic fillers improves them and reduces their shrinkage, tendency to crack at low temperatures and cost. The mixture of resin and cross-linking agent (known as a hardener) is called a system. An accelerator may be added, as well as various diluents, both reactive and non-reactive. Accelerators and promotors alter the speed of reaction and the pot life.

Typical potting formulations, in parts by weight, are:

(A) resin 100, hardener 10, mica flour filler, 15; and
(B) resin 100, hardener 82, quartz flour filler 375, accelerator 1.

Formulation A is satisfactory for small units of about 0.1 kg of mixture. It sets at room temperature, but is post-cured at an elevated temperature dependent on the heat resistance of the included components; 18 h at 65°C is usually satisfactory. Such a unit is suitable for small electronic packages containing small components, including semiconductors. Larger masses, depending on their geometry, may exhibit a strong exotherm (i.e. generate heat) which may damage the included components.

Formulation B is used hot (usually at about 65°C) and is very fluid at this temperature besides possessing a long pot life. This makes it suitable for the potting of transformers. Vacuum impregnation is essential to eliminate voids, which would result in ionisation and corona discharges. The large amount of filler greatly improves the thermal and mechanical properties and allows a larger casting to be produced without a high exotherm. Again, a post-curing cycle is essential to bring out the best properties; in this case about 18 h at 120°C is satisfactory and will produce a material with high volume resistivity and heat resistance. The pot life of this mixture is about two hours at 65°C, which may be compared with formulation A whose pot life is less than half an hour at room temperature.

In use the resin and dried filler of A are mixed together and stored under vacuum or in a desiccator until used. The hardener is added and thoroughly mixed just before pouring into the mould or other article to be potted. It stands at room temperature to set or 'gel' and is subsequently post-cured; this latter process may be carried out after removal from the mould if required.

For formulation B the resin, hardener and dried filler are all mixed and stored as before, and the accelerator is added to the mixture just before use and heated to 65°C. It is poured into the mould or other article. For transformers this is usually done under a vacuum of about 1 mmHg to 1 cmHg. Although a large amount of inorganic filler is used, this is filtered to some extent by the windings of the transformer so that the insulation between turns is mostly of unfilled resin, giving high breakdown strength and freedom from corona discharges. Good adhesion at the terminals is ensured by the nature of the resin system, but this also means that it sticks to the mould unless a release agent is used. Various preparations are used, including mixtures of high polymers, silicones, greases and waxes. The agent must be confined to the mould and not allowed to contaminate the components inside the casting.

The requirements of low shrinkage and low exotherm have been stated. The effect of the former is to damage components by compressional forces and is particularly severe on some nickel–iron alloys used in making inductors; thin-film resistors and capacitors are also vulnerable to this form of damage. Isolating the components from the resin by means of low modulus materials can appreciably reduce this defect. The heat of reaction (exotherm) is also damaging to organic materials and to semiconductors; this often results in a loss of volatile matter causing shrinkage and the effects already noted above.

All adhesives contain polar bonds in their molecular structure. These give rise to changes in dielectric properties as the frequency and temperature varies, causing the electrical behaviour of the components inside the casting to change as the frequency and temperature changes.

Another disadvantage of potted circuits is that they are irreparable: they must be designed as 'throw away' subunits and made at an economical price. For this reason they are rarely used in domestic applications such as radio or television, but are of particular use in military electronics, for machine control and similar purposes where the utmost reliability is essential and first costs are relatively unimportant.

7.6 Varnishes, enamels, paints and lacquers

Numerous liquid materials, which form solid films, are used extensively in the manufacture of insulating materials and for protecting windings, etc.

7.6.1 Air-drying varnishes, paints, etc.

One class of air-drying varnishes and lacquers consists of plain solutions of shellac, gums, cellulose derivatives or resins which dry (e.g. in 5–30 min) and deposit films by evaporation of the solvent. Other air-drying varnishes and paints form films which harden by evaporation of solvent accompanied by oxidation, polymerisation, or other chemical changes which harden and toughen the film. These processes take several hours.

7.6.2 Baking varnishes and enamels

Where the toughest and most resistant coatings are required, baking varnishes and enamels are used, the evaporation of solvents and hardening of the material being effected by the application of heat. Typical varnishes of this class require baking at, say 90–110°C in a ventilated oven for 1–8 h. During the baking (or 'stoving') the hardening is usually caused by oxidation, but in some cases polymerisation takes place. The latter process does not require oxygen and, provided that the solvents are first removed, drying can take place within the interior of coils. In consequence these varnishes are preferred to the oxidising types which skin over and leave liquid varnish underneath. Such thermosetting impregnating varnishes are used extensively for treating coils, and the windings of small machines and transformers.

7.6.3 Solventless varnishes

Thermosetting synthetic resins are used for impregnating windings and, owing to the manner of hardening during which little or no volatile matter is evolved, spaces in the interior of windings can be filled completely with non-porous resin. One type consists mainly of oil-modified phenolic resins, similar to the oleo-synthetic resinous varnishes but without solvents; they are consequently termed *solventless varnishes*. When used for impregnating they are in a hot liquid condition, the material solidifying within the winding by baking after impregnation.

Specially formulated low-viscosity resins of the polyester or epoxy type are also used for impregnation. It is possible to use solventless resins in conveyorised plants where the parts are dipped in the resin, allowed to drain and then passed through a heated tunnel to cure the resin. In other cases, particularly for tightly wound apparatus, the parts are treated in the resin using a vacuum–pressure process to ensure that a high level of impregnation is attained.

The most recent development in treating industrial machine windings is the 'trickle' process. The wound part is heated and mounted at a slight angle so that it can be rotated slowly about its axis. A metered quantity of the resin is allowed to trickle on to the winding and under the action of gravity and the rotational forces, the resin penetrates to all parts of the winding and completes the impregnation. It is possible to completely fill the interstices of the winding without loss of resin by draining. Radiant heat may be applied to complete the cure while the parts are rotating, or heating currents may be circulated through the winding. Trickle impregnation can be performed automatically and the process can be completed in a few minutes without removing the wound parts from the production line.

7.6.4 Silicone varnishes

Varnishes based on silicone resins are in general use. Some include other resins, etc., such as alkyd resins. Silicone varnishes are used for impregnating and coating cloths, tapes, cords, sleevings, papers, etc., made from glass fibres and polyethylene-terephthalate fibre; for bonding mica, glass cloths and papers, e.g. for slot insulation; for coating and bonding glass; for 'enamelling' wires for windings; and for all manner of impregnating, bonding, coating and finishing purposes such as the treatment of windings for classes B, F and H.

7.6.5 Properties of varnishes, etc.

The properties of the solid films formed after drying and hardening varnishes, paints, etc., naturally depend mainly on the principal basic materials, e.g. gums, resins and oxidising oils. Other materials added are: 'driers' to accelerate drying; 'plasticisers' to improve the flexibility; and pigments to provide the required colour and improve the hardness and filling capabilities.

Treatments of windings and insulation parts, by varnishes, enamels, lacquers or paints, take the form of (a) application of external coatings—chiefly for providing protection against moisture and oils, or (b) impregnation of windings and absorbent materials, for rendering them less susceptible to moisture and improving their electrical and heat-resisting properties. Both air drying and baking materials are used for (a), but only baking varnishes and enamels are suitable for (b). In practically all cases, thorough drying prior to application of the varnish is essential as for compound treatment, but with varnishes, extra care is required to remove excess varnish by proper draining and to extract solvents thoroughly before hardening the material by baking. This is usually done in well-ventilated ovens at temperatures of 80–150°C and sometimes by the application of vacuum for a period to assist the removal of solvent. These processes are not required when solventless varnishes are used. In the case of silicone varnishes, the solvents must be removed and baking must then be carried out at temperatures in the range 150–260°C.

Insulating varnishes are the subjects of BS 5629:1979, BS 7831:1995 and BS EN 60464:1999. Typical properties are listed in *Table 7.15*.

7.7 Solid dielectrics

The many solids that can be used for insulating purposes are considered in groups according to their form.

7.7.1 Breakdown in solids

The breakdown strength measured for a solid is very dependent on the time of application of the voltage. Strengths of 100 kV/mm to 1.5 MV/mm can be obtained for short pulses under carefully controlled laboratory conditions, but in practical insulation systems breakdown strengths of only about 20 kV/mm are obtained. The high laboratory value is known as the *intrinsic electric strength* of the material, with breakdown due to electronic processes. In practical insulation, however, breakdown is due to: erosion, thermal effects, electromechanical effects, or treeing.

Discharge/Erosion breakdown originates at internal voids in the insulation. These are left at the fabrication stage and all solids contain them, although in a high-quality material they will be very small (*microvoids*). These voids will be filled with air or another gas with a permittivity and dielectric strength much lower than that of the solid. Electric field enhancement in the micro-void takes place due to the relative permittivity difference in the two media. Consequently, the gas filling the void will break down at a voltage lower

Table 7.15 Properties of typical varnishes

Air drying
A1: spirit shellac/methylated spirit
A2: oil and resin/petroleum spirit

Baking
B1: oil/petroleum spirit
B2: black bitumen/petroleum spirit
B3: synthetic resin and oil/toluol

Air drying and baking
AB; —; pigmented oil/white spirit

Silicone
SA: cured for 16 h at 250°C
SB: cured for 16 h at 150°C

Property		*A1*	*A2*	*B1*	*B2*	*B3*	*AB*
Density	kg/m^3	935	920	910	890	990	1400
Body content	%	38	62	54	44	63	73
Viscosity	c.g.s.	0.66	5.1	2.5	1.9	4.9	—
Drying time at 15°C	h	0.5	6–8	—	—	—	4–6
105°C	h	—	—	2	2.5	4	2
Electric strength at 90°C, r.m.s.	kV/mm	16	60	44	73	64	27
Electric strength damp at 20°C, r.m.s.	kV/mm	13	24	21	31	24	11

Property		*SA*		*SB*	
		Dry	*Wet**	*Dry*	*Wet**
Relative permittivity (25°C)	100 Hz	3.0	3.0	3.0	3.0
	1MHz	2.9	2.9	2.9	2.9
Loss tangent (25°C)	100 Hz	0.0077	0.0069	0.0084	0.0085
	1 MHz	0.0039	0.0053	0.0043	0.0047
Electric strength r.m.s.	kV/mm	60	40	60	40

* After 24 h immersion in distilled water.

than the normal breakdown strength of the material and this *partial discharge* will erode the solid at the ends of the spark on the surface of the void. These discharges will occur twice for every cycle of an alternating supply, so over a period thousands of millions of these discharges will occur in a cavity. The erosion effects of these discharges can eventually be sufficient to cause complete failure of the insulation. This is the primary reason why the insulation of a component may fail after a number of years in service.

Thermal breakdown is breakdown caused by internal heating in the insulation due to dielectric loss. Insulating materials are not perfect capacitors and they do have internal energy losses. The quality of the insulation is often measured in terms of how small these losses are, with the *loss factor* tan δ, where (90-δ) is the phase angle between the voltage and current for a sample of the material, commonly used as the parameter to assess the losses. This energy dissipated within the insulation will cause heating and electrical insulation is usually a poor thermal conductor, so a significant temperature rise in the insulation will be produced. In some circumstances the internal losses may increase as the temperature increases and a thermal runaway situation results. The temperature of the material will rise, losses will increase and the temperature will rise further until the material fails. Thermal breakdown is most likely to occur in power cables operated beyond their power rating, polymers operating near their softening points or in high-frequency applications.

Electromechanical breakdown results from the mechanical forces that an applied electric field produces. This force will reverse for every cycle of the applied voltage and can produce cracks or other damage in solid insulation.

There are two types of *treeing breakdown*. The first type originates at sharp points on electrodes which act as the root of fine branching channels that propagate through the insulation. Over a period of time these fine channels gradually extend towards the other electrode until complete failure of the insulation occurs. This is another mechanism which can produce failure after equipment has been in service for months or years. The second type of treeing is known as *water treeing* and only occurs when there is water in contact with the surface of the insulation. Microscopic damage to the insulation gradually spreads out from a high-field region until failure of the material occurs. A major difference between this type of treeing and the previous one is that the damage cannot be seen once the material has dried out.

7.7.2 Rigid boards and sheets

7.7.2.1 Panels and simple machined parts

Asbestos-cement boards (3–100 mm) are incombustible and arc-resistant, and are used as barriers and arc-chutes. *Glass-bonded mica,* in sheets 0.5 m × ~~0.~~4 m of thickness 3–30 mm, are especially good for high-frequency, high-voltage and high-temperature (400°C) application. *Micanite*, mica splittings bonded with shellac or synthetic resins, in thicknesses

Table 7.16 Properties of rigid mica-paper sheets and tubes*

Type of mica and resin		*P*	*M*		
		Shellac	*Epoxy*	*Epoxy*	*Silicone*
Sheets					
Density	kg/m^3	2270	2300	2200	2000
Bond content	%	5	13	20	9
Maximum operating temperature	°C	130	180	180	600–700
Water absorption	%	14	0.7	0.05	0.7
Tensile strength (15°C)	MN/m^2	125	275	310	158
Flexural strength (15°C)	MN/m^2	—	440	380	100
Coefficient of expansion $\times 10^6$:					
normal to sheet	per °C	60	60	60	60
plane of sheet	per °C	10–12	10–12	10–12	10–12
Thermal conductivity	W/(mK)	0.2–0.3	0.2–0.3	0.2–0.3	0.2–0.3
Electric strength (15°C)					
normal to laminate r.m.s.	kV/mm	30	32	32	32

Type of mica and resin		*P* *Silicone*	*M or P* *Epoxy*
Tubes			
Density	kg/m^3	1800	1600
Maximum operating temperature	°C	350	200
Water absorption	mg/cm^2	1.5	0.8
Cohesion between layers, wall 1.6 mm	N	44	267
3.2 mm	N	133	890
Electric strength, r.m.s.:			
normal to laminae in oil (90°C), wall 1.6 mm	kV/mm	12	20
3.2 mm	kV/mm	10	16
in air (15°C), wall 1.6 mm	kV/mm	10	16
3.2 mm	kV/mm	10	16
along laminae in oil (90°C), 25 mm length	kV	26	40

* M, Muscovite mica; P, phlogopite mica. Test methods of BS 2782-Parts 3 and 4:1995 and BS 6128 Parts 8 and 9:1982.

up to 25 mm or more; also *mica-paper* materials, comprising layers of mica-paper bonded under heat and pressure with a wide variety of resins, including high-temperature types (see *Table 7.16*).

7.7.2.2 Pressboard

This is a paper product, widely used in oil-immersed transformers in thicknesses up to 10 mm. Thicker boards are built up by bonding plies together with adhesives such as phenolic resins or casein. In the USA, pressboards are available with special treatments which are claimed to permit operation at higher temperatures than untreated materials. In one case the material is treated with an amine; in another the cellulose molecule is modified chemically by cyanoethylation.

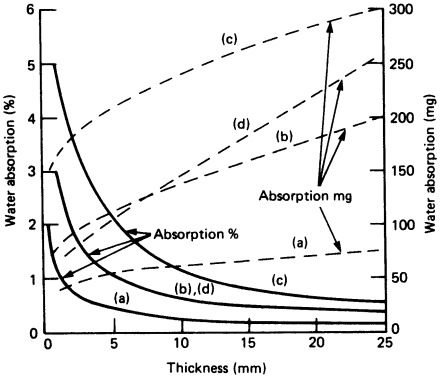


Figure 7.4 Water absorption of typical synthetic resin bonded paper and fabric boards: (a) type I; (b) type II; (c) type III; (d) type IIIA

7.7.2.3 Laminates

These are sheets of paper, fabric, etc., bonded with gums, shellac or synthetic resins under heat and pressure usually in hydraulic presses. Resins can be phenol formaldehyde, melamine formaldehyde, polyester, epoxy, silicone, polyimide, polyamide imide, etc. These are some of the most useful insulating materials for panels, terminal boards, coil flanges, packings, cleats, slot wedges and many other uses. The principal varieties are: *synthetic resin bonded paper* in thicknesses of 0.2–50 mm in several grades. *Synthetic resin bonded cotton* fabric (0.2–100 mm), the most common being bonded with phenolic resin, but an epoxy bonded type has been developed; both are tough and have good machinability.

Curves for water absorption for some types of s.r.b. paper and cotton fabric are given in *Figure 7.4*. *Synthetic resin bonded asbestos* paper, fabric and felt, used for low-voltage work at somewhat higher temperatures, e.g. 130–150°C. *Synthetic resin bonded glass fibre,* bonded under heat and pressure with synthetic resins of the melamine, epoxy, phenolic, polyester and silicone types in the thickness range up to 12.7 mm. Most of these materials have good electrical and mechanical properties and low water absorption (see *Table 7.17*). Similar materials are available where the reinforcement is a web of random laid glass fibres—known as mat. These materials are generally cheaper than those based on fabrics, although some of the properties are not as good. By arranging for a preponderance of the fibres to be in one direction, it is possible to produce boards with very good mechanical properties in certain planes. Such boards bonded with epoxy resins can have flexural strengths as high as 1300 MN/m^2 and flexural moduli approaching 5 GN/m^2.

In addition to the resins mentioned, high-temperature materials are now available which are thoroughly suitable for use with all forms of glass and other reinforcement. Some of these materials give thermal lives as good as the silicones but with mechanical properties more nearly equal to the epoxies. Typical materials in this class are resins based on acrylics, polyimides, polyamide imides and combinations of polyaralkyl ether with phenols, etc. Also, resins are now available for all types of laminate especially phenolic, epoxy and polyester where, in order to decrease the fire risk, the resin has flame-retardant properties (see *Table 7.18*).

Synthetic resin bonded (SRB) laminates of various types are made with a thin layer of copper (or other metals such as nickel and cupronickel) bonded to one or both surfaces. These materials are known as metal clad (or specifically copper clad) laminates and are used for printed circuit board applications.

A material in rigid form is produced by bonding high-density polyamide paper plies by heat and pressure only to give a tough, strong board with a long life at temperatures up to 220°C.

Other materials from the wide range of plastics are available as rigid boards or sheets although most of these materials are thermoplastic. Often such materials are moulded to produce finished products but it is possible to cut shapes and panels from sheets. Typical materials are based on polyvinyl chloride, acrylonitrile–butadiene–styrene, polyolefin, polymethylmethacrylate, etc. Many are flame-retardant or self-extinguishing; and reinforcing materials such as glass and asbestos fibres, glass spheres and mineral fillers can be added to improve mechanical properties and resistance to high temperatures. The well-known material ebonite, a mineral-filled rubber-based material, was a forerunner of this class.

The properties of typical sheets and boards are given in *Table 7.19*.

7.7.3 Tubes and cylinders

Tubes and cylinders can be produced by several methods. Materials capable of being moulded can be treated by compression or casting techniques. Many plastics can be extruded. Vulcanised fibre tubes are wound from paper treated with zinc chloride solutions. Laminated materials are generally wound on special machines with heated rollers

Table 7.17 Properties of synthetic resin bonded glass fabric laminates*

Property		*EP1*	*EP2*	*MF1*	*PR1*	*PR2*	*SIL1*	*SIL2*	*SIL3*
Water absorption†		10	20	118	20	75	9	11	47
Tensile strength	MN/m^2	173	207	103	138	173	84	117	110
Cross-break strength	MN/m^2	240	310	103	207	138	90	117	90
Impact strength†	Nm	2.8	4.8	4.1	4.1	5.5	2.8	4.1	5.5
Ins. resistance (wet)	MΩ	100	100	1	10	—	1000	100	10
Relative permittivity at 1 MHz	—	5.5	5.5	7.5	4.5	4.9	4.0	4.3	4.8
Loss tangent at 1 MHz	—	0.035	0.035	0.025	0.04	0.05	0.003	0.004	0.01
Electric flat-wise	kV/mm	6.3	6.3	2.8	8.3	7.1	—	—	—
Strength edge-wise‡⇐	kV/mm	30	30	15	35	30	30	25	20

* EP, Epoxy resins; MF, melamine formaldehyde resins; PF, phenol formaldehyde resins; PR, polyester resins; SIL, silicone resins.
Test methods of BS 2782:1995 and BS 3953:1990.

† Per 12.7 mm width, 3 mm thick.

‡⇐R.m.s. for 25 mm length at 90°C.

Table 7.18 Properties of typical high-temperature laminates*

Property		*A/AP*	*S/AP*	*P/GF*	*PI/GF*	*PEP/GF*	*PEP/AF*
Density	kg/m^3	1700	1720	1600	1800	1770	1650
Maximum operating temperature	°C	180	>220	300	280	250	250
Tensile modulus (*r*)	GN/m^2	—	—	19	24	37	14
Tensile strength (*r*)	MN/m^2	150	120	360	380	435	130
(250°C)	MN/m^2	—	—	275	—	300	—
Cross-break strength (*r*)	MN/m^2	280	190	440	450	690	190
(288°C)	MN/m^2	—	—	296	340	—	—
Impact strength‡⇐(*r*)	Nm	0.6	—	11	—	14	2.5
Relative permittivity (*r*, 1 MHz)	—	—	—	3.6	—	4.8	—
Loss tangent (*r*, 1 MHz)	—	—	—	0.012	—	0.011	—
Electric strength (*r*)‡⇐	kV/mm	9–20	10.5	—	—	27–34	—

* A/AP, acrylic + asbestos paper; S/AP, silicone + asbestos paper; P/GF, polyimide + glass fabric; PI/GF, polyamide imide + glass fabric; PEP/GF, PEP/AF, polyaralkye ether/phenol + glass or asbestos fabric. *r*, room temperature.

† Per 12.7 mm width.

‡⇐R.m.s.

Table 7.19 Properties of typical rigid sheets and boards

Property		*VF*	*PB*	*SP*	*SW*	*SF*
Density	kg/m^3	1300	1000	1330	1320	1330
Water absorption	%	20–50	150	0.3	0.8	0.7
Elastic modulus	GN/m^2	5	Low	10	17	6
Tensile strength	MN/m^2	80	40	60	100*	75
Shear strength	MN/m^2	75	—	36	40*	90
Cross-breaking strength	MN/m^2	—	—	100	150*	140
Crushing strength	MN/m^2	—	—	240	200	240
Compression at 70 MN/m^2	%	10	30	1.3	—	3.2
Relative permittivity	—	2.5–5	3.2†	4.5	4.5	10
Loss tangent at 50 Hz	—	—	0.015‡⇐	0.02	0.02	0.3
Electric strength (1-min) r.m.s.						
through laminae	kV/mm	1.4–4	8.8	8	4	1
along laminae (25 mm length)	kV/mm	1.2	2	1.4	2	0.6

VF, BS 6091 Parts 1 and 2:1995, Grey Vulcanised Fibre.
PB, BS EN 60641:1996, Absorbent Pressboard Grade II.
SP, BS 5102:1974, Synthetic Resin Bonded Paper Type I.
SW, BS 2572:1990 and BS EN 60893-Part 2:1995, Synthetic Resin Bonded Wood.
SF, BS 2572:1990 and BS EN 60893-Part 2:1995, Synthetic Resin Bonded Fabric. Type 3B.
* Varies considerably with relation of grain to lamination.
† Dry: 4 when dried and oil-impregnated.
‡ ~~D~~ry: 0.04 when dried and oil-impregnated.

while tension and pressure are applied to consolidate the layers. In this way SRB paper and fabric tubes and cylinders can be produced by rolling the treated material convolutely on heated mandrels. Some of the smaller tubes are moulded in a split mould while still on the mandrel. Cylinders (considered as tubes with internal diameters above 76 mm) are made by similar techniques. These tubes which are generally for use in large transformers may have diameters as high as 2 m and lengths of 4 m. SRB cotton fabric tubes and cylinders are produced by similar methods; SRB glass-fibre tubes and cylinders can be rolled from most of the rigid-board materials.

Pressboard tubes and cylinders (mainly for oil-immersed transformers) are produced by winding presspaper on mandrels under tension, applying adhesives resistant to transformer oil (gum arabic, casein, phenolic resins, etc.). Tubes previously rolled from shellac bonded micanite for high voltage use are now widely superseded by tubes rolled from mica paper bonded with epoxy or silicone resins.

In addition to the convolutely wound tubes, a wide range of products can be produced by helical winding of strips (papers, fabrics, film, etc.), adhesives being applied meanwhile. The edges of the strip are generally butted together and this technique makes it possible to produce tubes the length of which is limited only by the requirements of transport.

Yet another method of winding tubes, chiefly with glass fibres, is known as filament winding. Strands of resin treated glass fibre are applied to mandrels in special winding machines. Restrictions on shape are less stringent, and very good mechanical properties are obtained.

7.7.4 Flexible sheets, strips and tapes

In very many applications a certain amount of flexibility is necessary, mainly to enable the materials to be readily applied to conductors, coils and various shapes, often irregular. The following are the principal flexible insulating materials, used for such purposes as wrappings on conductors and connections of machines and transformers, bus-bars and other parts; interlayer and connection insulation of coils; and slot linings of armatures.

7.7.4.1 Micanite

Tapes and sheets of micanite (micafolium) are widely used for high-voltage and high-temperature machine windings. It consists of mica splittings bonded with gum, bitumen or synthetic adhesive, often backed with thin paper or fabric (especially glass fibre) to assist taping and to give mechanical support for micanite slot liners and coil insulation. Synthetic bonds, especially epoxies and silicones, are used for the higher temperature applications, i.e. for classes B, F and H. Similar sheet and tape materials are made from mica-paper produced by a paper-making process using minute particles of mica, the sheet being treated subsequently with shellac or synthetic resins; these mica-paper products are like micanite but are more adaptable and uniform.

7.7.4.2 Vulcanised fibre and presspaper

These are useful flexible materials for coils, slot liners and many uses in machines, transformers and other apparatus.

7.7.4.3 Papers

Chiefly made from wood pulp, cotton or manila fibres, these are employed in capacitor and cable manufacture, as insulations in coils and in the manufacture of SRB paper boards, tubes and bushings, also as backing material for flexible micanite products. Asbestos paper had uses for high-temperature conditions.

Papers are produced from other bases, such as ceramic, silica and glass fibre. A polyamide paper will not melt or support combustion, and has good electrical and mechanical properties; as with most cellulose papers it can be supplied in creped form to facilitate the taping of irregular forms, and it can be obtained in combination with mica platelets to give greater resistance to high-voltage discharges. Papers made

Table 7.20 Properties of typical flexible sheets*

Material		*VF*	*PP*	*CP*		*PF*	*MP*
				Dry	*Oil*		
Density	kg/m$^{3/}$	1260	1050	1050	—	950	1600
Thickness	mm	0.25	0.25	0.25	0.25	0.25	0.15
Maximum operating temperature	°C	90	90	90	120	220	600
Tensile strength, *m*	kN/m	31	19	12	12	30	1.5
c	kN/m	14	8	4	4	18	1.2
Tearing strength, *m*	kg	0.42	0.38	0.26	0.26	0.55	—
c	kg	0.46	0.41	0.29	0.29	0.9	—
Resistivity (20°C) (dry)	Ω-m	—	—	10^{16}	10^{12}	10^{14}	10^{13}
Relative permittivity (20°C, 50 Hz)	—	2.8	3.1	1–2.5	3–4	2.6	—
Loss tangent (20°C, 50 Hz)	—	0.05	0.012	0.0025	0.0025	0.01	—
Electric strength† in oil	kV/mm	16	60	—	50–80	—	—
in air (90°C)	kV/mm	15	11	9	—	30	20

VF, BS 6091:1987, Vulcanised Fibre.
PP, BS EN 60641:1996, Presspaper Grade 11.
CP, BS 4295:1968 and BS 5626 Parts 1, 2 and 3:1982, Cellulose Paper (dry, or impregnated in mineral oil).
PF, Polyamide Fibre Paper.
MP, Muscovite Mica Paper.
* *m*, in machine direction; *c*, cross direction.
† R.m.s.

from polyester fibres are used normally in combination with materials such as polyester film.

Properties of some flexible sheets are given in *Table 7.20*.

7.7.4.4 Fabrics

Cotton, nylon and polyethylene terephthalate (PETP) cloths are used mostly as bases for varnished fabrics for small coils where flexible dielectrics 0.1–0.25 mm thick are wanted. Woven asbestos and glass-fibre cloths are applied where temperatures are too high for organic textiles. The introduction of PETP fabrics with heat-resisting (e.g. alkyd and polyurethane) varnishes has provided a range for class E or even class B insulation. Varnished glass-fibre cloths and tapes—for which special coating materials have been formulated, including silicone resins and elastomers—are able to meet most of the high-temperature requirements of classes B, F and H apparatus. Fabrics of polyamide fibre can be coated with high-temperature resins and elastomers to give a material suitable for use at temperatures above 200°C.

Table 7.21 gives some typical properties.

7.7.4.5 Tapes

Pressure-sensitive adhesive tapes are used in the construction of all types of equipment and are invaluable for holding and positioning conductors, preventing relative movement, identification of parts, exclusion of moisture, etc. Many types of such material are available based on most of the papers, fabrics, films and metal foils together with adhesives which can be thermoplastic or can be made to cure on the application of heat. Extra care has to be taken to reduce corrosiveness since these tapes are often used in fine-wire coils where electrolysis can cause erosion of a conductor.

7.7.4.6 Films

Many plastic materials are available in film form and many are used for the production of composite materials. Films with thicknesses of about 0.008 mm up to about 0.5 mm are usual. Most are strong, have good electrical properties and good resistance to moisture, but some have certain limitations in operating temperature. The thermoplastics often soften or melt at comparatively low temperatures and some films which melt at very high temperatures are damaged by oxidation processes at much lower temperatures.

Cellulose acetate, triacetate and similar acetates These films have been available for a long time and still find use in machines and in coils at temperatures up to class Y conditions.

FEP This copolymer of tetrafluoroethylene and hexafluoropropylene (which is hardly affected by any known chemicals and solvents) has a service temperature range from about −250°C up to more than 200°C. It can be bonded to itself and to other materials by heat and pressure and has excellent high-frequency characteristics.

PTFE and PTFCE Polytetrafluorethylene and polytrifluorochloroethylene have excellent chemical stability and electrical properties. The former material does not soften, but degrades above 300°C; the latter is thermoplastic at high temperatures. Both materials suffer from flow at high pressures even at moderate temperatures. Operating temperatures are otherwise similar to those given for FEP above.

PVF and PVF2 Polyvinyl fluoride and polyvinylidene fluoride are produced by substituting fluorine for some of the chlorine in PVC. Both have the excellent temperature and chemical stability shown by fluorine substituted materials. Electrical properties are not as good as those of PTFE and FEP.

Polyethylene This is not widely used as electrical insulation, mainly because of the low softening temperature. Controlled radiation with high-energy electrons produces a film which has similar properties to the original film but because cross-linking has occurred, there is no sharp melting point. This enables the material to be used at higher

Table 7.21 Properties of typical varnished textiles*

Property		*Cotton Y/B*	*Nylon Y/B*	*PETP Y*/B*	*Glass Y*/B*
Thickness	mm	0.25	0.15	0.15	0.15
Tensile strength, wp	kN/m	8/9	9	9	21
wf	kN/m	6/7	5	5	16
Tearing strength, wp	kg	0.25/0.22	0.12	0.1	0.15
wf	kg	0.26/0.23	0.3	0.21	0.34
Electric strength (1 min):† large electrodes at:					
20°C	kV/mm	—	36/38	34/40	44/42
90°C	kV/mm	24/32	32/32	30/34	32/40
150°C	kV/mm	—	—	28/30	18/19
6.35 mm diameter electrodes (20°C), tapes stretched by:					
0%	kV/mm	—	52/58	52/62	42
5%	kV/mm	—	50/56	50/60	40
10%	kV/mm	—	50/54	50/54	28
20%	kV/mm	—	26/48	22/20	—

* Test methods of BS 419:1966. PETP, polyethylene terephthalate; Y, yellow; B, black. wp, Warp; wf, weft.

‡ Special heat-resistant varnish.

† R.m.s.

temperatures without damage; also, with this treatment, the material can be made heat shrinkable.

Polyethylene terephthalate These films are widely used for slot insulation in motors, as the dielectric in capacitors, for coils, etc. For slot insulation the film is frequently combined with polyester fibre paper but may be used alone in smaller equipments. It is suitable for class E temperatures although some manufacturers claim it can be used at class B temperatures. When protected from oxygen by films or varnishes with higher temperature resistance, some users claim temperatures up to class F are suitable for this material.

Polyamide (nylon) Films show moisture absorption in humid atmospheres, but are strong, solvent resistant and have a high melting point; they cannot be used at temperatures exceeding 80°C in air because of oxidation.

Polycarbonate The film is strong, flexible and highly stable against temperatures up to 130°C. Electrical properties are excellent and the film is finding use as a capacitor dielectric.

Polypropylene This dielectric film has a higher softening point than polyethylene film and the dielectric properties are similar. The material is resistant to hot mineral oils and chlorinated polyphenols and is finding widespread use in low-voltage and power capacitors.

Polyimide An outstanding film of recent development. The material has no melting point and can withstand exposure at temperatures above 500°C for several minutes. Life at 250°C is over 10 years, and more than 10 h at 400°C. It remains flexible down to the temperature of liquid helium. The film is replacing glass and mica insulation in motors especially for traction. As thin layers can be obtained, considerable space saving (often as high as 50%) can be achieved. The film cannot be bonded easily, but can be supplied with a very thin layer of thermoplastic FEP on one or both sides, and layers can be bonded at temperatures approaching 300°C.

Polyvinyl chloride These films have low water absorption and resist most chemicals and solvents. The high elongation makes it possible to apply tapes neatly and tightly over irregular shapes. This class of material is often used as a pressure-sensitive tape.

A summary of the properties of materials available in the form of thin film is given in *Table 7.22*.

Silicone elastomers Silicone rubbers or elastomers are a range of heat stable elastic silicone materials used for electrical insulation as sheet, tape, wire and cable coverings, extruded sleevings and mouldings, unsupported, but more extensively as coated glass-fibre cloths, tapes and braided glass sleevings. Such fabric tapes impregnated with a silicone elastomer are available straight or bias cut in thicknesses of 0.07–0.5 mm. Other fabric tapes, coated on one side with partially cured silicone elastomer, are used for taping bars and coils and the curing subsequently completed by baking. The cured taping forms a homogeneous coating having good resistance to moisture, discharges and heat, e.g. for continuous use at about 180°C and up to 250°C for short periods, and possessing high electric strength and low dielectric losses. Most grades of silicone rubbers remain flexible at temperatures down to −60°C and some can be used down to −90°C.

7.7.4.7 Composite sheet insulations

These combine two or more materials, especially varnished cloth, presspapers, vulcanised fibre, plastic films bonded with special adhesives. Particularly widely used have been cellulose fibres combined with polyester or cellulose acetate film, polyester fibre paper combined with polyester film and various combinations of mica, varnished glass fabrics and polyester film. Among the newer developments are combinations of aromatic polyamide paper with polyimide film, giving a material suitable for use at temperatures above

Table 7.22 Properties of materials as thin films*

Property			*CA*	*CT*	*FEP*
1 Density		kg/m^3	1290	1290	2150
2 Tensile strength		MN/m^2	50–80	60–110	21
3 Elongation at break		%	15–60	10–40	300
4 Resistivity (20°C)		Ω-m	10^{13}	10^{13}	$>10^{16}$
5 Relative permittivity (20°C)	1 kHz	—	3.6–5	3.2–4.5	2.0
6	1 MHz	—	3.2–5	3.3–3.8	2.0
7 Loss tangent (20°C)	1 kHz	—	0.015–0.03	0.015–0.025	0.0002
8	1 MHz	—	0.025–0.05	0.03–0.04	0.0002
9 Electric strength (20°C), r.m.s.		kV/mm	60	60	200

Property	*PTFE*	*PTFCE*	*PA*	*PVC*	*PET*	*PI*	*PVF*	*PVF2*
1	2200	2100	1140	1260	1400	1420	1380–1570	1750
2	20	40	50–80	13–30	170	170	100–140	40–45
3	200–300	200–350	250–500	150–350	100	70	100–200	150–500
4	$>10^{10}$	10^{16}	$>3 \times 10^{11}$	10^9–10^{12}	10^{16}	10^{16}	$>10^{11}$	2×10^{12}
5	2.1	2.8	3.8	3–7	3.2	3.5	8.5–9	8.0
6	2.1	2.5	3.4	—	—	—	—	6.6
7	0.0005	0.016	0.01	0.01–0.02	0.005	0.003	0.015	0.018
8	0.0002	—	0.025	—	—	—	—	0.17
9	60	120–200	60–70	10–40	280	280	160	50

* CA, cellulose acetate. PA, polyamide. PVF, polyvinyl fluoride. PVF2, polyvinylidene fluoride. PTFE, polytetrafluorethylene. PTFCE, polytrifluorochloroethylene. FEP, fluorinated ethylene propylene. PET, polyethylene terephthalate. CT, cellulose triacetate. PI, polyimide. PVC, polyvinyl chloride (plasticised).

220°C. Another combination is aromatic polyamide paper with polyester film, having good class B performance.

7.7.5 Sleevings, flexible tubings and cords

Tubular sleevings made of cotton, polyester, asbestos, glass, ceramic, quartz, silica, polyamide and other fibres supplied untreated in flat tubular form are suitable for low-voltage insulation of connections of coils, etc. Glass sleeving given a high-temperature treatment before use has the fibres set in place, thus preventing unravelling during assembly. Varnishing of all these sleevings during treatment of the coils is the usual practice.

Similar sleevings may be varnished or treated with resins and polymers before use; these types are known as coated sleevings. Most colours can be produced, thus helping in identification of circuits. The coating materials in general use for cotton and rayon sleevings are natural or synthetic varnishes. For glass fibres, high-temperature polyvinyl chlorides, polyurethanes and silicone elastomers are widely used as well as high-temperature varnishes such as those based on acrylics, polyesters, silicones or polyimide resins. Coatings of fluorinated resins on glass fibres are available.

Flexible tubings can be extruded from most of the materials mentioned in Section 7.7.4.6, but as many of these materials are thermoplastic and also liable to oxidation problems, care is required at temperatures above 90°C.

Other methods of making tubing are by helical winding with adhesives as a bond, when very thin walls (e.g. 0.025 mm) are required or only small quantities are needed. Many of the extruded tubings and some of the helix types can be caused to shrink by applying heat after the tubes have been assembled. Most shrinkage takes place in diameter but there is sometimes a small amount in length. These sleeves are useful for insulating connections and joints, for covering capacitors, resistors, diodes, etc., for colour coding, sealing and moisture proofing. The two main shrinkage processes are: (1) thermoplastics stretched during manufacture, and allowed to cool so that the mechanical strains are frozen in, shrink next time the material is heated; and when certain materials are exposed to high-energy radiation, molecular cross-linking occurs, and they shrink when the temperature is subsequently raised.

Cords for lashing bundles of switchboard wiring, holding leads in position and for tying down coils in machines and transformers, may be made from yarns laid together and twisted or braided, or from narrow woven or slit tape. The trend is away from cellulose materials (linen and hemp) to synthetics, especially glass and polyester fibres. Because of the poor abrasion resistance of glass fibres, glass cords are often pretreated with epoxy or silicone resins or with various polymers. Normally the type of treatment depends on the application.

For high-voltage machines, a useful material is a round cord 3–5 mm diameter comprising a central core of straight polyester fibres surrounded by a braided polyester sheath. This material shrinks slightly on heating, and tightens. For smaller windings, polyester mat can be slit into tapes and several tapes twisted together. Polyester fibre cords of all kinds have the advantage over glass cords that knots can be made without appreciable loss in strength. Glass and polyester bands can be treated after application with varnishes and both have good resistance to abrasion and to mould growth. For rotors straight unidirectional glass-fibre bands pretreated with epoxy, polyester, or acrylic resins are used. These bands must be applied under controlled tension to ensure that the windings are held against centrifugal force. Such bands can be placed safely near voltage carrying parts and (unlike wire bands) are not affected by magnetic fields.

For switchboard wiring, cable harnesses, etc., neat lashings can be made with small-diameter extruded polyvinylchloride threads. These materials do not burn readily and are obviously well suited for use with PVC insulated switchboard wiring.

7.7.6 Wire coverings

Wires for coils and armature conductors are insulated with lappings of cotton, asbestos, glass fibre and polyamide fibre, all of which are hygroscopic and require treatment with oils, compounds or varnishes, usually after winding. Relevant specifications are BS 1497, 1933, 2479, 2480 and 2776. A recent addition to the range of wire and strip coverings is lapped polyimide film, bonded to itself and to the conductor with FEP polymer: the films are thin and flexible, electrically good and workable at 220°C.

Orthodox oleo-resin (BS 156) and polyvinyl formal and acetal enamels (BS 1844) are tough, resistant to abrasion and to softening by varnishes. Enamelled wires have a better space factor than those with fibre covering.

Some wires are covered in separate operations with synthetic enamels of widely differing characteristics, giving a range of finely balanced properties. The inner coat can be considered as the insulant; the outer coat can give improved resistance to solvents and abrasion, better cut-through resistance, heat-bonding of turns into a solid coil, etc. Wires coated with polyurethane polymers are useful because they can be soldered without first removing the enamel (BS 3188).

For high-temperature working, the performance of PTFE coverings is limited by cold flow of the insulant under mechanical pressure. Silicone resin gives a high-temperature covering but the solvent resistance and mechanical properties are not very good. Better mechanical characteristics are given by polyimide, polyamide imide, polyester imide and polyhydantoin, but these materials are still uncommon.

7.7.7 Moulded and formed compositions, plastics, ceramics, etc.

Articles of various shapes, often quite complicated, which cannot readily or economically be matched or built up from sheet, rod or tube materials may be obtained by forming, moulding, coating or casting one of the 'plastics' (which are all essentially organic), or a ceramic such as porcelain, or glass or other inorganic materials. The materials can be grouped approximately as follows.

Organic thermoplastic Examples are: compounds made of natural gums, bitumen, etc., not specially processed; synthetic resin products such as polyvinyl chloride, nylon, polystyrene; cellulose derivatives such as cellulose acetate.

Organic thermosetting These are, chiefly: cured shellac products including micanite (for simple shapes, e.g. commutator cones); rubber sulphur and similar vulcanisable compositions; compounds made from synthetic resins of the phenolformaldehyde type, urea formaldehyde, silicones, polyesters, alkyds, epoxies and polyurethanes.

Inorganic The principal materials are listed below.

(1) *Asbestos–cement* compositions, mainly for high heat resistance, particularly for parts exposed to arcs;
(2) *Concrete*, cast or moulded, for inductors and switchgear where strength and fire-resistance are needed;
(3) *Porcelain*, mainly for out-door use and other cases where dust and moisture collect readily, also special grades for high temperatures;
(4) *Steatite*, for uses similar to those of porcelain;
(5) *Special ceramics* for radio capacitors, sparking plugs, etc.;
(6) *Fire-clay* for holding electric heating elements;
(7) *Glass* for out-door insulators, lamp bulbs, valves and high-frequency insulation;
(8) *Mica-glass* composition for high-temperature applications requiring good electrical properties, especially at high frequency.

7.7.8 Methods of moulding and forming materials

7.7.8.1 Organic materials

The principal methods of manufacturing parts from the organic materials are as follows.

Forming of laminated and other sheet materials in open moulds or other forming tools with moderate pressure, the material being made plastic by a suitable liquid or, more usually, by heat. Such forming is generally restricted to relatively simple shapes such as channels, cones, tubes, collars, spools and wrappings. Heat is usually applied during the forming operation, and the material sets either by cooling, heat-treatment in the mould, or subsequent drying, baking, etc. (e.g. vulcanisation).

Moulding under pressure in closed metal moulds ('compression moulding'), using a powder, dough, treated paper, treated fabric or other form of moulding material and heat treating during or subsequent to moulding.

Injecting material, made plastic by heat or other means, into moulds under pressure and setting by cooling, or heat processing.

Extruding plastic material as for injecting but not into moulds, the shape being determined by the orifice or die through which the material is extruded and the setting being generally due to cooling, but may be followed by a further process such as vulcanisation.

Casting a liquid or molten material into moulds (without pressure) and conversion to a solid condition by cooling or heating (thermosetting materials), or by the action of chemicals with or without heat.

Coating with polymers in powder form, spread on heated metal components by several methods including electrostatic fields or fluidised bed techniques. Thick coatings with good electrical properties can be applied overall. Conversion to the solid is by cooling with thermoplastic materials, or by further heat treatment for thermosetting materials. Some polymers can be applied to surfaces (which need not be metal) in the form of dispersions in a suitable liquid carrier, followed by curing by heating.

Properties of typical organic moulding compounds are listed in *Table 7.23*.

7.7.8.2 Inorganic materials (ceramics, etc.)

The methods used for producing articles from the inorganic materials are, briefly, as follows.

Asbestos-cement compounds Made from asbestos and other minerals, e.g. powdered silica, mixed wet with lime or Portland cement and moulded *cold* by compression moulding. After removal from the mould the parts are cured in live steam.

Concrete The large mouldings for inductors are made of high-grade concrete, made from specially selected Portland cement, sand and aggregate, the wet mixture being poured and 'puddled' into suitable moulds and, after preliminary setting, the parts being cured in live steam.

Porcelain Made from china clay *(kaolin)*, ball clay, quartz and felspar, finely powdered and mixed with water. Small parts (e.g. tumbler switch bases) are made by the *dry process*

Table 7.23 Properties of typical organic and inorganic moulded and formed materials*

Organic	*Binder: Filler*	*PF Wood filler*	*PF Wood filler*	*PF Asbestos*	*PF Mineral powder*	*UF Wood*	*Rubber Mineral*	*Cellulose Acetate*
Density	kg/m^3	1400	1360	1900	1880	1500	1700	1300
Plastic yield temperature	°C	>100	>140	>180	>140	>100	80	60–80
Coefficient of expansion $\times 10^6$	per °C	40	40	30	30	45	80	160
Water absorption†⇐	%	0.3–0.4	0.3	0.05	0.02–0.05	0.5	<0.01	1.5–3
Elastic modulus	GN/m^2	5	5	7	3	7	3	2
Tensile strength	MN/m^2	48	48	35	35	63	24	31
Cross-break strength	MN/m^2	63	70	49	52	105	67	49
Crushing strength	MN/m^2	240	240	160	120	230	140	140
Impact strength	Nm	0.3	0.3	0.2	0.3	0.3	0.17	0.14
Resistivity	Ω-m	10^9	5×10^{10}	10^9	5×10^8	10^6	10^{14}	5×10^6
Surface resistivity	MΩ/sq.	5×10^4	10^6	10^4	2×10^7	5×10^5	3×10^5	4×10^6
Relative permittivity	—	7–10	4–8	8–18	5	9	4.1	4–6.5
Loss tangent, 50 Hz	—	0.25	0.04	0.2–0.6	—	0.08	0.016	0.016
1 kHz	—	0.2	0.04	0.1–0.4	0.02	0.06	0.012	0.03
1 MHz	—	0.15	0.035	0.1	—	0.04	0.01	0.06
Electric strength‡⇐	kV/mm	1.2–4	9	3–4	15	3	15	12

Inorganic		*Porcelain*	*Steatite*	*Aluminium oxide*	*Glass*	*Glass mica*	*Asbestos cement*	*Glass ceramic*
Density	kg/m^3	2400	2750	3650	2250	2680	1600	2600
Plastic yield temperature	°C	>1200	1400	—	600	450	>700	1250
Coefficient of expansion $\times 10^6$	per °C	4	6–8	6.2	3.2	9.8	—	5.7
Water absorption (20°C, 24 h)	%	0	0.01	—	0	0	10–15	0
Specific heat	J/kg	900	840	750	840	840	—	750
Tensile strength	MN/m^2	35	56	75	—	42	7	150
Flexural strength	MN/m^2	70	100	330	—	93	28	140
Crushing strength	MN/m^2	420	840	1670	970	270	55	—
Resistivity (20°C)	Ω-m	10^{16}–10^{13}	10^{13}	10^{14}	>ς10^{12}	10^{11}	—	10^{14}
Relative permittivity (20°C), 1 MHz	—	5–7	4.1–6.5	10	4.5–4.9	6–7.5	—	5.6
Loss tangent (20°C), 1 kHz	—	—	0.005	0.008	0.005	0.007	—	0.0025
1 MHz	—	0.006	0.0045	0.0006	0.003	0.002	—	0.0015
Electric strength§	kV/mm	6–16	8–15	48	14	20	0.8–4	—

* Test methods of BS EN 69243 Parts 1, 2, and 3:2001 and BS 771:1992. PF, phenol formaldehyde; UF, urea formaldehyde.

† Cellulose acetate, 24 h; remainder 7 days.

‡ R.m.s., 3 mm at 90°C.

§ R.m.s., 20°C.

(or *die pressing*), in which a slightly damp mixture is compressed in steel moulds. Many high-voltage parts, particularly large pieces such as transformer bushings, are made by the *wet process* from a wet plastic mixture, shaped on a potter's wheel and turned on a lathe. Others, such as overhead line insulators, are made by pouring a *creamy* mixture into plaster moulds in which partial drying occurs. Parts formed by the foregoing processes are then dried and usually coated with glaze, after which they are fired in a kiln at temperatures such as 1200–1400°C.

Steatite Consists of powder soapstone (talc) die pressed dry or moulded wet, similar to porcelain, and finally fired at about 1400°C.

Special ceramics A number of ceramics are made, similar to porcelain and steatite, from such materials as rutile (a form of titanium dioxide)—having low losses and high permittivity, and suitable for high-frequency capacitors—and aluminium oxide, mainly for sparking plugs. A ceramic composed of barium titanate has exceptionally high permittivity.

Fire-clay refractory ceramics Made from special grades of clay, usually die pressed and fired, somewhat as in the case of porcelain.

Glass Made from powdered silica mixed with metallic bases (soda or potash) and a flux (e.g. borax). The mixture is fused at temperatures of the order of 1200–1400°C, and the molten mass 'blown' into moulds, or forced into moulds under pressure, the parts being removed when cool.

Developments have been made in which the glass instead of remaining in its usual super-cooled liquid state is devitrified, with the result that a fine-grain structure develops and the mechanical properties are much improved. These new materials are known as glass ceramics and can have specially controlled properties. In particular the coefficient of thermal expansion can be selected from high positive to negative values.

Mica glass composition Produced from powdered mica and glass, heated to a semi-molten condition and moulded in steel moulds by compression or injection at high pressure.

Properties of typical ceramics and other inorganic formed or moulded compositions are summarised in *Table 7.23*.

7.8 Composite solid/liquid dielectrics

Combinations of solids and liquids are widely used for electrical insulation. The solid normally consists of sheets of paper or of polymer film with an insulating liquid, typically a mineral oil, as the liquid component of the composite. Sometimes a synthetic oil is used instead of a mineral oil, especially when polymer films are used as the solid component. Particular applications of composite insulation are power cables, transformers, bushings and capacitors. The solid provides mechanical separation between the conductors and the liquid gives high electric strength with a low dielectric loss.

7.8.1 Breakdown mechanisms in composite dielectrics

The construction of composite dielectrics usually starts with the solid material in tape form that is wound around one of the conductors. Several layers of tape are used and there is a small gap between adjacent tapes, with the gaps staggered between layers (*Figure 7.5*).

The most widely used composite insulation system has the spaces between the tapes filled with a mobile insulating liquid. The system that is used in many power engineering applications such as transformers and oil-insulated switchgear is paper and mineral oil. The dielectric constant of paper and mineral oil is very similar, approximately 2.5, but very different from that of air. It follows that if there are air bubbles in the system there is a much higher electric field in the air bubbles than in the rest of the insulation. Bubbles may become trapped in the gaps between the tapes. The air in such bubbles will break down at a much lower applied voltage than the rest of the insulation, producing additional gas and eroding the solid insulation in contact with the bubble perimeter. This can eventually lead to breakdown of the complete insulation structure.

7.8.2 Oil/paper systems

This insulating system is excellent for high electric strength over long periods of time, provided the insulation is free from significant partial discharges. It is therefore essential that gas bubbles are prevented from forming. This is done by applying a vacuum to the insulation to remove dissolved gas where this is possible, and in insulation systems such as high-voltage power cables by applying several atmospheres pressure to inhibit bubble formation.

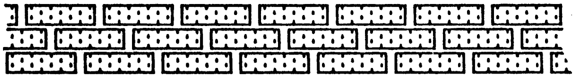


Figure 7.5 Tape arrangement for composite insulation

In laboratory-scale tests it was found that increasing the pressure of the oil from one atmosphere to 14 atm increased the electric stress that could be applied before partial discharges could be detected from 47 kV/mm to 69 kV/mm, or from 60 kV/mm to 90 kV/mm when a thicker oil was used. There is always a problem in relating the results of laboratory investigations to the design of practical apparatus because of the reduction in strength obtained as the volume of the insulation increases. In extreme cases the breakdown strength may be halved by two orders of magnitude increase in volume under stress.

In a series of tests on samples of oil/paper insulation of reasonable size the 50 Hz strength always exceeded 40 kV/mm and the 1/50 µs impulse strength was always greater than 90 kV/mm. The presence of moisture in the paper will give a 40% reduction in breakdown strength when the moisture content is increased from zero to 8%. The impulse strength of oil-impregnated paper increases as the thickness of the paper tapes decreases. In one set of tests the impulse strength was 130 kV/mm for 0.13 mm thick tapes and 160 kV/mm for 0.03 mm thick tapes.

The proven long-life-times of oil/paper systems mean that this insulation is the normal choice for insulating power transformers, cables and power-factor-correction capacitors.

7.9 Irradiation effects

Electrical equipment is being used increasingly in situations where it will be exposed to the effects of nuclear and other types of radiation, often at high energy levels. Such radiation can change the characteristics of many materials and may cause severe deterioration; on the other hand, radiation can have beneficial effects, especially by causing synthetic polymers to cross-link. Particular cases are the irradiation of polyethylene and other thermoplastic materials giving new products with improved properties, especially resistance to heat and mechanical failure; the vulcanisation of rubber; the polymerisation of plastics; the curing of resin and varnish films and the manufacture of heat-shrinkable films and sleevings.

7.9.1 Type of radiation

The more common forms of radiation encountered are: neutron and γςfrom nuclear reactors, neutron and γςfrom isotopes and electrons and X-rays from particle accelerators. The unit of absorbed radiation is the rad (=100 erg/g) or the megarad (=10^8 erg/g of material). The megarad is equivalent to 10 kJ/kg.

Although the effects of radiation on materials are cumulative and, therefore, dependent on the total dose, the dose rate (generally expressed as megarads per hour) may have some further effect. This point must be considered when experimental work is being carried out. Fortunately, it has been found that changes in most materials due to irradiation are practically independent of the type of radiation encountered. Hence, various sources of irradiation may be used for experimental work and those normally employed are:

(1) 'hot' fuel elements and other parts of nuclear reactors;
(2) radioactive isotopes (e.g. cobalt 60);

Table 7.24 Radiation resistance of organic insulating materials: probable useful-life dose

Material	Mrad	*Material*	Mrad
Gases		*Liquids*	
Sulphur hexafluoride	5000	Polyphenyls	5000
Difluorodichloromethane	1000	Radiation-resistant petroleum oil	2000
Trifluoromonochloroethylene	500	Transformer oil (naphthenic)	1000
Perfluoropropylene	100	Transformer oil (paraffinic)	500
		Silicone oil	200
Moulded or laminated plastics (filled)		Pyrochlor	100
Diphenyl silicone/glass	10 000		
Mineral-filled epoxyphenolics	10 000	*Resins, bitumens, etc. (unfilled)*	
Mineral-filled phenolics	4000	Diphenylsilicone	5000
Epoxy/glass cloth	4000	Polystyrene	5000
Cellulose-filled phenolics	1000	Polyvinyl carbazole	4000
Cellulose-filled urea formaldehyde	1000	Bituminous compounds	2000
		Nylon	2000
Elastomers		Polyethylene	2000
Polyvinyl chloride (plasticised)	500	High-impact polystyrene	2000
Polyurethane rubber	400	Polyurethane	1000
Butadiene styrene + antiradiant	300	Alkyd resins	500
Phenylmethyl silicone	200	Phenol formaldehyde resins	500
Polychloroprene	150	Polyethylene terephthalate	500
Natural rubber	150	Cellulose nitrate	100
Acrylonitrile	100	Cellulose butyrate	50
Polysulphide	80	Cellulose acetate	50
Dimethylsilicone	30	Methylmethacrylate	50
Polyisobutylene	20	Polytetrafluoroethylene	5

(3) Van de Graaf electron accelerator; and
(4) microwave linear electron accelerator.

The most convenient is the linear accelerator, in which a small-diameter beam can be made to scan the parts to be irradiated with uniformity of dosage. The physical conditions—temperature, ambient atmosphere (air, carbon dioxide, nitrogen, etc.), and humidity—must be carefully selected and controlled.

7.9.2 Irradiation effects

The effects of radiation are usually assessed in the first place by visual examination, as many materials discolour, crack, disintegrate or melt, while flexible or soft materials may become hard and brittle.

More advanced work relies on the determination of changes in measured properties after certain periods of irradiation. Results of tests on unirradiated samples are compared with those on similar samples after subjection to known energies for different times. Non-destructive tests (loss tangent, permittivity, resistivity, changes in weight and changes in dimensions) are particularly useful because they can be repeated on the same specimen as a function of the total dose. Other tests used to determine the effects of irradiation are electric strength; tensile, shear and impact strength; elongation; hardness; flexibility; and water absorption.

In general, organic insulating materials deteriorate mechanically and electrically as the result of irradiation, the mechanical properties usually being impaired at a greater rate than electrical ones. These results are mainly due to chemical changes which occur in the materials consequent upon certain rearrangements of their molecular structure, usually with the evolution of one or more gases, such as methane, carbon monoxide, carbon dioxide or hydrogen. The probable life-dose for typical insulating materials is given in *Table 7.24*. The dose figure is that at which marked deterioration is observable, there is a 50% reduction in mechanical and electric strength, or some similar factor. The life-dose is approximate: it depends on the actual formulation of the material and irradiation conditions such as dose-rate and ambients.

7.9.2.1 Gases

The spark-gap breakdown voltage of air is reduced by about 20% under intense nuclear radiation. A gas having good radiation stability is sulphur hexafluoride (SF_6), whereas halogenated hydrocarbons slowly polymerise with evolution of corrosive products, and other gases, such as perfluoropropylene, polymerise rapidly to form liquids when irradiated.

7.9.2.2 Liquids

The viscosity of hydrocarbon oils increases and most liquid dielectrics polymerise when irradiated, the electrical properties being lowered considerably. Gases are usually evolved by liquids during irradiation and can create difficulties due to increase of pressure in containers of capacitors, transformers, etc. Silicone oils polymerise to form elastomers, those of high molecular weight yielding elastomers of low tensile strength, but the solids produced from the low molecular weight (e.g. 300–20 000) oils crumble on handling.

7.9.2.3 Semi-fluid and fusible materials

Silicone compounds made from fluids with mineral fillers react to irradiation in the same manner as the fluids; they rapidly harden and then gel. Petroleum greases are affected similarly. Dimethyl silicone fluids of low viscosity are to be preferred for use where doses do not exceed about 100 Mrad.

7.9.2.4 Organic solids

Most of the organic solid insulating materials in general use are synthetic resins or are based upon such resins and similar plastics. The performance of the materials in the form in which they are used under irradiation conditions depends largely upon the resistance of the basic materials to the effects of radiation. Cellulose derivatives, such as cellulose acetate, have poor resistance, mechanical deterioration being rapid; hence lacquers, adhesives and moulded or extruded parts made from them also deteriorate rapidly under irradiation. On the other hand, diphenyl silicone and products made from this resin, have relatively good resistance. The combination of mineral materials, whether as powdered or fibrous fillers or as sheet reinforcement (e.g. woven glass cloth), with the organic base materials usually results in products having superior radiation resistance.

7.9.2.5 Synthetic resins: thermoplastic

Of the thermoplastic synthetic resins *polystyrene* is one of the most stable. Styrene copolymers are generally poorer in resistance to effects of radiation, and high-impact-strength polystyrene loses some of its impact strength when irradiated. A good deal of work has been done on the effects of irradiating *polyethylene*, the principal change being the elimination of its melting point due to cross-linkage: this can be beneficial. Some increase in tensile strength occurs at first but, with continued irradiation at high dosages, the tensile strength decreases and the material ultimately becomes brittle and cheesy. The high-density varieties are slightly better. *Nylon* behaves somewhat similarly when irradiated, cross-linkage occurring with consequent increase in tensile strength in the case of sheet, but rapid reduction of strength occurs in nylon fibres, with decrease of elongation and impact strength.

Of the *vinyl polymers* and *copolymers*, polyvinyl carbazole has good resistance to radiation; *polyvinyl chloride* (PVC) has radiation resistance equivalent to that of polyethylene but liberates hydrogen chloride when irradiated; *vinyl chloride acetate* has a lower radiation resistance than PVC, softening and turning black at low dosages.

Irradiation of *polyethylene terephthalate* in fibrous form causes rapid loss in strength, and the films become brittle and darken. *Cellulose plastics* such as cellulose acetate deteriorate mechanically to a serious extent by low dosages, but electrical properties of the latter are not appreciably affected. *Acrylic resins* also have comparatively low radiation resistance.

The thermoplastic resins most seriously affected by radiation are the fluoroethylene polymers such as polytetrafluoroethylene and *monochlorotrifluoroethylene*, the latter being superior. Decrease of tensile strength and elongation of these is rapid and they become very brittle with only moderate dosages.

7.9.2.6 Synthetic resins: thermosetting

Of the amino resins, the *melamine formaldehyde* type is slightly superior to the *urea formaldehyde,* but with cellulose fillers both deteriorate rapidly and become brittle when irradiated. The radiation resistance of *epoxy* resins is above the average for plastics but it depends largely on the hardener used. *Epoxyphenolic* resins are better than ordinary epoxies and phenolics for radiation resistance. *Phenolic* and *polyester* resins without fillers have low resistance but this is increased considerably by the addition of fillers, especially minerals such as asbestos.

Silicone resins are more resistant to radiation than silicone fluids and elastomers and do not rapidly deteriorate physically, the major electrical properties of most resins being maintained even after subjection to high dosage.

7.9.2.7 Solid materials: inorganic

In general, inorganic materials are much more resistant than organic materials to irradiation. The principal effects are to cause colour changes and to induce conductivity in glasses, ceramics, fused silica and mica, followed by disintegration of natural mica irradiated by high dosages at elevated temperatures (e.g. 150°C). Synthetic mica seems to be more resistant than the natural form to mechanical and electrical degradation. Built-up mica products made from flake mica or reconstituted mica paper are affected according to the bonding medium—which is usually organic. Wires insulated with swaged magnesium oxide have shown high resistance to radiation at a temperature of 815°C.

7.10 Fundamentals of dielectric theory

7.10.1 Basic definitions

An ideal dielectric is a material or medium which has no free electrons so that no conduction can take place. It is an ideal insulator. Electrically it can be represented as a pure capacitor

Real dielectrics, however, have some free electrons but much less compared with conductors. Their equivalent circuit can be represented by an RC parallel circuit.

Dielectrics have two main electrical applications:

(a) *Insulation*: to isolate live conductors or conductors of different potentials.
(b) *Energy storage*: some dielectrics can store large amounts of energy when an electric field is applied to them.

7.10.2 Types of dielectrics

Dielectrics can have three forms: gas, liquid and solid.

In gases, the atoms/molecules do not interact but can contain free electrons which are responsible for the conduction process. Conduction mechanisms due to positive and negative ions as well as charged molecules are also found in some gases.

In liquids dielectrics, there is some interaction between molecules, and they are also known to contain impurities in real applications. This makes the conduction process essentially ionic in nature with the addition of contributions due to charged particles.

Solid dielectrics have more complex conduction mechanisms (such as thermal, tunnelling, hoping) governed by free electrons, holes and ions.

Solid dielectrics can be either:

(a) *Glasses:* are amorphous materials which have no three dimensional atomic ordering over distances greater than 2 nm (atoms are tenths of nm across).
(b) *Crystals*: these have long range ordering which could lead to a single crystal if the ordering is consistent throughout the solid material.

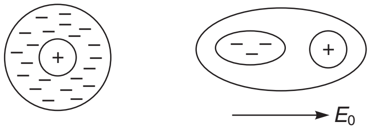


Figure 7.6 Electronic polarisation

In general, no solid or liquid is completely structureless. Ionic solids do not form glasses but covalent solids form glasses (e.g. polymers) e.g. MgO is an ionic solid (non glass) and SiO_2 is a covalent solid (glass).

7.10.3 Polarisation in dielectrics

When an insulating material is subjected to an electric field, a limited displacement of charge takes place at the atomic, molecular and bulk material levels. This charge displacement is known as polarisation.

(a) *Electronic polarisation*: in atoms, positive ions are surrounded by electron clouds. Since the electrons are very light, they respond rapidly to the action of an applied electric field. *Figure 7.6* shows a schematic of an atom being polarised due to the action of the field E_0.

(b) *Molecular polarisation*: within a molecule, the ionic bond is deformed when an electric field is applied resulting in increase of the dipole moment of the lattice:

Molecular polarisation can be:

- (i) *Simple ionic* where a simple separation of centre takes place.
- (ii) *Distorted ionic* takes place when large ions are distorted by other close ions in addition to the simple ionic polarisation.

(c) *Orientational polarisation*: in liquids and gases, whole molecules move into line with the acting electric field. Under weak static fields, the alignment is usually not complete. In solids, interfacial polarisation occurs at electrodes and at crystallites interfaces. In addition, in the presence of an interface with materials of different electrical properties (permittivity and conductivity), space charge polarisation occurs. *Figure 7.7* shows the various scenarios of orientational polarisation.

7.10.4 Quantification of dielectric polarisation

For a vacuum-filled parallel-plate capacitor, the surface charge density Q_0 is defined as $Q_0 = \varepsilon_0 V/d$ where V is the applied voltage, d the separation distance between plates, and ε_0 relative permittivity of free space ($\varepsilon_0 = 8.854$ pF/m).

If the vacuum insulating medium is replaced by a dielectric material of relative permittivity ε_r, the new surface charge density will be $Q_{ds} = \varepsilon_0 \varepsilon_r V/d$.

Thus, resulting in an increase of surface density $\Delta Q_s = Q_{ds} - Q_0 = (\varepsilon_r - 1)\varepsilon_0 V/d$.

This quantity is known as the dielectric polarisation P. In this case, the expression can be written as $P = (\varepsilon_r - 1)\varepsilon_0 V/d = (\varepsilon_r - 1)\varepsilon_0 E = D_d - D_0$.

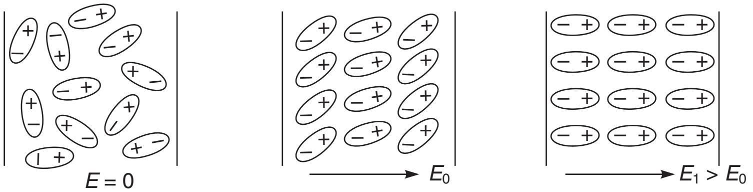
With D_d is the electric flux density in the dielectric case.
D_0 is the electric flux density in the vacuum case.

7.10.5 Properties of dielectric materials

Dielectrics can be grouped according to their structure and the way they react to the action of an electric field.

(a) *Non-polar dielectrics*: These consist of molecules that do not possess a permanent dipole moment, they are known as simple dielectrics. When an electric field is applied to simple dielectrics, it induces dipoles and orients them in the direction of the field.

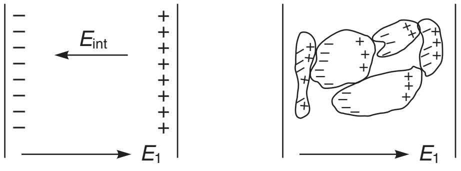
(b) *Polar dielectrics*: If a material has molecules with permanent electric dipole moments, in the absence of an external electric field, then it is called a polar dielectric e.g. Water and NaCl. Within these materials, the individual molecular dipoles are usually randomly oriented due to thermal agitation. The application of an external electric field will result in the alignment of the individual dipoles in the direction of the field. In general, polar dielectrics



a) No field: random distribution of molecules

b) Weak field: some alignment of molecules

c) Strong field: complete alignment



d) When alignment's complete we have interfacial polarisation at electrodes

e) In certain cases, interfacial polarisation occurs at crystallites

Figure 7.7 Orientational polarisation cases

Table 7.25 Examples of polar and non-polar dielectrics

Dielectric type	*Examples*	*Relative permittivity*
Non-polar	Transformer oil	2
Mildly polar	Chlorinated diphenyl	6
Strongly polar	Nitrobenzene Polypropylene carbonate Water Ethanol Hydrogen cyanide	≥20

have low values of relative permittivity and only small concentrations of permanent dipoles. *Table 7.25* gives examples of some polar and non-polar dielectrics.

(c) *Paraelectric dielectrics*: By contrast to simple dielectrics and polar dielectrics, paraelectrics contain a strong dipole in each unit cell with relative permittivity values more than 20 and up to 10 000.

(d) *Ferroelectric dielectrics*: By analogy to ferromagnetism, ferroelectrics contains domains (of about 1 μm) where permanent dipoles are oriented in the same direction. When an external electric field is applied to a ferroelectric material, the domains are aligned in the direction of the field. However, they are known to exhibit non-linear polarisation with applied field. Ferroelectrics have a relative permittivity, which increases with temperature until the 'Curie temperature' after which the permittivity decreases, and the domains cease to exist. The material becomes then paraelectric (see typical plot in *Figure 7.8*). Ferroelectrics exhibit hysteresis of polarisation as a function of the applied electric field (see typical curves on *Figure 7.9*). Therefore, the permittivity of the ferroelectric must be quoted for a given field value, *E*, a temperature *T*, and the 'history' of the material. *Table 7.26* gives examples of some ferroelectric materials.

(e) *Piezoelectricity*: When a ferroelectric material is subjected to an electric field at paraelectric temperatures (see *Figure 7.8* and *Table 7.26*) and then cooled, the domains in the material will align permanently. Mechanical shearing stress when applied to such materials can cause charge displacement, known as piezoelectricity. This property is also found in most crystals having an anisotropic structure e.g. quartz (simple dielectric which has no domains).

(f) *Electrets*: Any dielectric can contain electric charges at atomic or molecular level that can be oriented by an external electric field. If after cooling, the charges are immobilised, the dielectric will behave as an electrical equivalent of a permanent magnet. This is known as an electret (thermoelectrets in this case). An electret loses its volume charge exponentially but with a very large time constant (order of 100 years). Standard capacitors have time constants from few seconds up to six months (e.g. polystyrene).

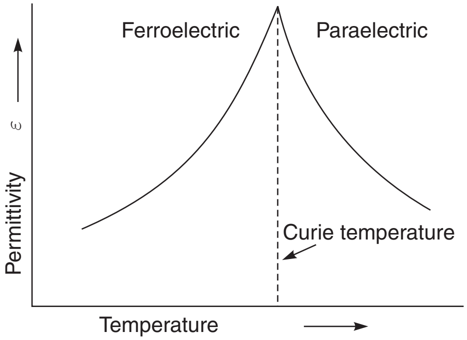
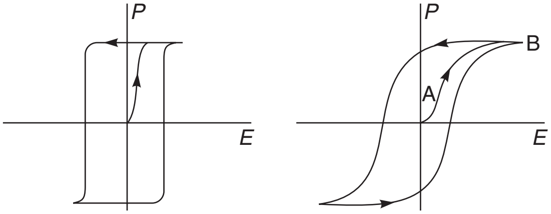


Figure 7.8 Typical dependence of permittivity on temperature in ferroelectrics



a) Single domain crystal
b) Polycrystal

Figure 7.9 Hysteresis in ferroelectric polarisation versus electric field

Table 7.26 Examples of ferroelectric materials

Ferroelectric material	*Symbol*	*Curie temperature* (°C)
Potassium dihydrogen phosphate	KH_2PO_4	−151
Lead titanate	$PbTiO_3$	487
Barium titanate	$BaTiO_3$	10 and 120
Cadmium titanate	$CdTiO_3$	−210
Lead niobate	$PbNb_2O_6$	570

7.10.6 Example of ferroelectric material: Barium titanate and its applications

Barium titanate is one of the most important ferroelectrics. It is formed from the reaction of a mixture of $BaCO_3$ and TiO_2 heated at 1250°C. The product is powdered and then worked by means of common ceramic techniques. Admixtures of other oxides are employed to modify the dependence of permittivity on temperature. Its high permittivity is exploited in ceramic capacitors for the range 500–10 000 pF for electronic equipment. Its piezoelectric property is used in transducer effect since it can be fabricated in a variety of complex shapes. The hysteresis in polarisation is used in timing control or carrier-frequency modulation of a voltage control.

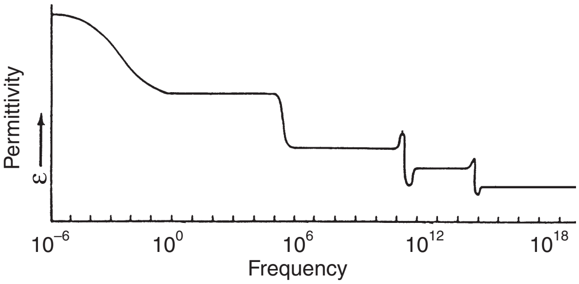


Figure 7.10 Typical relative permittivity, ε_r, dependence on frequency

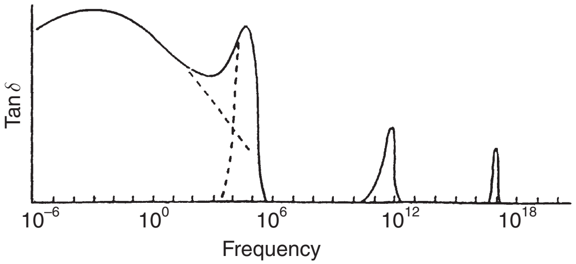


Figure 7.11 Typical loss tangent variation with frequency

7.10.7 Frequency response of dielectrics

Most materials are polarisable in several ways, producing complex frequency dependence. At the highest frequencies, only the electronic polarisation will 'keep up' with the applied field. At frequencies where permittivity varies rapidly, there is a peak in the dielectric loss. *Figures 7.10* and *7.11* show typical variations of permittivity and loss angle with frequency.

7.11 Polymeric insulation for high voltage outdoor applications

7.11.1 Materials

Traditional insulators are made of ceramic, either porcelain or glass. The design of these insulators for high voltage applications evolved from telegraph wire experience. These insulators are known to perform reliably in service for several decades. However, their bulky nature, poor pollution performance due to hydrophilic surfaces and susceptibility to vandalism raised a need for better materials.

Non-ceramic insulators (NCI), also commonly known as polymeric or composite insulators, are made of a compound of materials having one of several polymers as a base. Two main materials have been used extensively on outdoor high voltages power systems: silicone rubber and Ethylene propylene diene monomer (EPDM). Their combination (EPS) has been proposed to solve particular problems of surface properties and material strength. The main advantages of these materials are their light weight (since the insulator is made from a fibre rod with a polymeric outer sheath) and more importantly their surface hydrophobicity which inhibits the formation of continuous wet paths along the insulator surface. This is particularly advantageous under pollution condition when outdoor insulators suffer from dry-bands followed by localised discharges across the dry bands.

Room temperature vulcanised (RTV) coating is used to enhance porcelain insulators pollution performance by spraying/applying a thin layer of silicone rubber on the porcelain surface giving it a hydrophobic surface similar to that of silicone rubber insulators. RTV coatings contain typically 5% (by mass) cyclic LMW polydimethyl siloxanes.

IEC 1109:1992 Standard describes acceptance test procedures for composite insulators and IEC 507:1991 specifies artificial pollution tests procedures for high voltage insulators to be used on a.c. systems. However, these latter procedures are particularly suited for porcelain and glass insulators. At present, no standard procedure exists specifically for polymeric insulators. One difficulty is related to surface pollution because the hydrophobicity of polymeric insulators makes the conventional technique not suitable. Various methods have been suggested to obtain a uniform and repeatable pollution layer on polymeric surfaces. These include surface abrasion, kaolin spraying, multiple layer spraying and the use of small quantities of wetting agents, such as Triton X-100 as used in BS 5604:1986 (same as IEC 587:1984). International Standard IEC815:1986 gives guidance on the selection of insulators for high voltage applications taking into account the insulator shape and pollution severity.

7.11.2 Hydrophobicity loss and recovery

Extensive laboratory tests and field experience have shown that these materials can lose their hydrophobicity following a number of individual stresses or their combination. In particular, discharge activity on the insulator surface is found to reduce the hydrophobicity significantly. However, silicone rubber is found to recover its hydrophobicity after a certain time (which is still not well quantified) without the damaging stress. This process of recovery is thought to be due to the diffusion to the surface of low molecular weight (LMW) molecules and/or the changes in orientation of the methyl groups of the polymer.

7.11.3 Degradation ageing factors of polymeric insulator surfaces

Despite their excellent pollution performance, polymeric insulators are found to experience faster degradation on their surface compared with conventional insulators. Although such degradation depends on insulator material and design, a number of factors have been identified as affecting the surface properties of polymeric insulators. Some of these are described in the following processes.

7.11.3.1 Electrical process

The electric field distribution along a high voltage insulator unit is not uniform with the highest field regions located at the end terminals. In addition, the regions around the cores have higher electric fields than the shed areas. Under operating service conditions, the insulators are subjected to pollution that modifies the field distribution, reducing the magnitude at the end terminals. However, the wetting process is also non-uniform. The regions at the end terminals (especially at the top end) will wet more and at a faster rate than the regions under the sheds. Such combination of slow wetting and high field magnitudes encourages the initiation of discharges in these 'under-shed' core regions. With continuous intense discharging, the surface loses its hydrophobic properties. Furthermore, the discharging process generates ozone and nitrogen oxides which, when combined with water, produce nitrous and nitric acids. These acids attack the end fittings and make the polymer surface brittle forming a crazed pattern which can lead to splitting of the polymer sheath.

Surface currents on polluted insulators combined with the dry band discharging lead to tracking and erosion of the polymeric surface. Tracking and erosion resistance is improved by adding Alumina Trihydrate (ATH) as a filler to the polymer. After ageing, insulators with such materials exhibit a chalky white appearance caused by the diffusion of the ATH from the bulk to the surface.

7.11.3.2 Mechanical process

Direct mechanical stress on insulators can be tensile, compressive or cantilever loading. These stresses can lead to insulator failure by damaging the fibre reinforced plastic (FRP) core. Mechanical stresses usually affect the insulator in combination with other stresses.

Indirect mechanical stress is the form of surface tears caused by the release of stresses trapped during the manufacturing process. This provides arcing regions which would lead to tracking and erosion. Formation of fissures is also possible due to vibrations and the existence of material interfaces in the insulator unit.

Brittle fracture of the FRP rod has been observed under service conditions with insulator mechanical loads below the specified load. Various causes were suggested including faults from the manufacturing process, and water ingress combined with mechanical stress which transports hydrogen ions in solutions with a pH value of 3 or 4.

7.11.3.3 UV radiation

Outdoor insulators are exposed to sunlight, hence, ultraviolet (UV) radiation. UV radiation causes photo-oxidation and scission of molecular bonds in polymeric materials. Photo-oxidation is caused by ionisation of the surface molecules and attraction of the oxygen when the photon energy is sufficient. If the energy of the photon is higher than that of the bonds between the molecules or the polymer chains in the backbone of a single polymer, scission occurs.

Silicone rubber has a high resistance to damage by UV radiation because the siloxane (Si–O) bonds are of high energy. However, the hydrocarbon groups in the polymer can be damaged. UV resistance is usually increased by the use of carbon-based fillers which have the disadvantage of affecting the insulating properties of the material.

Intense UV exposure can lead to reduction of the silicone polymer on the surface of the material accompanied by an increase of the filler. Such surface depolymerisation was found to accelerate loss of hydrophobicity and increase of surface leakage current.

7.11.3.4 Chemical processes

Chemical attack occurs due to pollution products and following discharge activity on the insulator surface. Examination of field-aged insulators has found formation of uniform thin pollution layers on the surface. Sea or coastal pollution contains salts, while inland pollution comprises dust, industrial particles and agricultural fertilizers. When wetted, these pollution products react with the polymer under the action of the applied field. In tropical climates, micro-organism growth was found on insulator surfaces which was found to enhance the surface leakage current. The current increase warms the insulator surface, which further encourages the growth of the bacteria population. Partial arcs destroy the organisms but leave biological remains in the form of a slimy surface.

7.11.3.5 Water ingress processes

Water ingress in polymeric insulator occurs in three ways a) ingress through poor seals at end fittings, b) ingress through surface defects/damages, or c) through absorption of water into the polymeric material itself.

Corrosive chemicals and/or ionisable contaminants carried by the water affect the mechanical strength of the FRP rod and this is known to cause brittle fracture.

Water absorption causes depolymerisation as well as polarisation of the interfaces between the polymer and the fillers. More importantly, it increases the permittivity and loss tangent while it decreases the dielectric strength. As a consequence, heavy erosion and shed puncture can occur following moisture ingress.

8

Magnetic Materials

A G Clegg MSc, PhD
University of Sunderland
(Sections 8.1 and 8.7)

P Beckley BSc, PhD, DSc, CEng, FIEE, FIM, FInstP
European Electrical Steels, Newport
(Sections 8.2 and 8.3)

E C Snelling BSc(Eng), CEng, FIEE
Formerly of Philips Research Laboratories, Redhill
(Section 8.4)

R V Major MSc, PhD, FInstP, MIM, CEng
Formerly of Carpenter Technology (UK) Ltd, Crawley
(Sections 8.5 and 8.6)

Contents

8.1 Ferromagnetics 8/3

8.2 Electrical steels including silicon steels 8/3
  8.2.1 General 8/3
  8.2.2 Ancillary properties 8/4
  8.2.3 Chemistry and production 8/5
  8.2.4 Physical form 8/5
  8.2.5 Coatings 8/5
  8.2.6 Grain orientation 8/5
  8.2.7 Test methods 8/5
  8.2.8 Various 8/5

8.3 Soft irons and relay steels 8/5
  8.3.1 Composition 8/5
  8.3.2 Applications 8/6
  8.3.3 Physical forms 8/6
  8.3.4 Metallurgical state 8/6
  8.3.5 Magnetic characteristics 8/6
  8.3.6 Grades and specifications 8/6
  8.3.7 Heat treatment 8/7
  8.3.8 Ageing 8/7
  8.3.9 Test methods 8/7
  8.3.10 Other applications 8/7

8.4 Ferrites 8/7
  8.4.1 Magnetically soft ferrites 8/7
  8.4.2 Microwave ferrites 8/11

8.5 Nickel–iron alloys 8/11
  8.5.1 30% Nickel 8/11
  8.5.2 36% Nickel 8/11
  8.5.3 50% Nickel 8/11
  8.5.4 80% Nickel 8/11

8.6 Iron–cobalt alloys 8/13
  8.6.1 24/27% Cobalt iron 8/13
  8.6.2 50% Cobalt iron 8/13

8.7 Permanent magnet materials 8/13
  8.7.1 Alnico alloys 8/14
  8.7.2 Ferrite 8/14
  8.7.3 Rare earth cobalt 8/15
  8.7.4 Neodymium iron boron 8/16
  8.7.5 Bonded materials 8/16
  8.7.6 Other materials 8/16
  8.7.7 Properties, names and applications 8/16

8.1 Ferromagnetics

The choice of materials having ferromagnetic properties is very wide. Other than iron, a number of elements such as nickel and cobalt as well as some manganese alloys are ferromagnetic.

It has so far been impossible to obtain absolutely pure iron, in the true sense, even under laboratory conditions; but iron in which the impurities have been reduced to a mere trace has been found to have a maximum relative permeability of 50 000 and a very low hysteresis loss. However, even if it were possible to produce an iron of this degree of purity commercially, its low resistivity would be a disadvantage with alternating fluxes because of eddy currents.

The common impurities in iron are carbon, manganese, silicon, copper, sulphur, phosphorus and oxygen and the general effect of these impurities is to reduce the permeability and to increase the hysteresis loss. Sulphur, phosphorus and oxygen are particularly injurious and the presence of these has to be reduced to the lowest possible level. The presence of small amounts of silicon up to about 5% is beneficial. Small amounts of manganese are not injurious, but when the manganese content reaches 12%, a particularly non-magnetic steel is obtained. To neutralise the effect of impurities and to confer special properties, iron is alloyed with other elements, of which the following are the most important: nickel, cobalt, silicon, chromium, tungsten, molybdenum and vanadium.

The magnetic properties of ferromagnetic materials depend not only on their chemical composition, but also on the mechanical working and heat treatment they have undergone. For practical purposes magnetic materials fall into two main groups, high permeability or soft magnetic materials and permanent magnets or hard magnetic materials.

There are a number of groups of soft magnetic materials including soft iron, mild steel, silicon steels, nickel irons and ferrites. In static d.c. applications mild steel is the main bulk material, but where better properties are required with greater magnetic permeability, soft iron is used. Silicon steel is the bulk material for low frequency alternating fields and the nickel irons are used for more specialised applications with the soft ferrites being used at higher frequencies because of their higher resistivity.

There are four main classes of permanent magnet materials: steels, Alnicos, hard ferrites (also called ceramics) and rare earth alloys. Of these the steels are almost obsolete because of their inferior properties and the ferrites have gradually taken over from Alnico as the main bulk material, mainly due to their cost advantage. The rare earth alloys SmCo and NdFeB because of their superior properties have created new interest and have found a great many new applications. They have enabled considerable weight savings to be made in many instances.

8.2 Electrical steels including silicon steels

8.2.1 General

Electrical steels form a class of sheet material used for the flux-carrying cores of transformers, motors, generators, solenoids and other electromechanical devices. Almost always alternating magnetisation is employed and the material is designed to keep power losses due to eddy currents and magnetic hysteresis to a minimum.

8.2.1.1 Eddy current losses

Minimisation of eddy currents is achieved by using thin laminations. The e.m.f. generated in a lamination is proportional to the cross-section of the lamination for a given peak magnetic flux density and frequency; while the path length (and resistance) varies only slightly for a given lamination width, so that lamination of the core reduces the V/R ratio for the eddy currents (see *Figure 8.1*).

The current is proportional to (area of cross-section)/(path length).

For (i) in *Figure 8.1* this is $\frac{ab}{2b+2a} \approx \frac{ab}{2b} = \frac{a}{2}$

and for (ii) this is $\frac{2ab}{2b+4a} \approx \frac{2ab}{2b} = a$

Power rises as I^2R so that loss per unit volume rises rapidly as the sheet thickness increases. By increasing the resistivity of the steel, eddy currents may be further restrained and this is usually done by adding silicon as an alloying element. Silicon additions range from zero up to some 3.25%.

Over this range the resistivity may vary from about 12 to about 48–50 Ω-m $\times 10^{-8}$.

The addition of silicon has the effect of reducing the saturation flux density of the steel so that a larger core cross-sectional area is needed to carry a given magnetic flux.

8.2.1.2 Hysteresis losses

Magnetic hysteresis is another source of power loss which can be reduced by the addition of silicon as an alloying element. While the area of the very low frequency B–H loop is often taken as a measure of hysteresis loss, effects which impede magnetic domain wall motion alter with frequency. At 50 or 60 Hz the overall loss mechanism is a complex mixture of macro eddy currents relating to lamination thickness, micro eddy currents associated with the movement of the internal domain walls and energy dissipative hysteretical effects due to impediments to free domain wall motion. Such impediments may be non-metallic inclusions, grain boundaries or regions of stress within the crystal lattice of the material.

Table 8.1 gives an outline of the principal types of electrical steel produced in the UK, and some of their properties and applications.

As the percentage of added silicon increases, the steel becomes more difficult and expensive to process. Thinner material is more costly because it requires more rolling and processing. The key points of a material specification are given in *Table 8.1.*

Primary grading is by power loss in watts per kilogram at a peak magnetic flux density of 1.5 or 1.7 T and 50 Hz (60 Hz in the USA).

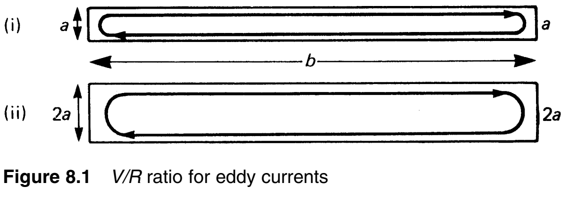


Figure 8.1 *V/R* ratio for eddy currents

Table 8.1 Selection of electrical steel grades produced by Cogent Power Ltd—typical properties

Grade identity	*Thickness* (mm)	*Typical specific total loss** (W/kg)	*Typical specific apparent power** (VA/kg)	*Nominal silicon content* (%)	*Resistivity* (Ω-m $\times 10^{-8}$)	*Stacking factor*	$\hat{B}$† (T)	*Typical applications*
Grain oriented: magnetic properties measured in direction of rolling only								
Unisil H								
M103-27P	0.30	(0.98)	1.40 (1.7, 50)	2.9	45	96.5	(1.93)	High efficiency power transformer
M111-30P	0.30	(1.12)	1.55 (1.7, 50)	2.9	45	96.5	(1.93)	
Unisil								
M120-23S	0.23	0.73	1.00	3.1	48	96	(1.85)	
M130-27S	0.27	0.79	1.10	3.1	48	96	(1.85)	
M140-30S	0.30	0.85	1.13	3.1	48	96.5	(1.85)	
M150-35S	0.35	0.98	1.24	3.1	48	97	(1.84)	
Non-oriented: magnetic properties measured on a sample comprising equal numbers of strips taken at 0° and at 90° to the direction of rolling								
Fully processed electrical steels								
M300-35A	0.35	2.62	23	2.9	50	98	1.65	Large rotating machines
M400-50A	0.50	3.60	19	2.4	44	98	1.69	Small transformers/chokes
M800-65A	0.65	6.50	14	1.3	29	98	1.73	Motors and fractional horse power (FHP) motors
Non-oriented: grades supplied in the 'semi-processed' condition and which require a decarburising anneal after cutting/punching to attain full magnetic properties								
Newcor M800-65D	0.65	6.00	8.6	Nil	17	97	1.74	Motors and FHP motors
Newcor M1000-65D	0.65	7.10	9.6	Nil	14	97	1.76	
Polycor M420-50D	0.50	3.9	6.5	Nil	22	97	1.74	
Tensile grades								
Tensiloy 250	1.60	—	—	Nil	—	—	1.60	Pole pieces large rotating machines

*At $\hat{B}=1.5\,\mathrm{T}$, 50 Hz. Values in parentheses are for $\hat{B}=1.7\,\mathrm{T}$ and 50 Hz.
† At $\hat{H}=5000$ A/m. Values in parentheses are for $\hat{H}=1000$ A/m.

8.2.2 Ancillary properties

8.2.2.1 Specific apparent power

Specific apparent power (in volt-amperes per kilogram) gives an indication of the r.m.s. ampere-turns which must be available to maintain a given state of alternating magnetisation, e.g. a peak value of 1.5 T at a specified frequency. Most of the current drawn is 90° out of phase with the supply voltage and so is non-dissipative, but copper or aluminium windings of sufficient cross-section must be used to keep copper losses low in the face of the current demanded.

8.2.2.2 Permeability

There are many varieties of permeability, but for electrical steels those of most interest are $\hat{B}_{1000}$, $\hat{B}_{2500}$, $\hat{B}_{5000}$, and $\hat{B}_{10\,000}$, that is the peak value of magnetic flux density attained under the influence of an applied field of 1000, 2500, 5000 and 10 000 A/m, respectively.

8.2.2.3 Surface insulation resistance

This is a measure of the quality of the surface insulating coating (expressed in ohm-metres squared) used to restrain the flow of current between laminations.

8.2.2.4 Stacking factor

This is the ratio of the height of a stack of laminations calculated from the volume of metal present to the measured height of the stack of laminations under a specified clamping pressure, for the same cross-sectional area of material. The stacking factor is expressed as a percentage and typical figures range from 96% to 98%.

8.2.2.5 Ductility

This is expressed as the number of 180° bends the material can sustain over a specified radius without cracking (typically >10 bends over a radius of 5 mm). Full details of specifications are given in: BSEN 10726, 10165, 10106 and 10107.

8.2.3 Chemistry and production

Electrical steels are produced by a complex process of casting, hot rolling, cold rolling and heat treatment. The end product must be very low in carbon and sulphur and free from non-metallic inclusions—that is metallurgically very 'clean'.

The magnetic properties are much improved if the grain size of the metal can be enlarged, up to a maximum of some 1 cm diameter for some grades and considerable metallurgical skill is employed to procure an end product that is chemically appropriate, metallurgically clean and of optimum grain size.

8.2.4 Physical form

Electrical steels are normally sold in coil form in coil weights up to several tonnes and widths up to 1 m wide, or in sheets cut to lengths of up to 3 m. Thickness normally ranges from 0.23 mm up to 0.65 mm (1.6 mm if high tensile grades are included). Strip can be supplied fully finished in a state of final anneal and carrying an insulating coating or semi-finished so that users can cut or punch the material and apply a final heat treatment and perhaps coating before building into a core. Electrical steels are susceptible to stress and mechanical damage so they must be handled with care and, if appropriate, re-annealed to recover properties damaged by stress. Appropriate heat treatments are normally an anneal at 800°C in a neutral atmosphere (non-oxidising) or a decarburising anneal at some 900°C in wet hydrogen.

When supplied as slit, narrow strip steel can be used to wind 'clockspring' cores directly without need for punching.

8.2.5 Coatings

The insulative coatings applied to electrical steels fall into two main classes: organic and inorganic. Organic coatings are inexpensive but cannot be used if further annealing is to be applied to the steel.

Inorganic coatings are resistant to annealing and some may be formed during the processing of the steel for the development of its magnetic properties.

Coatings having both organic and inorganic components can be used, so that the organic component which aids punching may disappear on anneal but a useful inorganic residual coating remains.

8.2.6 Grain orientation

The grains in electrical steel of the so-called 'non-oriented' type are, for the most part, randomly arranged so that the magnetic properties of sheet are broadly isotropic. This is convenient for many applications where flux must flow in varying directions in the steel.

However, if the grains are specially oriented so that an easy direction of magnetisation of the crystal lies along one direction of the sheet (normally the production rolling direction) then the magnetic properties in this direction are much enhanced at the expense of those at other angles to the rolling direction.

This, so-called, 'grain orientation' requires a costly and complex regime of rolling and heat treatments and the high-temperature anneal stage of production produces a glass film on the surface of the steel which acts not only as an insulating coating but also as a further means of enhancing magnetic properties. When steel bearing a glass coating cools from a high temperature the steel contracts more than the glass so that at room temperature the glass coating is holding the steel in a state of tension. This tension improves the magnetic properties of the steel, so that the material at its best functions as a composite system of a precisely grain oriented steel substrate held in a state of beneficial tension by a very thin (some 1–2 μm) layer of glass which also acts as an interlaminar insulant. A final phosphate coating is usually applied to complete the insulation and tensioning process.

8.2.7 Test methods

The test methods for electrical steels are fully described in BSEN 60404-2: 1998 and BSEN 10280: 2001 and in IEC standards 60404-2 and 60404-3. These tests seek to simulate the service conditions of electrical steels.

One of the best-known test methods is the Epstein test in which a set of strips of steel each 30.5 cm long and 3 cm wide are arranged in four limbs with double lapped corners to form a hollow square. Four double sets of windings enclose the strips and these are energised to produce a situation analogous to that in a small transformer at operating levels of magnetic flux density and under 'no load' conditions. Careful measurements of magnetic properties are carried out using a precision wattmeter and ancillary instrumentation.

There is a growing trend towards the use of single plates some 50 cm square for tests, and increasing use of such plate tests will reduce the labour involved in the lengthy Epstein test.

Cogent Power Ltd produces specialist test equipment for the measurement and quality control of electrical steel as do some instrument makers.

Steel producers offer a customer advisory service which in the case of Cogent Power Ltd is offered as well as the services of their Standards Laboratory which holds United Kingdom Accreditation Service (MKAS) approval for power loss testing at Newport in South Wales.

8.2.8 Various

Electrical steels cover a wide range of applications, some are made to have especially high strength for use where heavy mechanical loads arise. Some are of specially precise grain orientation where very low power loss and the highest permeability are vital. Some are simple but inexpensive for use in applications where cost and ease of manufacture of the final article may be paramount.

Electrical steels used inside nuclear reactors must behave well under conditions of high temperature and neutron flux.

Wherever alternating magnetic fields must be managed efficiently and economically, electrical steels of one variety or another can assist.

8.3 Soft irons and relay steels

8.3.1 Composition

The primary aim of soft irons is to be a close approximation to pure iron. Efforts are made to keep contaminant elements to a low level—particularly carbon and sulphur.

Carbon levels of 0.005% are advantageous, but expensive to produce, and material of intermediate quality can be obtained with carbon ranging from this low level up towards 0.05% and higher. At 0.1% carbon the material becomes a common 'mild steel' and while it can still be processed to have useful magnetic properties it is no longer a specialist material. Sulphur levels are kept as low as possible and <0.01% is a good aim.

8.3.2 Applications

Soft iron (or 'relay steel' as it is often called) is used largely for the magnetic circuits of electromechanical relays and solenoids, for the pole pieces of d.c. magnets and parts of the magnetic circuits of small generators. Quite a considerable amount of soft iron is used in the construction of particle accelerators for nuclear research where huge magnets are required to guide accelerated particle beams.

8.3.3 Physical forms

Soft irons come in four main physical forms:

(1) flat rolled, mainly in the thickness range 0.3–2.0 mm;
(2) round rod in the diameter range 2.0–20 mm;
(3) thick plate from 2.0–25 mm thick; and
(4) solid block up to 100 mm to 150 mm thick.

8.3.4 Metallurgical state

To achieve the best magnetic properties the material should be free from non-magnetic inclusions, have large grains and be free from internal stress.

8.3.5 Magnetic characteristics

The primary grading property of soft iron is coercive force—that is the magnitude of reverse field required to demagnetise the steel fully after a previous magnetisation to a specified value of magnetic flux density. This is shown as the distance *PQ* on the *H* axis of the *B–H* loop in *Figure 8.2*.

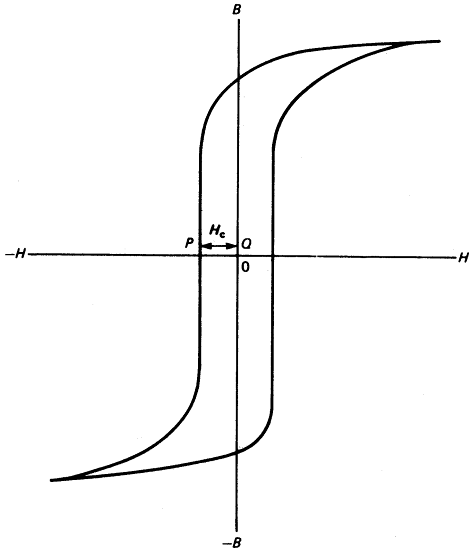


Figure 8.2

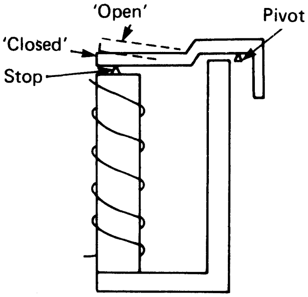


Figure 8.3 A simple electromechanical relay

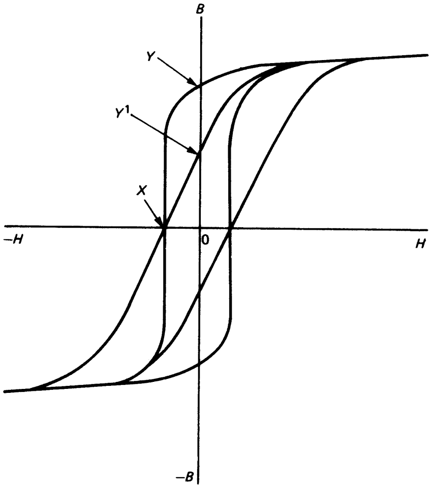


Figure 8.4 Shearing of the *B–H* loop

Coercive force is important as it indicates the amount of hysteresis which a material exhibits.

In an electromechanical relay (*Figure 8.3*) the magnetic circuit has quite a large air gap in it when the relay armature is in the relaxed 'open' state. When 'closed' the air gap is less but still of finite size (often set by a non-magnetic stop pip). It is well known that the magnetic reluctance of the air gap in a magnetic circuit produces 'shearing' of the *B–H* loop of the circuit material (see *Figure 8.4*). The coercive point *X* stays the same under loop shear, but the remanent point *Y* moves to Y'. It is apparent that material having a 'thin' loop with a small H_c will show a greater fall in remanence for a similar amount of 'shear' than will a material of high H_c and having a 'fat' loop. For this reason coercive force is used as a quality guide for material which ideally should have as low a hysteresis as possible.

Besides coercive force, permeability is important. Depending on the application the low-, mid- or high-field permeability may be important and strict specifications may be placed on the magnetic induction (tesla) achieved for a given applied magnetic field strength.

8.3.6 Grades and specifications

BS6404:8.10: 1994 is the UK Standard for grades of soft iron. Traditionally, a chemical specification has been used for the supply of relay steels and soft iron, but this can lead to wide variations of magnetic properties within a batch which meets the chemical specification.

Typically, steel is offered to a magnetic specification within five coercive force grades:

(1) $\lesssim 40$ A/m
(2) $\lesssim 60$ A/m
(3) $\lesssim 80$ A/m
(4) $\lesssim 120$ A/m
(5) $\lesssim 240$ A/m

(There is still much use made of the CGS unit, the oersted, (Oe) for soft iron grading). Material supplied to a magnetic guarantee is much more satisfactory to the relay designer and quality control engineer.

8.3.7 Heat treatment

It is important that iron should be in a fully annealed stress-free state in its final form in a magnetic circuit. Most material is punched, sheared, upset or machined in one way or another during fabrication, all of which induces high stress. An appropriate heat treatment of a piece—part will allow the material to give the best performance in service.

Optimum heat treatments are best worked out in conjunction with the steel supplier. Broadly, a 950°C anneal in a neutral atmosphere followed by a slow cool over several hours is satisfactory.

Where material is fairly thin (up to 1 or 2 mm) the final heat treatment can, with advantage be decarburising so that lowest coercive force is obtained as residual carbon is removed. Decarburising anneals employ furnace atmospheres of wet hydrogen and, again, require specialist advice.

8.3.8 Ageing

A final anneal takes residual carbon into solution and some of this may precipitate over a period of years leading to increases in coercive force as domain wall motion becomes impeded by carbon precipitates. Ageing may be minimised by keeping carbon content low and by use of special stabilising additions to the steel. When necessary, modern irons can be made ageing-free.

8.3.9 Test methods

British Standard BSEN 6404-4: 1997, now being amended, deals with test methods appropriate to relay steels. It covers the more lengthy procedures in which a welded ring is used as sample (or stamped rings from sheet), or the use of a d.c. permeameter with flux closure yokes. BS 6404: Part 7 (1986) gives details of the Vibrating Coil Magnetometer with which it is possible to obtain a coercive force reading in a few seconds. This is a very simple procedure compared with the more lengthy permeametric methods which require ballistic galvanometers, etc.

Cogent Power Ltd have pioneered the production of the Vibrating Coil Magnetometer for sale to industry in conjunction with the National Physical Laboratory.

IEC Standards 60404-4 and 60404-7 are equivalent to BS 6404 Part 4 and BS 6404 Part 7 respectively. Study of the UK and IEC Test Standards is of the greatest assistance in understanding how to grade soft irons for quality.

8.3.10 Other applications

Where relays must operate at the highest speeds it is necessary to raise the electrical resistivity of the magnetic circuit. This is necessary so that flux can penetrate the core iron rapidly. A low electrical resistivity gives rise to very vigorous induced eddy currents which oppose flux penetration to the centre of the material. The use of lamination is not often suitable so resistivity can best be raised by the addition of silicon, up to some 3% as required. This produces a fourfold increase in resistivity.

Thin iron is often used for screening. This may be applied to cables to reject outside interference or to load them inductively. Whole areas may be screened with iron enclosures where steady field or field-free conditions are essential. For this purpose screens consisting of layers of iron and air are the most effective.

8.4 Ferrites

In the context of electromagnetism, ferrites may be defined as magnetic materials consisting of compounds of metallic oxides and containing ferric ions as the main constituent. They usually have the structure of polycrystalline ceramic materials, but for some special applications the single crystal form is used. Ferrites are used in a wide variety of applications in electronic and communication engineering.

In the crystal lattice of ferrites the metal ions are separated by oxygen ions and this results in high electrical resistivities which suppress eddy currents and their usually undesirable effects. Another important consequence of the presence of oxygen ions is that the magnetic moments of the metal ions on the two constituent sublattices have anti-parallel alignment so that the net available magnetisation is the difference between the magnetisations of the sublattices (ferrimagnetism). This, together with the dilution due to the presence of the non-magnetic oxygen ions, inherently limits the saturation flux density of ferrite materials to about 0.4–0.6 T (compared to about 2 T for some magnetic alloys).

The usual manufacturing process consists of the following steps; mixing of raw materials (oxides, carbonates, etc.) in the required proportions, calcining at about 1000°C, crushing, milling, powder granulation, forming to the required shape by pressing the powder in a die or by extrusion, and finally sintering the piece parts at about 1250°C for up to 12 h in a controlled atmosphere. During sintering, crystallites of the required formulation are formed by solid-state reaction and this process is accompanied by a shrinkage of linear dimensions of between 10 and 25%. The product is a black brittle ceramic piece part having a density of about 4800 kg/m^3. Any subsequent shaping operations, such as pole face finishing, have to be done by grinding (and sometimes lapping).

Ferrites may be broadly divided into the magnetically soft category, those having high permeability and low losses, and the magnetically hard category, those having permanent-magnet properties. The hard ferrites are described later.

8.4.1 Magnetically soft ferrites

Magnetically soft ferrites have the cubic crystal structure of the mineral spinel. In ferrites, this structure is characterised by a very low magnetocrystalline anisotropy which results in the ferrite having a very low coercivity, high permeability and low magnetic power losses at frequencies extending up to 300 MHz, depending on composition. The chemical formulation is represented by $MeFe_2O_4$ where Me is most commonly a combination of manganese and zinc with some ferrous iron (abbreviated to MnZn ferrite), or nickel and zinc (NiZn ferrite). Minor constituents (additives or substitutions) are often included in order to enhance particular properties.

Ferrites are usually manufactured in the form of specific core shapes such as rings and a wide variety of pot-, E-, and U-shaped cores. Except for rings, the complete cores are usually assembled from pairs of half cores, the mating surfaces having been ground flat and sometimes lapped to ensure a low-reluctance joint. Where the application requires the core to contain an air gap in order to modify the properties of the wound assembly, one of the pole faces in the magnetic circuit is ground back by a specific amount during manufacture.

8.4.1.1 Properties

The properties of a ferrite material are strongly influenced by the composition, the presence of minor constituents and the details of the manufacturing process. For example, increasing the proportion of zinc lowers the Curie point and affects the saturation flux density. In MnZn ferrites the presence of a small proportion of the iron in divalent form critically controls the temperature coefficient of the permeability and influences the bulk resistivity and the magnetic losses. By varying these and other factors, a wide range of ferrite grades can be made, each specifically matched to a particular application. Typically a manufacturer's range of ferrite grades may extend to over fifty different specifications.

All ferrites exhibit a ferrimagnetic (or spin) resonance frequency at which the permeability falls to a very low value and the magnetic losses peak. For soft ferrites in general, the product of the low frequency initial permeability and the ferrimagnetic resonance frequency is approximately a constant (Snoek's law). So a high-permeability ferrite has a low cut-off frequency; where low losses are required at higher frequencies, ferrites having lower permeabilities must be used, see *Figure 8.5*

Generally, the MnZn ferrites have the higher permeabilities and lower losses in the frequency range up to about 2 MHz; their bulk resistivities lie between about 0.05 and 20 Ωm. The NiZn ferrites have lower permeabilities (dependent on the Ni/Zn ratio) and higher losses (relative to the MnZn ferrites) below about 2 MHz. However, they maintain their useful properties to much higher frequencies. Their resistivities are about three orders of magnitude higher than those for MnZn ferrites and so are virtually free of eddy current effects.

The initial permeability, μ_i, is the permeability at very low field strengths. High values of μ_i are desirable in principle, but Snoek's law dictates that the optimum initial permeability falls as the frequency characterising the application rises. If an air gap is inserted in the magnetic circuit of a core, the permeability apparent to a winding on that core is reduced to the effective permeability, μ_e. At the low flux densities appropriate to signal applications, the magnetic loss in ferrites is expressed in terms of a loss factor defined as the tangent of the loss angle divided by the initial permeability, $(\tan\delta)/\mu_i$. The loss tangent of a gapped core equals the loss factor multiplied by the effective permeability. In many applications the flux density is low enough for the hysteresis

Table 8.2 Typical properties of some magnetically soft ferrites (at 25°C unless otherwise stated)

	Conditions			*Ferrite classification*				
Parameter	*f* (kHz)	*B* (mT)	*Misc.*	*(i)*	*(ii)*	*(iii)*	*(iv)*	*Units*
Principal application				Low-frequency inductors	Wide band and pulse transformers	Power conversion	High-frequency inductors	
Basic composition (typical):								
$MnO/ZnO/Fe_2O_3$				27/20/53	25/22/53	34/14/52		mol%
$NiO/ZnO/Fe_2O_3$							32/18/50	
Initial permeability, μ_i	<10	<0.1		1200–3000	3800–18000	900–4000	70–150	
Residual loss factor, $(\tan\delta_r)/\mu_i$	30	<0.1		0.8–2.0	5–10	—	—	10^{-6}
	10^2	"		1.0–3.0	10–50	—	—	"
	10^3	"		—	—	—	20–40	"
	10^4	"		—	—	—	60–100	"
Temperature factor, $\Delta\mu_i/(\mu_i^2 \cdot \Delta\theta)$	<10	<0.1	5–55°C	0.5–2.0	—	—	0.8	$°C^{-1} \times 10^{-6}$
Hysteresis coefficient, η_B	100	1.5–3.0		0.5–1.0	0.3–1.0	—	6–40	$mT^{-1} \times 10^{-6}$
Saturation flux density, $\hat{B}_{sat}$			$H = 250$ A/m	0.35–0.45	0.35–0.45	0.4–0.5	0.3–0.4	T
Curie point, θ_c	<10	<0.1		130–210	100–200	180–240	250–400	°C
Resistivity, $\rho\varsigma$	d.c.			1–5	0.05–0.5	1–20	$>10^4$	Ωm

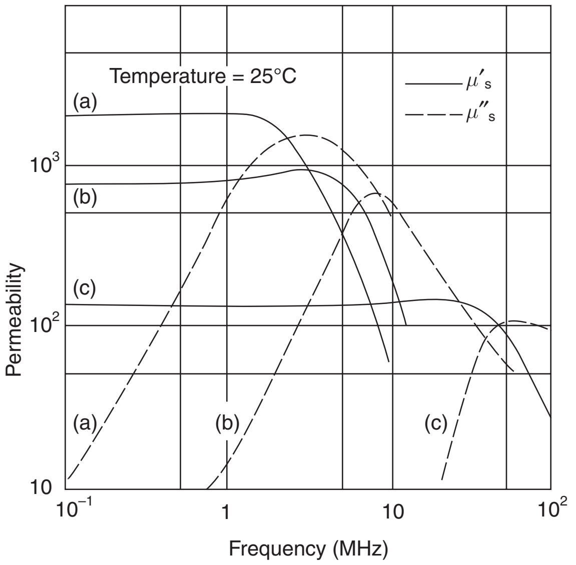


Figure 8.5 The inductive component of the permeability, μ'_s, and the corresponding loss component, μ''_s, as functions of frequency for three typical ferrites. (a) is an MnZn ferrite having an initial permeability of 2000 and intended for high performance inductors at frequencies up to about 0.3 MHz, (b) is an MnZn ferrite having a permeability of about 700 and is intended for inductors operating at frequencies of 0.3–0.7 MHz and (c) is an NiZn ferrite with a permeability of 130, applicable to inductors intended for the low MHz band

loss to be negligible. The main loss is then represented by the residual loss. This is low over the lower frequency region but rises rapidly as the frequency of ferrimagnetic resonance is approached. *Figure 8.6* shows the residual loss factor as a function of frequency for three typical ferrites: (i) is a low loss MnZn ferrite, (ii) is a high permeability MnZn ferrite intended for signal transformers, and (iv) is an NiZn ferrite with a permeability of about 130. Other properties of these typical ferrites are listed in *Table 8.2*

The temperature dependence of the initial permeability is important for inductor applications. As a material property it is usually expressed as a temperature factor $\Delta\mu_i/(\mu_i^2\Delta\theta)$, where θ is the temperature. If a core is gapped so that it has an effective permeability, μ_e, then the temperature coefficient of effective permeability (and therefore of the inductance of a winding on the gapped core) is μ_e times the temperature factor. So the temperature coefficient of inductance can be determined by the choice of the gap length. For an inductor used in filter applications, the negative temperature coefficient of the tuning capacitor can thus be compensated to achieve very high combined temperature stability.

In applications involving MnZn ferrites, the finite conductivity of the material gives rise to some eddy-current loss in the core. This loss depends on the size and shape of the core and, although small in relation to the total core loss at low frequencies, it can become significant at frequencies above about 50 kHz, depending on the ferrite resistivity and the size of the core. The low-amplitude hysteresis loss in a ferrite is expressed in terms of a coefficient, η_B, such that the hysteresis loss factor $(\tan\delta_h)/\mu_i = \eta_B\hat{B}$, where $\hat{B}$ is the peak flux density. The hysteresis coefficient of a ferrite determines the Total Harmonic Distortion (THD) generated by a core made in that material, e.g. for a pulse transformer.

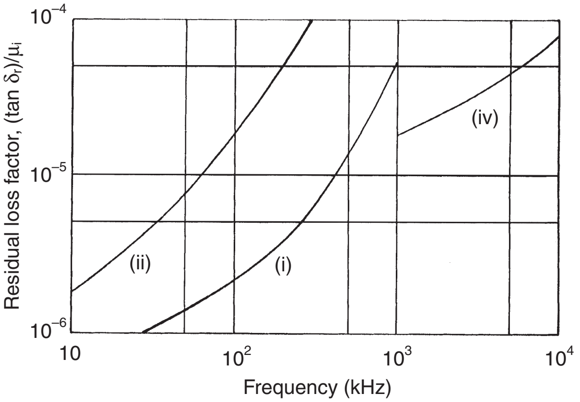


Figure 8.6 Typical residual loss factors as functions of frequency. (For material classifications see *Table 8.2*)

In modern digital networks THD is a critical parameter in that it can cause bit errors. Special MnZn ferrites with high permeability and very low hysteresis loss have extremely low THD levels and are to be recommended for such applications.

In power applications the important material property is the power loss per unit volume, *i.e.* the power loss (volume) density. The principal component of this is the magnetic loss which is defined as the loss due to all causes other than the eddy currents generated in the bulk of the core. The magnetic loss (volume) density $P_M = kf^m\hat{B}^n$, where k is the loss coefficient and n is the Steinmetz exponent. The values of k, m and n depend on the composition and microstructure of the ferrite and are temperature dependent. For MnZn ferrites intended for power applications, n usually lies between 2 and 3 (typically 2.5). The exponent m is unity at low frequencies, rises to typically 1.5 at 100 kHz and continues to rise as the ferrimagnetic resonance frequency is approached. These data apply both to sinusoidal and symmetrical square-wave excitation (in the latter case the flux waveform in the core is approximately triangular).

As the frequency of operation increases, the flux density must be reduced to keep the power loss within the limits imposed by thermal considerations. For MnZn ferrites operating at frequencies approaching 1–2 MHz, the rising value of m in practice imposes a cut-off frequency for the power application of these materials. To extend the range of MnZn ferrites to the highest possible frequencies, lower permeability grades have been developed. Lowering the permeability raises the frequency of the ferrimagnetic resonance and this, in effect, lowers the value of the exponent m. *Figure 8.7* shows the loss curves for two typical high-performance power ferrites and illustrates the improvement in the higher frequency performance that can be achieved by such means. The ferrite composition is usually chosen so that the minimum in the power-loss/temperature curve occurs at the typical operating temperature of the core, for example 85°C. To evaluate the total loss density it is necessary to add the eddy-current loss density to the magnetic loss density. The eddy current loss density, which is specific to a given core, may be expressed approximately by $P_F = (\pi f^2\hat{B}^2A_e)/(4\rho)$, where A_e is the effective cross-sectional area of the core and ρ is the bulk resistivity of the ferrite at the operating temperature. At low frequencies, the eddy-current loss can be relatively small, but at frequencies above about 50 kHz it becomes an increasingly significant fraction of the total.

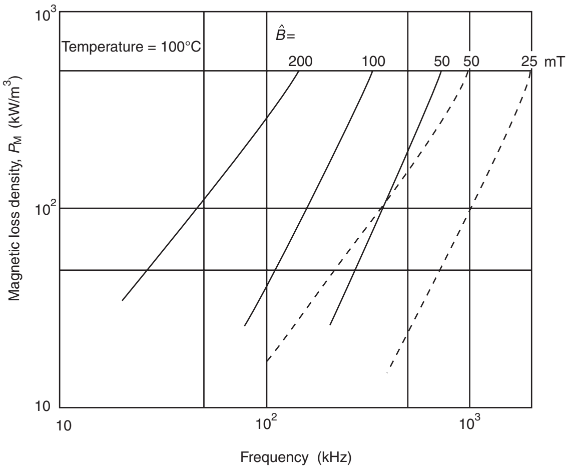


Figure 8.7 Magnetic loss (volume) density, P_M, for two MnZn ferrites intended for power conversion applications. The solid curves are typical of power ferrites for switching frequencies in the range 50–500 kHz. The broken curves show the properties of lower permeability power ferrites designed for the frequency range 0.5–2 MHz. Courtesy of Drs. J. W. Waanders, *Philips Components*

For a conventional transformer core, the temperature rise normally limits the total core loss density to about 100–400 kW/m^3, the larger figure relating to the smaller cores. The maximum allowable core loss density will set a limit on the flux density for a given frequency and this will enable the minimum number of turns required on a given core to be evaluated. The actual core loss is obtained by multiplying the total core loss density by the effective volume of the core as quoted in the data sheet for the core. In addition to providing loss data on their grades of power ferrite materials, some manufacturers quote the core loss in watts at given frequencies and flux densities for specific cores in their product range.

In the development of electronic power supplies, the priority is size reduction which leads to increasing throughput power densities. To accommodate this trend, switching frequencies are increased, the losses in the ferrites are reduced by material development and the power loss densities in the switching transformers are increased. To achieve higher loss densities in the transformers without exceeding a safe operating temperature, the trend is to use flat or planar geometries that can dissipate heat more easily. Such flat designs, which may be only a few millimetres high, are more compatible with pcb technology, can use windings based on printed circuit practice and can allow power loss densities as high as 700 kW/m^3.

For the lower frequency power applications where the design is saturation rather than loss limited, it is important that the saturation flux density of the ferrite should be as high as possible at the operating temperature. This allows a high value of peak flux density to be used and minimises the required number of turns. Except in saturable reactor applications, a flux density excursion into the saturation region must usually be avoided, even under the worst overvoltage conditions, since it causes a steep rise in magnetising current that can damage the associated semiconductors.

8.4.1.2 Applications

Magnetically soft ferrites are used extensively as cores for inductors and transformers in industrial and consumer electronics. Referring to *Table 8.2*, the data in the first three classifications indicate typical properties of MnZn ferrites. The first is a low-loss ferrite intended for cores in high quality inductors used at frequencies up to about 300 kHz, such as inductors in filters for telecommunication applications. These cores are usually made in the form of cylindrical or square pot cores with an air gap to provide a specific effective permeability. There is usually provision for adjusting the inductance over a range of about ±10%. Such components combine potentially high *Q*-factors (300–1000) and good stability with temperature and time. The evolution from the early analogue to the present digital electronics has greatly reduced the proportion of ferrite cores required for such inductor applications.

The second column in *Table 8.2* is representative of ferrite grades having high initial permeabilities and intended for cores used in low-power wide-band and pulse transformers. They are usually made in the form of rings and a wide range of E-core and pot-core shapes, some having low profiles compatible with pcb technology. The demand for wideband and pulse transformers increased dramatically with the introduction of digital networks such as Internet, and this in turn promoted the development of special high permeability ferrites having very low third harmonic distortion levels.

The third column refers to ferrites developed specifically for power applications. Initially the main application was the line scan (and high voltage) transformer in the domestic television receiver. The enormous growth in integrated circuit electronics led to a corresponding growth in the need for high frequency switched-mode power supplies This stimulated the development of ferrite materials and cores specifically for this purpose. Cores for use in power conversion are now second only in production volume to ferrite yokes used in cathode ray tube deflection systems. Originally switching frequencies were between 25 and 50 kHz, but the trend towards more compact power supplies and the attendant trend towards the use of resonant power conversion has progressively pushed these frequencies up towards 1 MHz and beyond. *Figure 8.7* shows the magnetic loss data for two typical power ferrites, illustrating the development of improved performance at higher frequencies. When switching frequencies exceed about 3 MHz, even the best MnZn ferrites are unsuitable due to their increasing eddy current and ferrimagnetic resonance losses. NiZn ferrites, at the higher permeability end of their composition range, then offer superior performance.

Cores developed for power applications are generally variations of the E-core or U-core shape and many different designs are available. With the higher frequency applications, power loss densities become very high and this makes flat (low profile) geometries preferable, as described in Section 8.4.1.1.

The largest use of ferrites, by weight, is for yokes in cathode ray tube deflection systems. Most line deflection frequencies are relatively low so magnetic loss has not been a significant parameter in the design. However, in response to the need for better image definition, the line frequencies in domestic television receivers and, more particularly, in computer monitors are increasing; 100 kHz and even higher is quite common. Power losses are therefore becoming important in this application.

In addition to the uses outlined above, magnetically soft ferrites are employed in a variety of other applications, e.g. as rods for radio receiver antennas and tuning slugs in radio frequency coils. Ferrites with relatively square hysteresis loops are used for magnetic amplifiers in power control. With the increased use of electronic equipment and the stricter EMC regulations, electromagnetic interference

(EMI) suppression has become an important application area for soft ferrites. Special ferrites have been developed to meet the requirements. In particular, ferrites in which the permeability and magnetic losses combine in such a way that the core presents an impedance that remains high over a broad frequency band. Above 1 GHz, ferrites having hexagonal crystal structures, such as those used for permanent magnets, are being introduced into the EMI suppression application. Traditionally, massive ferrite toroids have been and still are used as coupling cores in large particle accelerators. To conclude this very brief overview of the great range of ferrite applications, two very contrasting products can be included. High permeability ferrites, often in single-crystal form, are used in large quantities for magnetic recording heads in all types of analogue and digital recording equipment, and ferrites having the property of absorbing EM waves are produced as tiles, powders and granules for coating the surfaces of rooms and buildings to attenuate the reflection or transmission of EM energy at frequencies from hundreds of MHz to tens of GHz.

8.4.2 Microwave ferrites

Ferrites having special formulations are used in the non-reciprocal components (or gyrators) in microwave systems. This application of ferrites at microwave frequencies depends on the fact that the axis of the spinning electron (the magnetic element of the crystal lattice) will, if disturbed, precess with a rotational direction dependent only on the direction of the static magnetic field aligning the spins, and at a frequency that depends only on the static field strength. An electromagnetic wave, having positive circular polarisation relative to the static field direction and a frequency equal to the precession frequency, will couple to the spin precession to induce a resonance. There is no corresponding coupling or resonance for a negatively polarised wave. So a ferrite that is saturated by a static magnetic field and thus constitutes an array of aligned spins, will have properties that depend on the sense of the polarisation of an incident electromagnetic wave. Devices based on such material conditions will exhibit discrimination as to the directional properties of such a wave. This phenomenon gives rise to a family of waveguide and stripline devices used in microwave systems, such as unidirectional attenuators (isolators) and energy-routing components (e.g. circulators).

To meet the diversity of requirements found in microwave applications, many different magnetic oxides have been developed. Some are based on spinel (cubic) ferrites such as nickel ferrite, MnMg ferrite and NiCuCoMnAl ferrite. Magnetic oxides based on the garnet structure are extensively used in many applications, examples are yttrium iron garnet and gadolinium iron garnet each having a variety of substitutions to obtain the required properties.

8.5 Nickel–iron alloys

The nickel–iron alloys, ranging in composition from 30 to 80% nickel are the most versatile of soft magnetic materials. The design engineer can select from a wide range of available properties such as saturation flux density, permeability, coercive force and loss. The alloys can be subdivided into four main groups. Properties of a typical range of alloys are given in *Table 8.3*.

8.5.1 30% Nickel

The 30% nickel–iron alloy has a Curie temperature of the order of 60°C and exhibits a large linear change of permeability with temperature over the range −30°C to +40°C. This alloy property can be utilised in the form of a shunt to provide compensation for the change in the characteristics of a material, such as magnetic performance or conductivity, with temperature. Typical examples of such usage are watt-hour meters, speedometers and microwave circulators. The 31% and 32% nickel–iron alloys have Curie temperatures of 100°C and 150°C respectively and were developed for use in devices operating at higher temperatures.

8.5.2 36% Nickel

The alloy has a lower saturation flux density (1.2 T) and permeability than the 50% alloy but its lower cost and high resistivity make it an attractive alternative. The main applications are relays, high-frequency transformers and inductors. The alloy is also employed in devices, such as inductive displacement transducers, where its very low coefficient of expansion, 1×10^{-6}/°C, is a great advantage.

8.5.3 50% Nickel

This alloy has the maximum saturation in the nickel–iron range (1.6 T) and is used where a higher permeability and/or a better corrosion resistance than that of silicon–iron is required. Another advantage of the alloy is its high incremental permeability over a wide range of polarising d.c. fields. Applications include chokes, relays, small motors and synchros.

Special processing techniques enable a wide range of properties to be developed in this alloy. A cold reduction, in excess of 99%, produces a cube texture in the annealed strip and a square hysteresis loop. Applications utilising this property are magnetic amplifiers, inverters and pulse transformers. Annealing in a magnetic field, below the Curie temperature, can also alter the magnetic characteristics. If the field is applied in the conventional manner, both the initial and maximum permeabilities are substantially increased. When the field is applied in a transverse direction a very flat hysteresis loop with a low remanence is produced. These properties are ideal for thyristor firing circuits where a unipolar pulse is required and the transformer only operates between remanence and saturation.

8.5.4 80% Nickel

The high permeability and low coercivity of this group of alloys is due to the fact that both the magnetostriction and crystalline anisotropy are essentially simultaneously zero at this composition. Until the discovery of the cobalt–boron metal glasses and the nanocrystalline alloys with a similar combination of properties, the 80% nickel–iron alloys were unique for their ultra-high permeability performance.

The commercial alloys have additions of small percentages of molybdenum, copper or chromium to give improved magnetic performance over the binary alloy. Applications include sensitive relays, pulse and wide-band transformers, current transformers, current balance transformers for sensitive earth leakage circuit breakers and magnetic recording heads. Magnetic shielding is also a major area of application either in the form of fabricated

Table 8.3 Nickel–iron based alloys

	30%Ni *R 2799**	*31% Ni* *R 3100**	*32% Ni* *R 2800**	*36% Ni*[4] Radiometal 36*	*45% Ni*[3] Radiometal 4550*	*50% Ni*[3] Super Radiometal	*50% Ni*[3] HCR*	*54% Ni*[2] Satmumetal*	*77% Ni*[1] Mumetal*	*77% Ni* *Super Mumetal 250**
Initial permeability $\times 10^{-3}$	—	—	—	3	6	11	0.5	50	60	250
Maximum permeability $\times 10^{-3}$	—	—	—	20	40	100	100	120	240	400
Saturation flux density (T)	0.2	0.4	0.7	1.2	1.6	1.6	1.54	1.5	0.8	0.8
Coercive force (A/m)	—	—	—	10	10	3	10	2.5	1.0	0.5
Remanence (T)	—	—	—	0.5	1.0	1.1	1.5	0.7	0.45	0.5
Resistivity (μΩ-m)	0.85	0.84	0.83	0.8	0.45	0.4	0.4	0.45	0.6	0.6
Density (kg/m^3)	8000	8000	8000	8100	8300	8300	8300	8300	8800	8800
Curie temperature (°C)	60	100	150	280	530	530	530	550	350	350
Expansion Coefficient (°C $\times 10^{6}$)	10	7	5	1	8	10	10	11	13	13

*Carpenter Technology (UK) limited.

[1,2,3,4] BS 6064 Section 8.6, 1988. Classes E1, E2, E3, E4, respectively.

cans or screens for transformers and cathode ray tubes. Annealed tape is also used in the manufacture of shielded cables.

8.6 Iron–cobalt alloys

The addition of cobalt to iron results in an increase in saturation, up to a maximum flux density of 2.45 T at 35% cobalt. The high cost of cobalt, relative to that of iron, restricts the widespread use of these alloys, although they are widely employed in the aircraft and associated industries. Their high Curie temperature and magnetostriction are also of special importance. Properties of a typical range of alloys are given in *Table 8.4*.

8.6.1 24/27% Cobalt iron

The ductility drops very considerably when the addition of cobalt exceeds 27% and compositions in this area are chosen for applications such as magnet pole tips where a combination of good ductility, ease of machining and high magnetic saturation flux density is required. The permeability, coercive force and loss characteristics are, however, inferior to the 50% cobalt alloy.

8.6.2 50% Cobalt iron

The optimum combination of magnetic properties is obtained in the 50/50 cobalt–iron composition. The ductility of the binary alloy is so low that it is not possible to fabricate the alloy to any extent. The addition of 2% vanadium greatly increases the ductility and fortunately, from the eddy current loss viewpoint, also substantially increases the resistivity. The 49/49/2 Fe/Co/V alloy has been extensively used for stators and rotors in lightweight electrical generators for many years. Other applications include lightweight/small-volume transformers, special relays, diaphragms, loudspeakers and magnet pole tips.

With Curie temperatures in excess of 900°C, this group of alloys is often the only choice for designers of equipment which incorporates soft magnetic materials that are required to operate at exceptionally high temperatures. A typical example of such an application is in the construction of the magnetic pumps used to pump liquid metal in the production of automobile engine castings.

The demand for increased electrical power in civil aircraft has forced designers to consider changing from the conventional constant frequency type of electrical generator to variable frequency or variable speed, constant frequency, machines. Higher centrifugal forces are exerted on the rotors of these machines, which operate at greatly increased speed. High strength cobalt–iron alloys have therefore been developed to meet these and other applications, such as magnetic bearings.

8.7 Permanent magnet materials

The properties of a permanent magnet material are given by the demagnetisation curve (*Figure 8.8*) the second quadrant of the *B*–*H* hysteresis loop. This extends from the remanence B_r to the coercivity H_{cB}. It can be shown that when a piece of permanent magnet material is put into a magnetic circuit, the magnetic field generated in a gap in the circuit is proportional to BHV, where B and H are the corresponding points at a point on the demagnetisation curve and V is the volume of permanent magnet material. This means that, to obtain a given field with the minimum volume of magnet material, we require the product BH to be a maximum. The magnet is then designed so that its BH value is as close as possible to the $(BH)_{max}$ value. It is also useful to use the $(BH)_{max}$ value, which represents the useful available energy, to compare the characteristics of materials. Generally, the material with the highest $(BH)_{max}$ will be chosen but this has to be weighed against such considerations as cost, shape, manufacturing problems and stability.

When the magnet is fully magnetised in a completed magnetic circuit with no gap, the working point is the remanence B_r. When a gap is made in the circuit the magnet will be partly demagnetised and its working point will fall to, say the point P. If now a further demagnetising field is applied to the magnet corresponding to H_1H_2 the flux density will be reduced to B_2. If this extra field H_1H_2 is removed the field will recoil along the narrow loop QR to the point R. The corresponding flux density is B_3. This narrow loop is called the recoil loop and its slope is $\mu\varsigma\Leftarrow\not{\mu}_0\mu_r$ where μ_r is the

Table 8.4 Iron–cobalt alloys

Alloy	*Saturation flux density* (T)	*Initial permeability*	*Maximum Permeability*	*Coercive force* (A/m)	*Remanence* (T)	*Resistivity* (μΩm)	*Curie temperature* (°C)	*Yield strength* (Mpa)
24% Co (24 Permendur)* (Hiperco 27)*	2.34	250	2000	150	1.5	0.2	950	280
49/49/2 Co/Fe/V (49 Permendur)* (Hiperco 50)*	2.34	800	7000	65	1.6	0.4	940	340
49/49/2 Co/Fe/V + Nb/Ta (Rotelloy 8)* (Hiperco 50 H.S.)*	2.34	300	4000	180	1.8	0.4	940	600

*Carpenter Technology (UK) Limited.

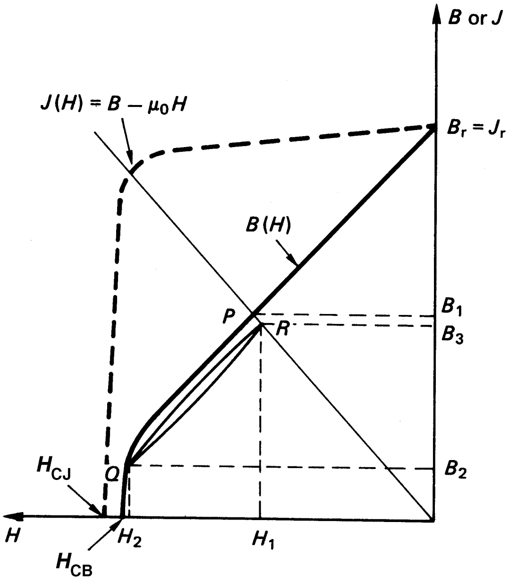


Figure 8.8 Demagnetisation curve and recoil loop

relative recoil permeability and μ_o is the magnetic constant $(=\cancel{4\pi\times10}^{-7})$. This is an important consideration in dynamic applications such as motors, generators and lifting devices where the working point of the magnet changes as the magnetic circuit configuration changes. It is also important when considering premagnetising before assembly into a magnetic circuit. If μ_r is nearly equal to unity, as it is for ferrites and for rare earth alloys, then the magnets can be premagnetised before assembly into a magnetic circuit without much loss of available flux. However, if μ_r is considerably greater than unity, care must be taken to magnetise the magnet in the assembled magnetic circuit.

The dotted curve in *Figure 8.8* is the *J–H* curve where $B=\cancel{\mu_0}H+J$. For materials with B_r much greater than μ_oH_{cB}, e.g. Alnico, the two curves are almost identical, but for ferrites and rare-earth alloys where B_r and μ_oH_{cB} are close in numerical value the curves become quite different. The J–H curve has only a gentle slope from B_r, this indicates that the magnetisation in the material is nearly uniform. A new parameter the intrinsic coercivity H_{cJ} is introduced. This intrinsic coercivity H_{cJ} is a measure of the difficulty or ease of demagnetisation. As the value of H_{cJ} is increased the more will the material resist demagnetisation due to stray fields, etc. However, as a field of at least three times H_{cJ} must be applied to the magnet for magnetisation, difficulties may be encountered in attaining the full magnetisation if H_{cJ} is too high.

Magnets in the three main groups Alnico, ferrites and rare-earth alloys are generally made anisotropic with the properties in one direction considerably better than in the other directions. This best direction is called the preferred direction of magnetisation. The curves and parameters applied to such a material refer to the properties in this preferred direction. To choose between these three materials for a particular application, Alnico must be used where magnetic stability is a requirement and this will apply in any type of instrument application. Ferrite will be used where cost is the main consideration and rare-earth magnets will be used if the highest possible strength is required or if miniaturisation is needed

8.7.1 Alnico alloys

A wide range of alloys with magnetically useful properties is based on the Al–Ni–Co–Fe system. They are characterised by high remanence and available energy and moderately high coercivity. They are very stable against vibration and have the widest useful temperature range (up to over 500°C) of any permanent magnet material. They are however mechanically hard, impossible to forge and difficult to machine except by grinding and special methods such as spark erosion.

The commonest alloys contain 23–25% Co, 12–14% Ni, about 8% Al, a few per cent of Cu and sometimes small additions of Nb, Ti and Si, with the balance being iron. They are cooled at a controlled rate in a magnetic field applied in the direction in which the magnets are to be magnetised. The properties are much improved in this direction at the expense of those in the other directions. There is finally a fairly prolonged and sometimes rather complicated heat treatment at temperatures in the range 550–650°C. With the aid of the electron microscope it has been shown that after this treatment the magnets have a two-phase structure consisting of fine elongated magnetic particles separated by a non-magnetic phase, this structure being oriented by the magnetic field applied during cooling.

The field cooling is more effective in producing anisotropic properties if the magnets are cast with a columnar structure. Casting the alloy into a very hot mould placed on a cold steel plate produces this columnar structure. The moulds are usually made from exothermic material and the molten alloy is poured into this mould. This is a rather expensive process and there are limitations on the lengths and shapes of magnets that can be made into this columnar form.

The coercivity of Alnico can be improved by a factor of 2 to 3 by increasing the cobalt content to 30–40% and adding 5–8% Ti with possibly some Nb. For the best properties it is necessary to hold these alloys for several minutes in a magnetic field at a constant temperature, accurately controlled to ±10°C, instead of the usual field cooling process. This meticulous heat treatment as well as the high cobalt content makes these alloys expensive and they are only used where the higher coercivity is required. Columnar versions of these alloys can also be made but they are very brittle and are only used for very specialised applications.

The Alnico alloys can be produced as castings of up to 100 kg and down to a few grams in weight but it is found more economical to produce the smaller sizes of 50 g and less by the sintering process. It is always advisable to contact the manufacturers before any design is considered.

8.7.2 Ferrite

The permanent magnet ferrites (also called ceramics) are mixed oxides of iron (ferric) oxide with a divalent metal oxide usually either barium or strontium. These ferrites have a hexagonal crystal structure, the very high anisotropy of which gives rise to high values of coercivity, e.g. 150–300 kA/m (compared with about 110 kA/m for the best Alnico alloys). The general formula is $MO.5.9(Fe_2O_3)$ where M is either Ba or Sr. These ferrites are made by mixing together barium or strontium carbonate with iron oxide in the correct proportions. The mixture is fired in a mildly oxidising atmosphere and the resulting mixture is milled to a powder of particle size of about 1 μm. The powder is then pressed in a die to the required shape (with a shrinkage allowance), anisotropic magnets are produced by applying a magnetic field in the direction of pressing. After pressing the compact is fired. This compound is a ceramic and can only be cut using high speed slitting wheels. Ferrite magnets are produced in large quantities in a variety of sizes for different applications. Flat rings are made for loudspeakers ranging up to 300 mm in diameter with a thickness of up to 25 mm. Segments are made for motors with ruling diameters from 40 to 160 mm and rectangular blocks are made for separators with dimensions of up to 150 mm × 100 mm × 25 mm. These blocks can be built up into assemblies or cut down into smaller pieces for a variety of applications. In each of these cases the preferred direction of magnetisation is through the shortest dimension. Generally the magnetic length does not exceed 25 mm and if a longer length is required, this is built up with magnets in series.

The great success of permanent ferrites is due to the low price per unit of available energy, the high coercivity, the high resistivity and the low density. The isotropic grades are the least expensive to manufacture and may be magnetised into complex pole configurations; they are used for a wide range of relatively small sized applications. The largest application in terms of market volume is for magnets for motors which use high coercivity anisotropic ferrite in the form of an arc-shaped segment. These segments are magnetised radially and a pair of segments surround the armature and provide the stator field. The application for these d.c. motors is particularly in the automotive industry, e.g. for windscreen wipers, window movers, blowers etc. Another large volume application is for loudspeakers where the high remanence grades are used. Other applications of anisotropic hard ferrites include magnetic chucks, magnetic filters and separators.

8.7.3 Rare earth cobalt

The $SmCo_5$ alloy was the first of the rare-earth permanent magnet alloys. The alloy powder is prepared by chemically reducing the rare earth oxide powder together with cobalt powder. The resulting powder is compacted and pressed almost invariably in a magnetic field to produce anisotropic magnets. The properties are due to a very strong magnetocrystalline anisotropy related to its hexagonal crystal structure. The range of sizes which can be produced is from about 1 mm cubes up to blocks of 50 mm × 50 mm × 25 mm, the short dimension being the preferred direction.

A more recent development is Sm_2Co_{17} which is the generic title for a series of binary and multiphase alloys with a number of transition metals replacing some of the cobalt. These alloys have a number of advantages over $SmCo_5$ alloys, they include a higher remanence B_r and energy product $(BH)_{max}$, greater temperature stability and lower material costs. This potential lower cost is however offset by the difficulties of production including a long heat treatment and inconsistency of properties. The good temperature stability, the best for the rare-earth alloys, makes it attractive for some applications.

Table 8.5 Characteristics of permanent magnet materials*

Material†⇐	B_r (T)	$(BH)_{max}$		H_{cB}		μ_r
		(kJm^3)	MGsOe	(kA/m)	Oe	
Alnico						
Normal anisotropic	1.1–1.3	36–43	4.5–5.4	46–60	580–750	2.6–4.4
Columnar	1.35	60	7.5	60	750	1.8
High coercivity	0.8–0.9	42–46	5.3–5.8	95–150	1200–1900	2.0–2.8
Columnar high H_{cB}	1.05	74	9.2	120	1500	1.8
Ferrites (ceramics)						
Barium isotropic (a)	0.22	8.0	1.0	130–155	1600–1900	1.2
Barium anisotropic (a)	0.39	28.5	3.6	150	1880	1.05
Strontium anisotropic (a)	0.36–0.43	24–34	3.0–4.3	240–300	3000–3800	1.05
La,Co substituted Sr ferrite (a)	0.42–0.44	33–36	4.1–4.5	320–330	4000–4100	1.05
Bonded ferrite						
Isotropic (a)	0.14	4.0–4.3	0.50–0.54	80–100	1000–1250	1.1
Anisotropic (a)	0.20–0.28	7–15	0.9–1.9	140–210	1700–2600	1.15
Samarium cobalt						
$SmCo_5$ sintered (b)	0.9	160–180	20–23	640–700	8000–8800	1.05
$SmCo_5$ bonded (b)	0.3–0.75	25–110	3–14	250–550	3000–6900	1.05–1.15
Sm_2Co_{17} sintered (c)	0.9–1.1	150–240	19–30	600–820	7500–10 000	1.1
Neodymium iron boron						
NdFeB sintered (d)	1.0–1.4	200–370	25–46	800–1050	10 000–13 000	1.05
NdFeb bonded (d)	0.45–0.75	35–90	4.4–11.3	300–540	3800–6800	1.15
Others						
CrFeCo anisotropic	1.3	32–48	4–6	45–50	560–630	2.5–3.5
FeCoVCr anisotropic	0.85	15	1.9	28	350	5

*B_r, remanence; $(BH)_{max}$, energy product; H_{cB}, coercivity; μ_r, relative recoil permeability. † Intrinsic coercivity H_{cJ}: (a) 160–400 kA/m, 2000–5000 Oe (b) 700–1300 kA/m, 8700–16 200 Oe (c) 700–1650 kA/m, 8700–20 600 Oe (d) 800–2500 kA/m, 10 000–31 200 Oe.

The Curie temperature of Alnico is 800–850°C, of ferrite 450°C, of SmCo over 700°C and NdFeB it is 310°C. The Alnicos have a resistivity of about 50×10^{-8} Ω-m, for the ferrites it is about 10^4 Ω-m and for the rare earths $90–140 \times 10^{-8}$ Ω-m.

Table 8.6 Applications of permanent magnet materials

Alnico	Watt-hour meters Moving coil instruments Weighing machines Loudspeakers (television)
Ferrite	Loudspeakers (hi-fi) Holding devices and chucks D.c. motors (e.g. for windscreen wipers, window shifters in cars)
Rare earth cobalt ($SmCo_5$)	Travelling wave tubes, klystrons and focusing assemblies Stepper and servo motors Centrifuges, gyroscopes Microspeakers and earphones
Rare earth cobalt (Sm_2Co_{17})	Couplings and frictionless bearings Stepper and servo motors Magnetic bearings and couplings
NdFeB	Computer applications including track shifters and compact disc motors. Focusing assemblies. Travelling wave tubes. MRI body scanners. Magnetic bearings and couplings Brushless synchronous and stepper motors, generators Micromotors and actuators
NdFeB bonded	Walkman motors and earpieces Small motors for hand tools, toys etc. Holding devices

8.7.4 Neodymium iron boron

The basic alloy is $Nd_{15}Fe_{77}B_8$; it came out of efforts to produce high energy magnets which do not contain cobalt and from interest in amorphous alloys produced by very rapid cooling from the melt. When alloys of the appropriate NdFeB composition are recrystallised, a material with very good permanent magnet properties is obtained. The crystal structure of this alloy is tetragonal, the long axis being the preferred magnetic direction. The alloy powder can be produced by ball milling or by reduction of the neodymium oxide using calcium. However, a more popular development is to reduce the particle size of the alloy by contact with hydrogen. In this HD (hydrogen decrepitation) process, the hydrogen is absorbed with considerable expansion of the alloy causing it to break down into powder. The hydrogen is then pumped away. This is followed by jet milling to reduce the size of the powder to the optimum size. The resulting powder is then pressed in a magnetic field to produce an anisotropic compact and sintered. Throughout the processing the powder is kept either under a protective atmosphere or under vacuum to protect the neodymium from oxidation.

The advantages of magnets made from this alloy are that they have higher remanence B_r and $(BH)_{max}$ than the SmCo alloys and they are also cheaper because Co and Sm have been replaced by Fe and Nd. The disadvantages are that they are subject to corrosion and to a rapid change of magnetic properties with temperature, particularly coercivity changes. Coating the magnets with polymer or metal films, depending on the application can prevent corrosion. The elevated temperature properties can be improved by additions of small amounts of dysprosium, gallium, niobium or vanadium. These last two elements may also give improved corrosion protection.

A further development of the HD process is the HDDR (hydrogen, disproportionation, desorption and recombination) process. This can produce very fine powder after a heat treatment under vacuum. This powder is suitable for the production of anisotropic magnets by pressure or for mixing with a polymer to produce bonded magnets.

8.7.5 Bonded materials

Each of the above materials can in the powdered form be mixed with a bond pressed to shape and cured either cold or in an oven. The bonds can be rubber, polymer or plastics and they may be flexible or rigid. The flexible rubber bonded ferrites have found wide application in holding and display devices and the best quality anisotropic material is used in small motors. Alnico and rare earth cobalt are also made in the bonded form, they have the advantage of a uniform level of properties and freedom from cracking. Bonded magnets are not so good as the cast or sintered materials, at best the $(BH)_{max}$ is about 50% of the solid forms.

For bonded NdFeB magnets the powder is very often made in the form of flake. Rapidly quenched melt spun ribbon is produced by ejecting molten NdFeB alloy through a fine orifice onto a rapidly rotating water cooled wheel. The ribbon is crushed into fine flake. The bonded magnets are produced by mixing the annealed flake with epoxy resin, pressing the mixture into the required shape and subsequently oven curing. The moulding procedure may be either injection moulding or compression moulding. Injection moulding is the best for bulk production but compression moulding provides better magnetic properties. This MQ1 flake material is isotropic and is used for a wide range of applications including hand held tools, holding devices and motors. Powder produced by the HDDR process mentioned above may be better for some applications as a higher $(BH)_{max}$ is possible than with the flake.

8.7.6 Other materials

There are a number of other permanent magnet materials with minority applications. This is often because they can be mechanically formed to shape, which is not possible for Alnico, ferrite or rare-earth alloys. It must be emphasised that these other materials are only available in a limited range of sizes and shapes.

Steels, which were the original permanent magnets, are now almost obsolete because their properties are inferior to those of the materials mentioned above. They do, however, find applications in hysteresis motors where lower coercivities are required.

CrFeCo is an alloy which can be rolled and drawn into wire, it has properties similar to those of Alnico. The alloy FeCoVCr can be drawn into wire, which produces the anisotropy. Platinum cobalt (PtCo) which has a very high coercivity was much used in the early days of space research. For most applications it has now been replaced by rare-earth alloys. However it is still used for medical and other applications where resistance to corrosion is required.

8.7.7 Properties, names and applications

The range of properties for each class of material is given in *Table 8.5*. These ranges include deliberate variations in properties obtained by small changes in composition and heat treatment.

Permanent magnets have for many years been used for a wide range of applications. The ferrite (ceramic) magnets provide cheap magnets which are used for small motors and many other applications in the automobile industry. The ferrites are also used as ring magnets for radio and hi-fi loudspeakers and are prepared as blocks which can be cut up for a wide range of applications. The very great improvement in the available properties provided by the rare-earth magnets has focused interest on permanent magnet motors and the associated electronic equipment. In recent years there has been a movement from the SmCo alloys to NdFeB alloys but for some applications Sm_2Co_{17} is preferred because of its better corrosion resistance and elevated temperature properties. *Table 8.6* lists the important applications of the various permanent magnetic materials.

References

BECKLEY, P. *Electrical Steels* published by European Electrical Steels (now Cogent Power Ltd), Newport, Gwent, UK (2001)

BOLL, R. (Ed.), *Soft Magnetic Materials*, Vacuumschmelze Handbook, Heyden (1977)

CAMPBELL, P., *Permanent Magnet Materials and their Application*, Cambridge (1994)

COEY, J. M. P. (Ed.), *Rare-earth Iron Permanent Magnets*, Oxford (1996)

GOLDMAN, A., *Modern Ferrite Technology*, Van Nostrand Reingold, New York (1990)

JILES, D., *Introduction to Magnetism and Magnetic Materials*, 2nd edition, Chapman and Hall, London (1998)

LAX, B. and BUTTON, K. J., *Microwave ferrites and Ferrimagnetics*, McGraw-Hill Book Co., New York (1962)

McCAIG, M. and CLEGG, A. G., *Permanent Magnets in Theory and Practice*, 2nd edition, Pentech and Wiley, New York (1987). (Gives a list of permanent magnet makers throughout the world)

PARKER, R. J., *Advances in Permanent Magnetism*, Wiley, New York (1990)

SNELLING, E. C., *Soft Ferrites: Properties and Applications*, 2nd edition, Butterworths, London (1988). (Gives a global list of soft ferrite makers)

WOHLFARTH, E. P. (Ed.), *Ferromagnetic Materials*, 6 vols, North Holland, Amsterdam (1980–1990) continued with BUSCHOW, K. H. to 10 vols (1998)

For more recent information it will be found useful to refer to:

IEEE Transactions on Magnetics which contains the proceedings of Intermag Conferences.

Proceedings of International Workshops on Rare Earth Permanent Magnets, from the University of Dayton Dayton, OH, USA

MagNews published by the UK Magnetics Society, Berkshire Business Centre, Wantage, Oxon, OX12 8SH

IEC Standards on Magnetic Materials, No. 60404 parts 1 to 9 on the classification, methods of measurement and specification of magnetic materials. These standards have also been published as parts of British Standards 6404 (1984–2000)

9

Electroheat and Materials Processing

P L Jones PhD, CEng, MIMechE
Petrie Technologies Ltd, Chorley, Lancs
(Sections 9.1 to 9.9)

S Taylor BSc, PhD, MEng, ACGI, CEng, MIEE
University of Liverpool
(Section 9.10)

S Nakai
Institute of Laser Engineering, Osaka University
(Section 9.11)

J Jennings

Contents

9.1 Introduction 9/3

9.2 Direct resistance heating 9/3
  9.2.1 Metals 9/3
  9.2.2 Glass 9/4
  9.2.3 Water 9/4
  9.2.4 Salt baths 9/4
  9.2.5 Other fluids 9/4

9.3 Indirect resistance heating 9/5
  9.3.1 Metallic elements 9/5
  9.3.2 Sheathed elements 9/5
  9.3.3 Ceramic elements 9/5
  9.3.4 Terminals and leads 9/5
  9.3.5 Aggressive environments 9/6
  9.3.6 Infra-red heaters 9/8

9.4 Electric ovens and furnaces 9/9
  9.4.1 Heating-element construction for ovens and furnaces 9/9
  9.4.2 Ovens 9/9
  9.4.3 Furnaces 9/10

9.5 Induction heating 9/10
  9.5.1 Power sources for induction heating 9/12
  9.5.2 Load matching 9/13
  9.5.3 Coil design 9/13
  9.5.4 Through heating of billets and slabs 9/14
  9.5.5 Strip heating 9/14
  9.5.6 Surface and localised heating 9/14
  9.5.7 Semiconductor manufacture 9/14
  9.5.8 Indirect induction heating for non-metals 9/14

9.6 Metal melting 9/15
  9.6.1 Introduction 9/15
  9.6.2 Arc furnace 9/16
  9.6.3 Coreless induction furnace 9/17
  9.6.4 Channel induction furnace 9/19
  9.6.5 Resistance furnaces 9/20

9.7 Dielectric heating 9/20
  9.7.1 RF dielectric heating systems 9/21
  9.7.2 Microwave power sources and applicators 9/23

9.8 Ultraviolet processes 9/24

9.9 Plasma torches 9/24
  9.9.1 Types of plasma-torch design 9/24
  9.9.2 Electrical connection of plasma torches 9/25
  9.9.3 Performance 9/25
  9.9.4 Plasma furnaces or reactors 9/26

9.10 Semiconductor plasma processing 9/26
  9.10.1 Basic mechanisms in plasma processing 9/27
  9.10.2 Power supplies for plasma production 9/29
  9.10.3 Current trends and future developments 9/30

9.11 Lasers 9/30
  9.11.1 Introduction 9/30
  9.11.2 Gas lasers 9/32
  9.11.3 Solid-state lasers 9/32
  9.11.4 Application of high-power lasers 9/34
  9.11.5 Laser pumping methods and electric power supplies 9/37

9.1 Introduction

The three major uses of electricity in industry are motive power, lighting and in the provision of heat for processes. The latter is perhaps the most interesting since it covers such a wide variety of products ranging from metals, glasses and ceramics to textiles, paper, food and drink. The processes covered involve aspects of the whole frequency range from d.c. to ultraviolet and power levels from a few watts to many megawatts (*Table 9.1*). With the changes away from 'smoke stack' industries some electroheat processes have decreased in importance but others involved in the newer higher value operations are widely accepted.

9.2 Direct resistance heating

Direct resistance heating is used in the iron and steel industry: for heating rods and billets prior to rolling and forging; for ferrous and non-ferrous annealing; either alone or in combination with other fuels for melting glass; in electrode boilers, for water heating and steam raising; and in salt baths, for the surface heat treatment of metallic components.

9.2.1 Metals

The resistivity of several common materials is shown in *Table 9.2*, and *Figure 9.1* shows the variation in resistivity with temperature of some of these. The relatively high resistivity of steel allows billets of up to 200 m^2 to be heated efficiently, provided that the length is several times greater than the diameter. The heating time is of the order of seconds to a few minutes, so that heat losses (e.g. radiation from the surface and thermal conduction through the contacts) are small. The efficiency of the process can be of the order of 90% or better. The workpiece resistance is normally low, and for these efficiencies to be achieved the supply resistance must be much lower. The low resistivity of copper and similar materials implies that the length/diameter ratio should be considerably higher than 6 if the process is to be successful, and with these materials the more common application is the annealing of wire and strip. In all cases the cross-sectional area of the current-flow path must be uniform, otherwise excessive heating, with the possibility of melting, will occur at the narrower sections. With the normal 50-Hz supplies non-uniform current distribution caused by skin effect leads to higher heating rates at the surface, but this is counteracted by increased surface heat loss.

Table 9.1 Electric heating processes

Technique	*Frequency range*	*Power range*
Direct resistance	0–50 Hz	0.01–30 MW
Indirect resistance	50 Hz	0.5–5 kW
Oven, furnace	50 Hz	0.01–1 MW
Arc melting	50 Hz	1–100 MW
Induction heating	50 Hz–450 kHz	0.02–10 MW
Dielectric heating	1–100 MHz	1–5000 kW
Microwave heating	0.5–25 GHz	1–100 kW
Plasma torch	4 MHz	0.001–1 MW
Laser CO_2	30 THz	0.1–60 kW
Infra-red	30–400 THz	1–500 kW
Ultraviolet (mercury arc)	750–1500 THz	1 kW

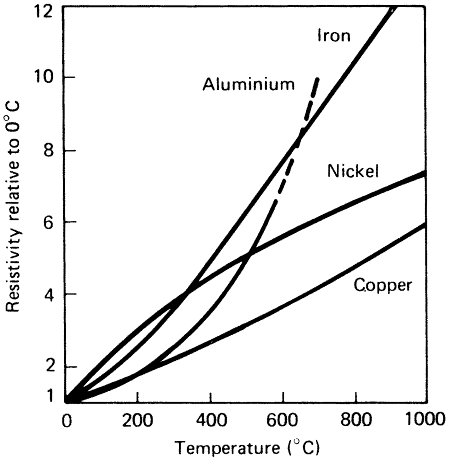


Figure 9.1 Variation of resistivity with temperature (relative to 0 °C) of aluminium, copper, nickel and iron

Table 9.2 Resistivity of typical metals

Metal	*Resistivity* (μΩ-cm)
Copper	1.6
Aluminium	2.5
Nickel	6.1
Iron	8.9
Mild steel	16.0
Stainless steel	69.0

A typical supply circuit is shown in *Figure 9.2*. The high-current transformer will usually have off-load tappings to take account of the major changes in workpiece resistivity that occur over the heating cycle. The reactance of this transformer and its associated connections should be minimised and the whole arrangement must be designed to withstand the magnetic forces at the high currents involved which, in the larger units, can be in excess of 100 kA. Although this is a resistance-heating application, the inductance of the circuit may result in a power factor as low as 0.4 at the start of the process, and power-factor-correction capacitors are often incorporated.

A high contact resistance between the supply and the workpiece leads to excessive voltage drop and local overheating. Large-volume contacts behave as heat sinks with

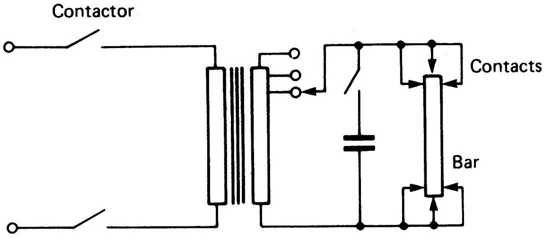


Figure 9.2 Electric circuit for direct resistance billet heater

the possibility of cold ends, and welding of the contact to the bar may occur. Allowance needs to be made for the longitudinal thermal expansion during heating. One commercial arrangement uses a number of hemispherical contacts of copper alloy at each end of the bar, applied by hydraulic pressure. Liquid contacts have been employed for the continuous heating of wire and strip. Sliding or rolling metal–metal contacts are also used.

Process control can be achieved on the basis of a simple timed cycle with the transformer tappings adjusted during the cycle to maintain the required current. A refinement involves weighing the billet before the process is started and using this information to adjust the cycle time to give the required temperature. Other installations are controlled by an optical pyrometer to monitor the surface temperature and feed back a signal to the control system. Wire and strip heating installations usually employ throughput speed as the control variable.

The direct resistance heating of bar or billets is a one-phase load that is switched at frequent intervals. This results in voltage unbalance at the point of common coupling and in transient voltage disturbances. The load can be phase balanced by inductive and capacitive components. In large units the switch-on disturbances may be compensated by a soft-start arrangement.

9.2.2 Glass

Glass at temperatures above 1100°C, has low viscosity and a resistivity low enough (*Figure 9.3*) for direct resistance heating at acceptable voltages to be considered.

In the UK, electricity is used in mixed melting units where, typically, electrodes are added to a fuel furnace to increase the output for a relatively low capital outlay. Current is passed between electrodes immersed in the molten glass. These electrodes must withstand the high temperatures involved and the movement of molten glass across their surface, and must be protected from exposure to the atmosphere. Contamination of the glass by pick up of electrode material must also be avoided and either molybdenum or tin oxide electrodes are used with current densities of the order of 1.5 kA/m^2. A one-, two- or three-phase electrode system may be employed; the two-phase connection uses Scott-connected transformers. The three-phase arrangement is preferred for phase balance and low cost, and also produces electromagnetic forces in the glass which, together with the thermal forces, lead to significant movement in the melt. Provided that it is not excessive, the movement improves quality and melting rate.

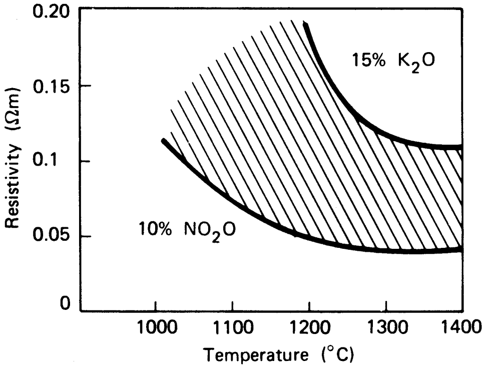


Figure 9.3 Variation of resistivity with temperature of glasses

Existing electric furnaces have usually followed a traditional design pattern with a rectangular tank, but circular designs have now been developed which are claimed to give improved stirring in the melting zone.

The power input to the furnace, and hence the melting rate, is controlled by varying the input voltage by use of tapped transformers or saturable inductors. An alternative method is to change the effective surface area of the electrodes by raising them from the melt.

9.2.3 Water

The generation of steam and hot water by passing current between electrodes in water is now common practice. As with molten glass, the conductivity is dependent on the ions that the water contains, but in many districts it is sufficient to use normal tap water. Electrode boilers range in size from a few kilowatts up to 20 MW. The larger capacity units operate at high voltages up to 6 kV, and the water is sprayed on to the cast-iron electrode. The heating rate is regulated by moving a porcelain shield over the jets to divert the water. Output control in the smaller units is achieved by varying the surface area of the electrode in contact with the water, either by vertical movement of the electrodes or, more simply, by changing the water level in the boiler. It is important to ensure that the conductivity of the water is maintained at the appropriate value: too low a conductivity reduces the power, while if the resistivity is too high the result is a deposition of insoluble salts on the electrodes. In large units the conductivity and pH are continuously monitored and their values automatically controlled.

9.2.4 Salt baths

Salt baths can be used for heat treatment of metal components. The heated salt reacts chemically with the surface layer of the workpiece to give the required surface properties. At temperatures above about 800°C direct resistance heating as opposed to the use of sheathed elements or external heat sources is the only usable method. Although the salt is a good conductor when molten, the bath must be started up from cold by using an auxiliary starting electrode to draw a localised arc. The electrodes have to withstand the corrosive effects of the salt and are manufactured from graphite or a corrosion-resistant steel alloy. The currents involved can be up to 3 kA at voltages of 30 V. Both one- and three-phase units are available.

9.2.5 Other fluids

In recent years the principles of electrode boilers have been extended to the heating of a number of other fluids, under the generic term of *ohmic heating*. Units having a heating capacity of up to 200 kW operating at voltages up to 3.3 kV have been used to heat foodstuffs containing solid particulates (with characteristic dimensions of up to 25 mm) to temperatures of 140°C, thus cooking and sterilising the foodstuff in a single very rapid (in some instances a few tens of seconds) operation, which preserves the quality and taste of the food for subsequent packaging and unrefrigerated storage with a shelf-life of many months.

The potential of the technique for heating other fluids which tend to foul conventional heat-transfer systems, but without the expense of microwave or r.f. generators, has also been exploited in the heating of clarified sewage and sewage sludge to maintain process temperatures during the treatment at sewage works, and as a possible method of pasteurising the final effluent from the works.

The technique also lends itself to the heating of highly corrosive conducting liquids, provided materials can be identified for the electrodes and heater enclosure which are capable of withstanding the conditions in the heater for reasonable periods of time, without introducing unacceptable levels of contamination into the liquid stream.

9.3 Indirect resistance heating

In this method, electricity is passed through a suitable conducting material or element which then transfers its heat by conduction, convection and/or radiation to the target material or process. The element becomes the hottest component of the system and the factors which determine the choice of element materials depend on the nature of the heat transfer process employed and the physical and chemical characteristics of the process environment.

All such elements have the following characteristics:

(1) they are electrically conducting materials;
(2) they are supplied with electricity via a suitable contact, cold end or terminal block, and leads;
(3) they need mechanical support;
(4) they are solid materials; and
(5) they have properties enabling an economic operating life for the environment or process chosen.

9.3.1 Metallic elements

Traditionally, metallic elements take the form of wire, strip or tape. This calls for reasonable ductility in the manufacturing process. For a given operating temperature, the element life tends to decrease with decreasing cross-section of element caused by progressive oxidation of the surface or reduction in mechanical strength and so, in practice, an optimum element thickness for economic life exists for each application. The choice of metal composition will depend upon the operating temperature required, the material resistivity, the temperature coefficient of resistance, high temperature corrosion resistance, mechanical strength, formability, and cost. Metallic elements can be manufactured to close tolerances and, provided the material is in the fully annealed, slowly cooled condition, will have nominal resistivities within ±5%.

Whilst many metals and alloys are used as element materials, the most common for industrial applications are alloys based on nickel–chromium, iron–nickel–chromium alloys, or iron–chromium–aluminium alloys. All depend on an adherent non-spalling, self-sealing oxide layer forming on the surface of the alloy. Further oxidation is limited by diffusion of reacting species through this oxide layer. The iron–chromium–aluminium alloys rely on an alumina film developing on the surface, whereas the nickel–iron–chromium alloys depend on chromium oxide. Generally, the iron–chromium–aluminium alloys can operate at higher temperatures than the nickel–chromium based alloys, but are not as readily fabricated. The most exotic metals (platinum, tantalum, molybdenum, etc.) are used for special laboratory or high temperature vacuum work. *Table 9.3* summarises some of the characteristics of these materials.

When used in a freely radiating state, the elements need to be supported on ceramic tubes, pins or correctly insulated metallic pins. If ceramic supports are chosen, it is important that they have sufficient mechanical strength to prevent sagging and have a sufficiently high electrical resistance at the working temperature to prevent excessive leakage current. Elements can also be supported in grooves in bricks, partially or fully embedded in ceramic fibre vacuum formed blocks or cast refractory blocks. It is important to reduce the furnace wall loadings to suitable levels when employing these techniques, and that the electrical loadings, often quoted in terms of watts per square centimetre of element surface, and the geometry of the element and supports comply with the element manufacturer's recommendations.

Advice and literature on this subject is readily available from suppliers.

9.3.2 Sheathed elements

The elements may be protected from the working environment by the use of a suitable insulation layer separating the element from an outer sheath. In many domestic appliances, for example cooker rings, immersion heaters and kettle elements, a purified magnesium oxide powder separates the helical element coils from copper, stainless steel or nickel based alloy sheath material. In these cases the element is often rated in terms of watts per square centimetre of sheath. Such mineral insulated elements are also used in industrial applications as cartridge heaters, radiant panels and immersion heaters.

Thin strip or band heaters are available which use a mica insulation between the element and the sheath.

Higher rated industrial units employ a protective metallic or thermally conducting ceramic sheath insulated from the elements by an air gap created by suitable ceramic supports and spacers. As before, careful consideration should be given to selection of materials for use in each application regarding the heat transfer, electrical and mechanical properties and corrosion characteristics.

9.3.3 Ceramic elements

Silicon carbide, molybdenum disilicide, lanthanum chromite and hot zirconia are examples of ceramic materials which have sufficient electrical conductivity to act as element materials. The silicon carbide and molybdenum disilicide elements depend on a protective, self-sealing silica layer on the surface. These materials tend to be brittle and have to be handled with care, but they can achieve much higher temperatures and surface loadings in air than can conventional metal elements. They also can have unusual temperature coefficients of resistance as shown in *Figure 9.4*. Note that the values shown in this figure are for illustrative purposes only. The actual resistivity will depend on purity, grain size and method of manufacture. Careful selection of element size, shape and working resistance is required in practice, and advice regarding the choice of element, support, insulation and electrical supply characteristics should be sought.

Graphite is another recognised non-metallic element material which, of course, should be operated in the absence of oxygen or gaseous oxygen compounds such as steam and carbon dioxide.

Ceramic elements generally consist of a hot zone either created by a thin section or a spiral cut supported by two or more cold ends which are either thicker in cross-section or which have been impregnated with a metallic phase to lower the resistance locally. Examples, together with their metallic counterparts, are shown in *Figure 9.5*. Ceramic elements must be free to expand and contract in their support system, otherwise tensile failure may occur.

9.3.4 Terminals and leads

The means of connecting the elements to the electrical supply is a very important design requirement and the correct choice of location, materials and joining method can

Table 9.3 Materials for resistance-heating elements*

Material	$\theta\varsigma$(°C)	$\rho\varsigma$(10^{-8}/m)	$\alpha\varsigma$(10^{-3}/K)	*Principal applications*
Nickel based alloys†				
80 Ni/20 Cr	1200	108	+14	Furnaces, resistance heaters, mineral insulated elements for domestic and industrial use
60 Ni/15 Cr/bal Fe	1150	111	+18	Firebar and convector heaters. Domestic and furnace applications up to 1100°C
35 Ni/20 Cr/bal Fe	1100	104	+29	Some domestic appliances and general heating equipment at moderate temperatures
20 Ni/25 Cr/bal Fe	1050	95	—	Terminal blocks
Iron based alloys†				
22 Cr/5.8 Al/bal Fe	1400	145	+3.2	Furnaces for heat treating glass, ceramics, steel, electronics; crucible furnaces for melting/holding aluminium and zinc
22 Cr/5.3 Al/bal Fe	1375	139	—	Industrial furnaces
22 Cr/4.8 Al/bal Fe	1300	135	4.7	Furnaces for moderate temperatures, appliances
Exotic metals				
Platinum	1300	10.58	3.92	Laboratory furnaces, small muffle furnaces
90 Pt/10 Rh	1550	18.7	—	
60 Pt/40 Rh	1800	17.4	—	
Molybdenum‡⇐	1750	5.7	4.35	Vacuum furnaces, inert atmosphere furnaces
Tantalum‡⇐	2500	13.5	3.5	Vacuum furnaces
Tungsten‡⇐	1800	5.4	4.8	Incandescent lamps, vacuum and inert atmosphere furnaces
Non-metallic materials				
Graphite‡⇐	3000	1000	−26.6	Vacuum, inert gas, reducing-atmosphere furnaces
Molybdenum disilicide	1900	40	1200	Glass industry, ceramic kilns, metal heat treatment, plus laboratory furnaces
Silicon carbide	1650	1.1×10^{5}	—	Furnaces for heat treatment of meals, ceramic kilns, conveyor furnaces
Lanthanum chromite	1800	2100	—	Laboratory furnaces and special ceramic kilns
Zirconia§	2200	—	—	Laboratory furnaces and special ceramic kilns

*θ, Maximum element operating temperature; ρ, electrical resistivity at 20°C; α, mean temperature coefficient of resistance at 20°C.

† Approximate compositions.

‡ Not to be used in atmospheres containing oxygen, oxides of carbon, water vapour, etc.

§ Becomes sufficiently conducting at temperatures in excess of 1000°C.

maximise reliability. The temperature of the joint should be kept as low as possible. This can be achieved by using large terminations of high thermal conductivity arranged such that the contact resistance is low and any heat generated in this region can be dissipated correctly. Materials with a good electrical conductivity and high oxidation resistance at the expected temperature are recommended. Several types of nickel alloy can be employed for high-temperature connections, and aluminium braid, held in place with special clips, is used to connect to the cold ends of ceramic elements.

For terminals exposed to wet conditions it is important to ensure adequate corrosion resistance and this should include consideration of avoiding electrochemical corrosion couples by choice of incompatible alloys as well as the danger of electrical shorting across the terminals themselves.

Leads should be of low resistance and should have sufficient section to dissipate any joule heating (I^2R) incurred. Note that the use of thyristor-control systems may require the use of leads and cables up-rated to cater for harmonic currents.

9.3.5 Aggressive environments

Carbon and sulphur are harmful to nickel based alloys, and chlorine or molten chloride or liquid metal splashes can damage both nickel and iron–chromium–aluminium based alloys. Element materials which rely on a self-sealing oxide for protection may be damaged by reducing atmospheres and the maximum operating temperature of the elements for such conditions, or the element surface electrical load (W/cm^2), may be reduced accordingly. Protective tubes or sheaths may be used for such environments. However, this may impose an additional space requirement in the process and should be considered during the design stage.

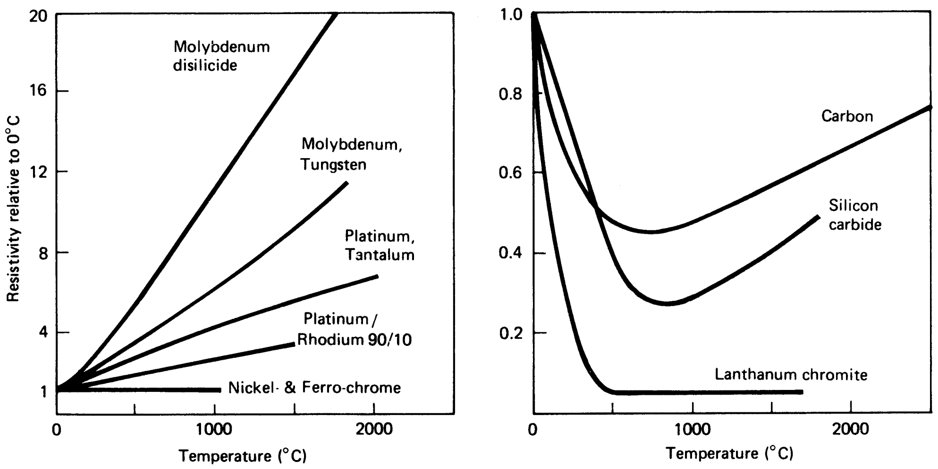


Figure 9.4 Resistivity (relative to 0°C) of some resistance materials

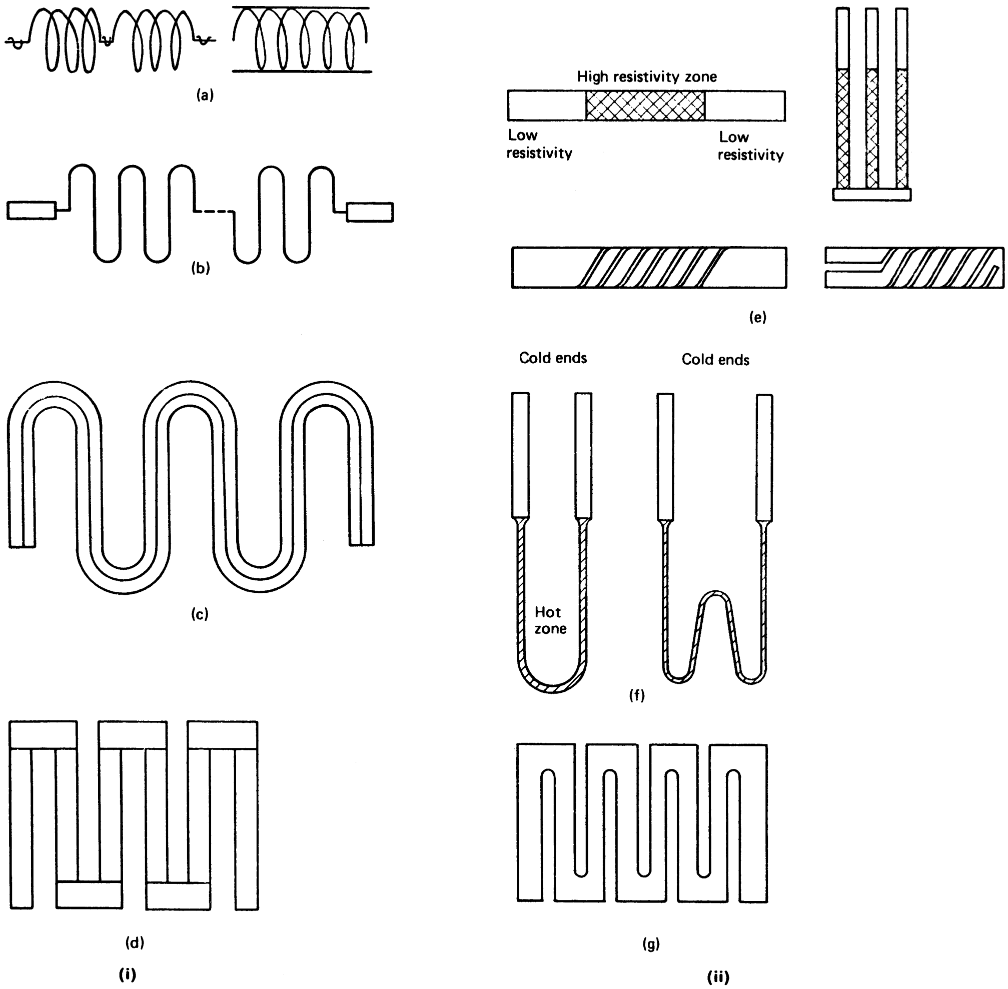

(a)
(b)
(c)
(d)
(i)
High resistivity zone
Low resistivity
Low resistivity
(e)
Cold ends
Cold ends
Hot zone
(f)
(g)
(ii)

Figure 9.5 Construction of heating elements used in ovens and furnaces: (i) metal heating elements; (a) wire coil supported on hooks or in grooved tile; (b) strip; (c) cast element; (d) tubular heating elements; (ii) non-metallic heating elements; (e) rod and tubular silicon carbide elements; (f) molybdenum disilicide; (g) graphite

Table 9.4 Examples of immersion-heater sheath materials

Iron, steel
Grey cast iron
Ni resist cast iron
Aluminium
Copper
Lead
Monel 400*
Nickel 200*
304, 321, 347 stainless steel
316 stainless steel
Type 20 stainless steel
Incoloy 800*
Inconel 600*
Titanium
Hastelloy B†
Quartz
Graphite
Teflon
Nitride bonded silicon carbide
Silicon carbide

* Registered tradenames of Inco Europe Ltd.
† Registered tradename of Cabot Corporation.

Immersion heater sheath materials should be matched to the process environment too. *Table 9.4* lists some of the materials commonly used for this purpose.

9.3.6 Infra-red heaters

Infra-red is one of the most widely used electrical process heating techniques in industry, finding applications in all sectors. The intensity distribution of radiation with wavelength from an infra-red source varies with its temperature. Proportionately more of the power is radiated at shorter wavelengths as the temperature of the source is increased (*Figure 9.6*). Typical input power densities to different types of infra-red sources are given in *Table 9.5*, with many of the comments given earlier in Section 9.3 applying also to their use for infra-red heating. The lower temperature infra-red sources (long and medium wave) are normally used for heating and drying non-metallic materials as these generally absorb at the longer wavelengths. Short-wave heaters are used where higher product temperatures are required (e.g. metal heat treatment), high intensities and through heating of suitable materials.

Efficient use of infra-red radiation for heating is assisted by matching the spectral output of the infra-red source to the reflective and absorptive properties of the product in the infra-red range. Materials considered to be good absorbers of infra-red radiation, because they are opaque, can show a significant degree of reflection, particularly at short wavelengths. Materials of thick section which show a considerable degree of transparency to some infra-red wavelengths, can advantageously be through heated by choosing an infra-red source producing most of its output at these wavelengths.

Control of infra-red heating can be effected in a number of ways. Infra-red heaters are generally of low power rating (less than 10 kW). Thus an infra-red heating installation will often use many heaters which can either be switched on/off individually or in groups, or the voltage/current to each or groups can be adjusted. As control of heating systems increases, contact or remote (pyrometers) temperature sensing of the product is used, with information fed back to the control system for the heaters. One aspect of altering the power input to an infra-red heater is that its temperature, and hence its spectral output, will be changed which can have implications for the efficient heating of those materials which show a variation in infra-red absorption with wavelength.

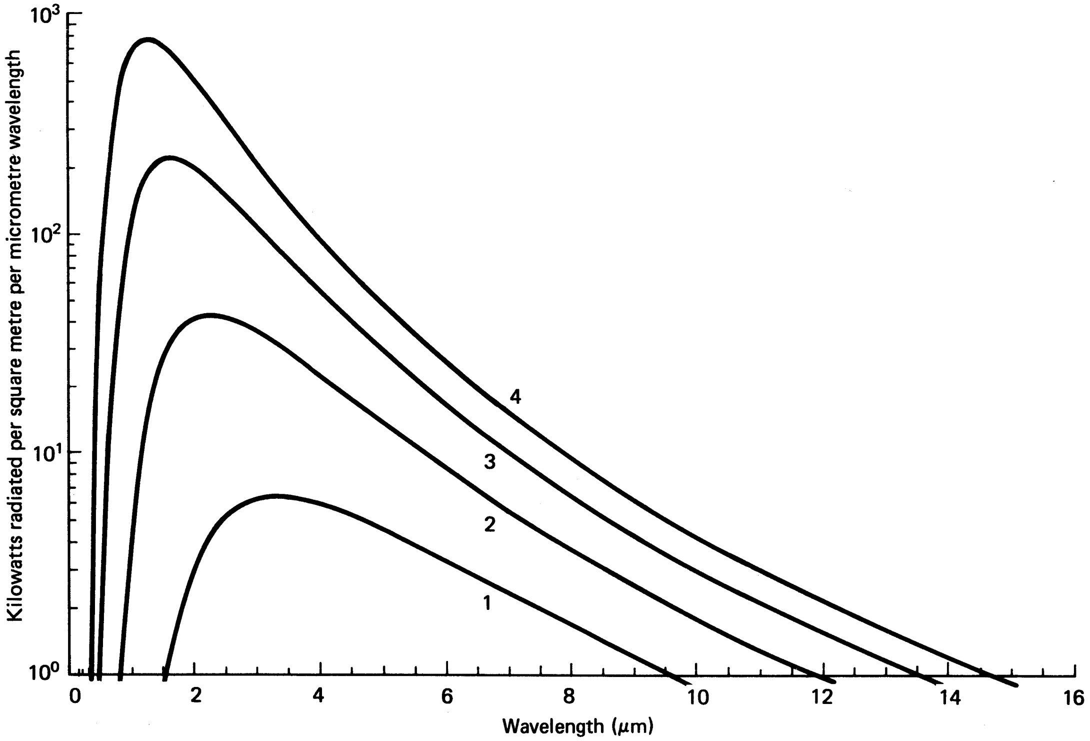


Figure 9.6 Black-body radiation: 1, 600°C; 2, 1000°C; 3, 1500°C; 4, 2000°C

Table 9.5 Characteristics of infra-red heaters

Type	*Maximum surface temperature* (°C)	*Maximum power density* (kW/m^2)	*Heater wavelength*
Embedded ceramic heater	600	68*	Long
Mineral insulated element	800	40*	Long
Tubular quartz-sheathed element	900	50*	Medium
Circular heat lamp (375 W)	2100	30*	Short
Linear heat lamp (1 kW)			
Parabolic reflector	2100	50*	Short
Elliptical reflector	2100	90†⇐	
Quartz-halogen linear heat lamp (12 kW)			
Parabolic reflector	2700	200*	Short
Elliptical reflector	2700	3000†⇐	Short

* Average power density.
† Power density at focus.

9.4 Electric ovens and furnaces

Electric ovens and furnaces are used for a great variety of different processes, ranging from sintering ceramic materials at temperatures up to 1800°C to drying processes close to ambient temperature at power ratings varying from a few kilowatts to more than 1 MW. A substantial part of the electric heating load is taken by electric ovens and furnaces of conventional design used for heat treatment (*Table 9.6*) and similar processes at temperatures up to about 1100°C.

9.4.1 Heating-element construction for ovens and furnaces

The materials used for resistance heating elements have been listed in *Table 9.3*. Forms of construction of furnace heating elements are shown in *Figure 9.5*. Metal resistance heating elements for furnaces are normally in the form of wire, strip or tube. Heavy-section low-voltage high-current elements are made with an alloy casting or with corrugated or welded tube. Helically wound wire heating elements are manufactured with wire-to-mandrel diameter ratios of between 3:1 and 8:1, limited by collapse of the helix. The helix is then expanded so that its length is about three times the close-wound length. Coiled-wire or strip heating elements may be inserted in ledges or grooves supported at intervals by nickel alloy or ceramic pegs. The end connections of the heating elements, normally made of a material different from that of the element to reduce attack from oxidation and chemical reaction with the refractories, have a lower resistance to reduce heat dissipation where the leads pass through the furnace wall; this lower resistance is achieved by making the diameter of the ends greater than that of the heating zone or by using a material of high electrical conductivity.

Table 9.6 Heat-treatment processes

Metal	*Treatment process*	*Temperature range* (°C)
Aluminium and alloys	Annealing	250–520
	Forging	350–540
	Solution heat treatment	400–500
	Precipitation hardening	100–200
	Stress relief	100–200
Copper	Annealing	200–500
Brass	Annealing	400–650
Magnesium and alloys	Annealing	180–400
	Solution heat treatment	400
	Precipitation hardening	100–180
Nickel and alloys	Annealing	650–1100
Carbon steel (0.6–1.5% C)	Annealing	720–770
	Forging	800–950
	Tempering	200–300
	Hardening	760–820
Stainless steels	Tempering	175–750
	Hardening	950–1050
Cast iron	Annealing	500

Silicon carbide rod or tubular heating elements are mounted vertically at the side or arranged to span the roof of the furnace. The cold ends of these rod elements are impregnated with silicon or made by joining end sections of higher conductivity. The hot zone of a tubular element is obtained by cutting a helix so that the current path length is increased and the cross-section is reduced which increases the resistance. Single-ended double-spiral and three-phase rod heaters can be suspended vertically from the roof of the furnace.

Molybdenum disilicide is normally available only in the hairpin or W-form of construction for suspending vertically or supporting in a horizontal plane. The heating zone is of reduced cross-section to increase the resistance. Graphite in the form of machined rods or slabs, or as tubes, is used in muffle furnaces.

9.4.2 Ovens

An oven is usually defined as having an upper temperature limit of about 450°C. Ovens, using natural or forced convection, are widely employed for drying and preheating plastics prior to forming, curing, annealing glass and aluminium baking and a host of other applications. Coiled nickel–chrome wire or mineral-insulated metal sheathed heating elements are distributed around the oven so as to obtain as uniform a temperature distribution as possible. Heat transfer rates may be increased by using a fan to circulate air over the heating elements onto the workpiece, the air being recirculated through ducts. An important advantage of convective ovens is

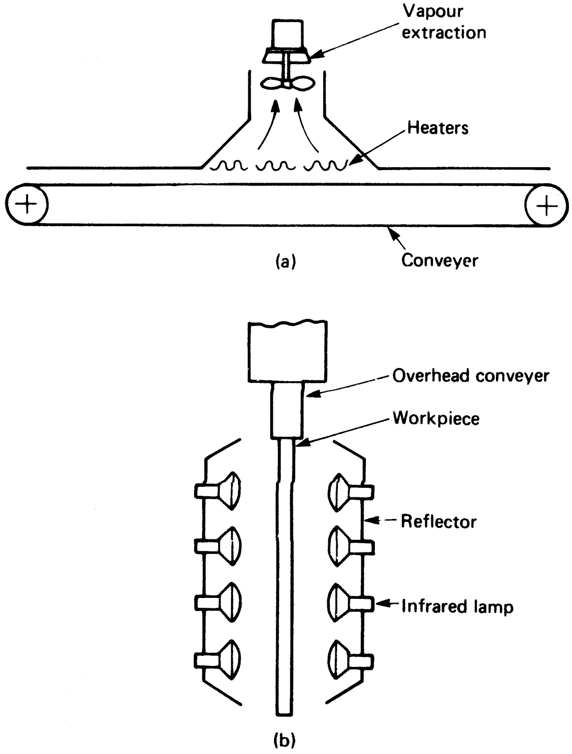


Figure 9.7 Infra-red conveyor ovens. (a) Horizontal conveyor; (b) Vertical conveyor

that the operating temperature is normally the element temperature and the maximum temperature is never exceeded as the process is self-limiting, which prevents overheating if the material is left in the oven too long. This is particularly important for temperature-sensitive materials such as plastics.

High heating rates can be achieved by direct radiation from heating elements in infra-red ovens. The oven walls are made of sheet metal which reflects the radiation, and vapour is easily removed. Two forms of continuous infra-red ovens are illustrated in *Figure 9.7*, the low thermal mass and weight allowing a light-weight structure and a very high power density. The high-intensity lamps employed have a low thermal inertia and can be switched to a low level of stand-by power, full power being used only when the workpiece is inside the oven. Infra-red heating processes are advantageous when only the surface layer is required to be heated, for example when curing coatings, so that the overall efficiency may be very high compared with other methods which heat also the substrate.

9.4.3 Furnaces

Furnace construction depends on the application. Some examples of batch furnaces for heat treatment are shown in *Figure 9.8*. The heating elements are arranged around the sides of the furnace and, where very uniform heating is required, also in the roof, door and hearth. Radiation from the heating elements and from the refractory lining occurs, so that the internal surface of the furnace approximates to a black-body enclosure.

Forms of work handling used with batch furnaces include bogie or car-bottom hearths and elevated hearths raised into the furnace on hydraulic rams to facilitate loading and unloading. The box furnace is normally used at temperatures where radiation is dominant. Where it is necessary to avoid exposure to air during the quenching process, the sealed quench furnace is used. Forced convection furnaces allow high heating rates and with careful design, good temperature uniformity can be achieved; but usually these are limited to maximum temperatures of 700–900°C. In the pit furnace, capable of operating at higher temperatures, radiation is the dominant mode of heat transfer and, by using a retort, the process can be carried out in a controlled atmosphere.

The bell furnace may be used as a hot-retort vacuum furnace. By reducing the pressure inside the bell, very low partial pressures of oxygen are obtained. Since heat losses by convection are greatly reduced at low pressures, the bell can be raised when the required temperature is reached, one bell then being used to heat several retorts.

The rectangular bell furnace is a form of box furnace that allows unimpeded access to the furnace hearth. It is useful for heat treating large components such as mill rolls or fragile items such as ceramics. The cold-retort vacuum furnace allows very high temperatures to be achieved. It is used for heat treatment and brazing in controlled low-pressure atmospheres using vacuum interlocks to achieve high throughputs. The hydrogen muffle furnace is used for sintering alumina at 1700°C, zone refining where a precise temperature profile is required, and the continuous annealing of wire and strip in controlled atmosphere by use of an open-ended muffle.

Continuous furnaces use a conveyor mechanism enabling in-line processes to be carried out combining heating and cooling at a controlled rate (*Figure 9.9*). Some examples of different conveyor mechanisms are illustrated in *Figure 9.10*. The choice depends on the nature of the process and the size of the workpiece. Applications range from bright annealing of fasteners to normalising steel billets.

9.5 Induction heating

Induction heating makes use of the transformer effect. The workpiece is placed in the alternating magnetic field of a coil. The field produces eddy currents in the workpiece, which is heated as a result of I^2R losses. The induced current density and consequent heating effect is always non-uniform; it depends on the magnitude and frequency of the inducing field and the physical properties of the workpiece. The current density in the workpiece is a maximum nearest to the surface adjacent to the coil conductors and, in the case of a solid workpiece in an enclosing coil, is zero at the centre. The distribution of heating in the body of the workpiece can be controlled by choice of frequency. If the frequency is high, most of the heat is developed in a thin layer, while lower frequencies give a more uniform distribution. Hysteresis loss heating also occurs in ferrous metals; it is normally small compared with the eddy-current effect but it is applicable in heating metal powders at high frequencies. Induction techniques are used for both through heating and surface heating of metallic materials at frequencies in the range 50 Hz to 1 MHz. They are used for melting and also, at very high frequencies, in the manufacture of semiconductor materials, and the hot working of glass.

The eddy-current power P per unit length of a cylindrical workpiece of diameter d, resistivity $\rho\varsigma$ and absolute

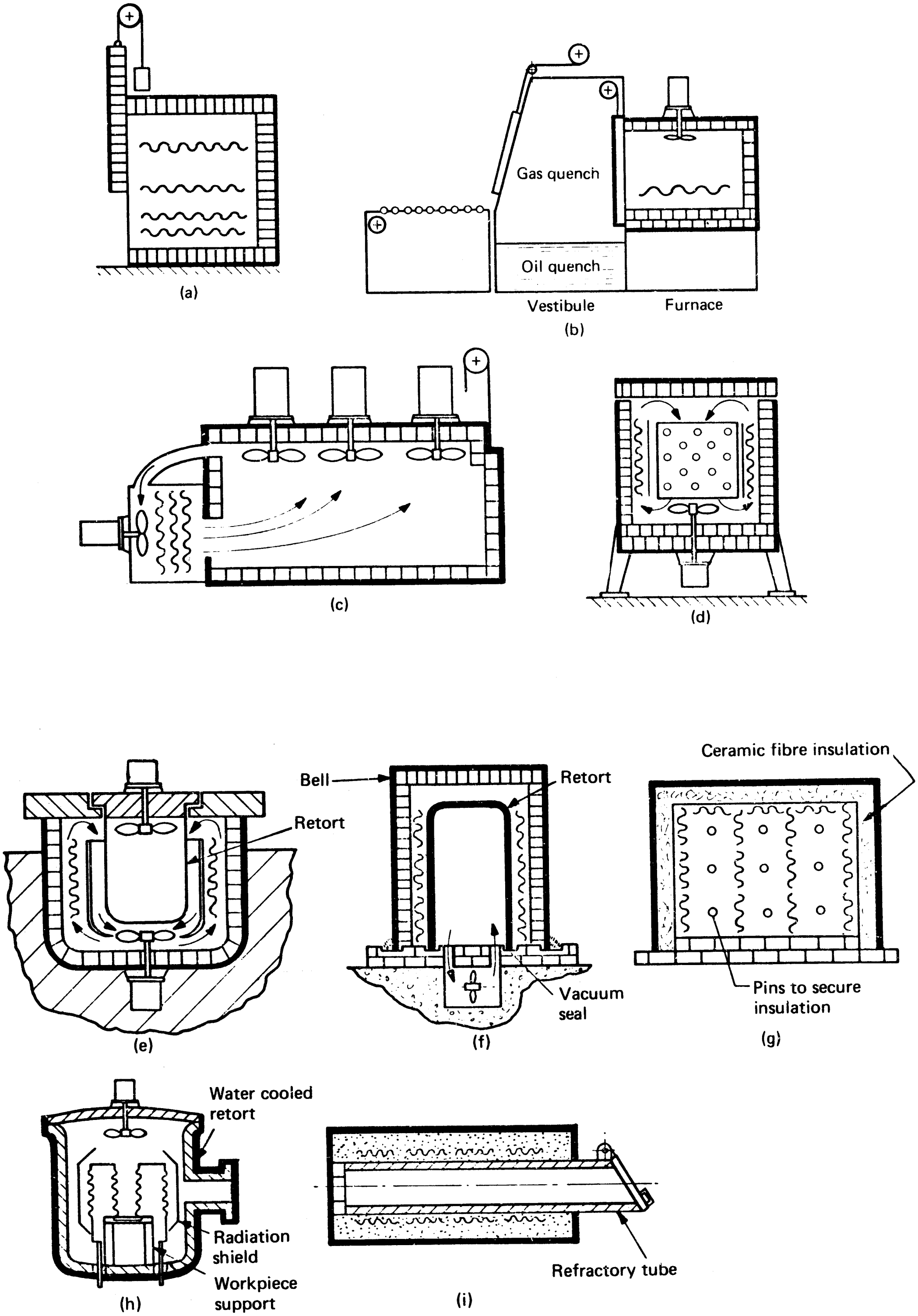

(a)
Gas quench
Oil quench
Vestibule
Furnace
(b)
(c)
(d)
Bell
Retort
Retort
(e)
Vacuum seal
(f)
Ceramic fibre insulation
Pins to secure insulation
(g)
Water cooled retort
Radiation shield
Workpiece support
(h)
Refractory tube
(i)

Figure 9.8 Batch furnaces for heat treatment: (a) box furnace; (b) sealed quench furnace; (c) horizontal forced convection furnace; (d) vertical forced convection furnace; (e) pit furnace; (f) bell (hot retort) vacuum furnace; (g) rectangular bell (low thermal mass) furnace; (h) cold retort vacuum furnace; (i) muffle furnace

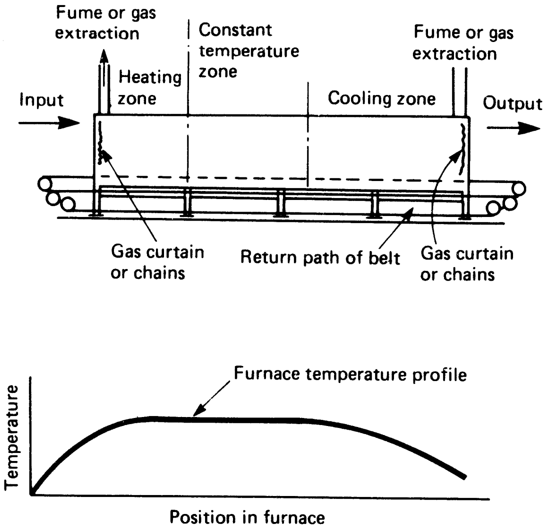

Fume or gas extraction
Constant temperature zone
Fume or gas extraction
Heating zone
Input
Cooling zone
Output
Gas curtain or chains
Return path of belt
Gas curtain or chains
Furnace temperature profile
Temperature
Position in furnace

Figure 9.9 Conveyor furnace and typical temperature profile

permeability μ, induced by an alternating field of strength H (r.m.s.) and frequency f is

$$P = H^2 \frac{\pi ς}{2}\left(\frac{d}{\delta ς}\right)^2 \rho Q \qquad (9.1)\Leftarrow$$

where $\delta ς$is the skin depth of the induced current

$$\delta ς= (\rho/\pi f \mu)^{1/2} \qquad (9.2)\Leftarrow$$

Q is a factor depending on skin depth and the shape of the workpiece, and is plotted in *Figure 9.11* for a cylinder in terms of a normalised (dimension/skin depth) parameter. Values of δ, based on room-temperature resistivities and assuming, where appropriate, a relative permeability of 100 (i.e. $\mu ς= 100\mu_0$), are given for a number of materials in *Table 9.7*. The relative permeability is a function of the applied magnetic field strength and can be as low as 10 in high-intensity heating applications. For design purposes, the value of resistivity appropriate to the average temperature of the workpiece is used. High frequencies (i.e. small skin depths) are needed if workpieces of small dimensions are to be heated.

9.5.1 Power sources for induction heating

There is still an economic advantage in being able to use power at the frequency of the mains supply rather than converting to a higher frequency. However, the increased power rating of solid-state converters, their reducing cost/unit power, ease of control and the fact that they present balanced loads to the supply system, means that each application must be carefully assessed. Loads which can be effectively heated at 50 Hz include slabs, large billets and cylinders, and process vessels. Depending on the load rating, the power input is controlled by either an off-load tap changing transformer or an autotransformer. Power-factor correction is usually provided on the primary side of the heater supply transformer, and phase-balancing networks are used to correct the imbalance of large single-phase loads. Voltage transients on the supply network are minimised by the use of soft-start arrangements when switching large loads.

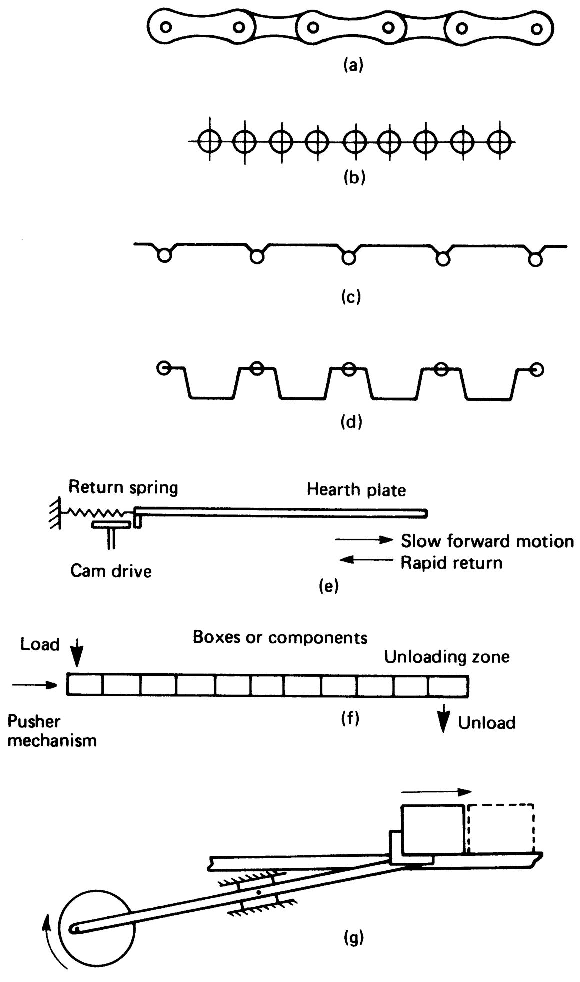

(a)
(b)
(c)
(d)
Return spring
Hearth plate
Slow forward motion
Rapid return
Cam drive
(e)
Load
Boxes or components
Unloading zone
Pusher mechanism
(f)
Unload
(g)

Figure 9.10 Some examples of conveyor furnace mechanisms. (a) Cast link conveyor; (b) Roller hearth; (c) Salt belt conveyor; (d) Slat pan conveyor; (e) Shaker hearth; (f) Pusher furnace mechanism; (g) Walking beam furnace

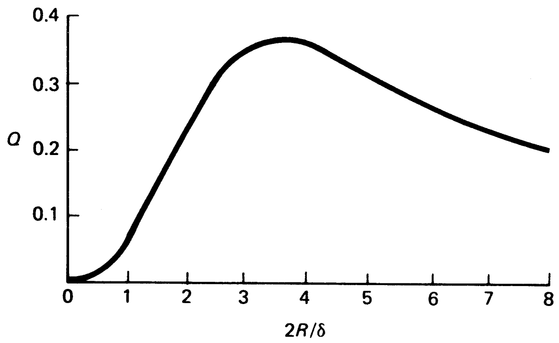


Figure 9.11 Variation of Q with the normalised diameter for a solid cylinder

Table 9.7 Typical skin depth (mm) and frequency relation

		Frequency				
*Metal**		*50 Hz*	*150 Hz*	*1 kHz*	*10 kHz*	*450 kHz*
Copper		9.3	5.4	2.1	0.66	0.098
Aluminium		11.7	6.8	2.6	0.83	0.12
Grey iron	(a)	6	3.4	1.3	0.42	0.063
	(b)	78	45	18	5.5	0.83
Steel	(a)	2.9	1.6	0.64	0.20	0.03
	(b)	71	41	16	5	0.75
Nickel	(a)	1.9	1.1	0.42	0.13	0.02
	(b)	34	20	7.6	2.4	0.36

*(a) Below Curie temperature ($\mu_r = 400$, ρ at 20°C). (b) Above Curie temperature ($\mu_r = 1$, ρ at Curie temperature).

Power supplies for frequencies above that of the mains supply are now almost invariably based on solid-state convertors. Some large installations previously utilised static frequency triplers to provide a 150 Hz supply for through heating slabs prior to rolling, derived from three one-phase saturated transformers with secondaries connected in open delta (*Figure 9.12*). Harmonics generated in the tripler are filtered by an arrangement of inductors and capacitors. A 20 MW installation of this type is operated in Canada.

Higher frequencies give faster heating rates and, because of skin-effect phenomena, enable the effective induction heating of a smaller cross-section of workpiece. Power at these frequencies, before the advent of the thyristor, was obtained from rotating machines designed to produce power at frequencies up to 10 kHz. They are normally vertically mounted and, in the larger sizes, water cooled. Conventional starting equipment is used for the induction motor. Output power is controlled by varying the d.c. excitation of the alternator field either manually or with an automatic voltage regulator. They have a fixed frequency output and present a balanced load to the supply; the efficiency is comparatively low and falls with decreasing load and increasing frequency. These machines are noisy and were often installed in sound-reducing enclosures.

The availability of high-power thyristors capable of operating at frequencies up to 50 kHz has opened up a new field to the designers of medium-frequency heating equipment. Sources based on these devices are commercially available with power outputs up to about 6 MW at the lower frequencies. The major applications are through heating and melting, and this source has the advantage over the machine of a flexible operating frequency. The efficiency is around 90%, even on partial load. The operating principle is familiar: the rectified 50 Hz supply is chopped by thyristors and fed into a resonant load circuit formed by the resistance and inductance of the loaded work coil together with a tuning capacitor. The firing pulses may be applied at a pre-set frequency; alternatively, the control signal may be derived by feedback from the load circuit. Principal variants of this basic arrangement are commonly referred to as *voltage source, current source* and *variable mark/space ratio generators*. A fourth variation, the cyclo-converter, converts the 50 Hz supply directly to the higher frequencies without the intermediate d.c. stage.

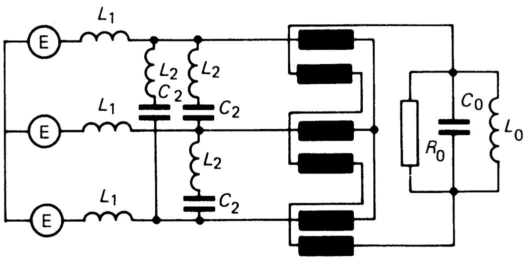


Figure 9.12 Three-phase transformer with open delta secondary for generation of the third harmonic

At frequencies above 50 kHz, the power source uses a vacuum triode feeding into a tank circuit of which the load forms a part (*Figure 9.13*) either directly or through a coupling transformer. Water-cooled inductors are used in the tank circuit and, in all but the smallest ratings, the valve is also water cooled. Conventional industrial valves have been and still are used for this application in ratings of up to 500 kW operating in the class C mode with conversion efficiencies greater than 50%. Even higher efficiencies are obtained with the magnetically beamed triode, which has an inherent low grid dissipation. This valve is robust and offers more flexible control than the conventional unit, owing to the lower grid power requirements which enable semiconductor control circuits to be used. Closed-loop control of the process variables, i.e. power and temperature, is possible.

An important version of the triode based r.f. generator uses a high efficiency coupling transformer in the triode circuit, with the tuning capacitors connected directly across the work coil. This 'aperiodic' generator can operate over a wide frequency range, supplying remote workstations at a higher efficiency than the more conventional oscillator.

9.5.2 Load matching

The matching of the loaded work coil to the source is extremely important in the successful application of induction heating processes. The impedance of the work coil depends on the geometry and physical properties of the workpiece, the frequency of operation, and the geometry and number of turns of the coil. A matching transformer may be required if the correct impedance cannot be achieved by selection of coil turns. The rating of the power source is minimised by tuning the reactive component of the load impedance with capacitors. These may be connected in series or parallel directly with the load coil, or on the primary side of the matching transformer, depending on the coil voltage and type of power source.

The load impedance will change during a heating cycle which might necessitate re-tuning or re-matching to the source.

9.5.3 Coil design

Coil designs depend on the operating frequency and the application. The basic objectives are to produce the

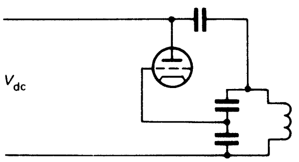


Figure 9.13 Basic oscillator circuit for r.f. induction heaters

required magnetic field strength over that section of the workpiece to be heated and to insulate the coil both to prevent electrical breakdown and to reduce heat transfer from the workpiece to the coil. The heat generated by the I^2R loss in the coil usually requires some form of forced cooling, and it is customary to use hollow water-cooled copper tubular conductors. The current distribution in the coil conductors is non-uniform. It is frequency dependent, and affected by the proximity of adjacent conductors, the presence of the workpiece and the geometry of the winding.

Although most units employ only one layer of winding, there is now widespread use of multilayer windings for the heating of non-ferrous billets at supply frequency. With the correct choice of conductor cross-section, multilayer coils offer significant energy savings in these applications.

9.5.4 Through heating of billets and slabs

Induction heating is used extensively for the through heating of both ferrous and non-ferrous metal billets prior to rolling or forging. The billets, of circular or rectangular cross-section, are either heated individually or passed in line through a series of induction coils. The frequency of the current, the power input and the length of time in the coil are chosen to give the required throughput rate with an acceptable temperature distribution over the cross-section of the workpiece. A compromise is sought between high heating rates produced by high frequencies, and an acceptable skin depth. A typical example is a 350 kW, 1 kHz unit for heating steel transmission forgings to a temperature of 1150°C; the energy consumption is of the order of 400 kWh/t and the heating time can be less than 5 min. Heaters with power outputs of up to 6 MW are readily available and, although the frequency for most applications is in the range 50 Hz to 3 kHz, higher values (up to 50 kHz) are occasionally employed. The efficiency of the process is a function of the coil design and the coupling between it and the workpiece. In some instances a tapered coil has been used to allow for the changing parameters of the workpiece material as the temperature increases during the process. The workpiece resistivity at the final working temperature may be four or five times that at 20°C and, in the case of ferromagnetic materials, the relative permeability will fail to unity when the Curie temperature is reached.

Metal slabs are also heated by induction processes. One of the largest single installations is in the USA, where a 200 MW, 60 Hz power supply heats large steel slabs prior to rolling, at a maximum rate of 600 t/h. Average energy consumption is reported as 340 kWh/t for a 22 t slab. A more recent installation in Sweden rated at 37 MW heats 15 t slabs, for rolling, at the rate of 85 t/h. Thin slabs, from continuous casting machines or at an intermediate rolling state, are heated at medium frequency.

9.5.5 Strip heating

The heating of continuous-strip materials by the conventional induction method requires the use of frequencies above 10 kHz and efficiencies are low for non-ferrous materials. A preferred technique is the transverse flux method where, for example, efficiencies above 70% can be achieved heating thin aluminium strip at 250 Hz. The strip is passed between two flat inductors comprised of windings in a laminated iron, or ferrite, core which form a series of magnetic poles. The flux passes transversely through the sheet and currents are induced in the plane of the sheet. The winding and pole arrangement must be designed to give uniform heat distribution over the moving strip. Installations rated at 1.8 and 2.8 MW are operating in Japan and Belgium, respectively, for the heat treatment of aluminium strip.

Development is also being carried out on the travelling-wave induction heater, which utilises a three-phase winding operating at 50 Hz similar in construction to that of a linear motor with the secondary (workpiece) held stationary. Applications of interest are the heating of metal cylinders and tanks; this technique could also be used for the heating of sheets and slabs.

9.5.6 Surface and localised heating

Induction heating at high frequencies (up to 500 kHz) is used extensively in the engineering manufacturing industries for the hardening of bearing surfaces, welding and brazing and similar surface heat treatment processes. The workpiece is placed in a coil (*Figure 9.14*) designed to cover the area to be heated. Heating takes place within a few seconds and, owing to the high frequencies, is confined in the earlier stages to a thin surface layer. The heated area is then cooled rapidly, so that the required surface hardening effect is obtained while still leaving the centre ductile. Contoured shapes can be heated, as can the internal surfaces of tubes, etc., with suitably designed coils. Flat spiral coils are used for the treatment of localised areas in sheet materials. An extension of the application is to soldering and brazing in which case the field distribution is modified by the use of a field intensifier, a suitably shaped piece of magnetic material such as a ferrite core. The technique is also used for seam welding of steel tubes from flat strip, bent into shape and the two edges welded. The field and, therefore, the heat is concentrated at the two surfaces which are to be welded. In this latter application high throughput rates are achieved by using power sources with ratings as high as 1 MW at 450 kHz.

9.5.7 Semiconductor manufacture

Since the energy from the heating coil can be generated in the workpiece without any heat transfer medium, the whole process can be carried out in a vacuum chamber. This is particularly useful in the manufacture of semiconductor materials. In one technique the material is placed in a conducting crucible in the vacuum space and is heated by induction. Semiconductor materials can also be heated directly by induction without the need for the crucible: in this case very high frequencies up to 4 MHz are used.

9.5.8 Indirect induction heating for non-metals

Induction heating is generally used to heat the 'work' directly, that is the induced currents flow in the electrically conducting object being heated. However, there is an important and growing series of applications where induction heating is used to heat a vessel or container from which heat is transferred by conduction to a non-metal product. Some examples include:

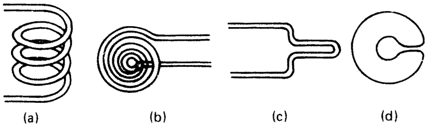


Figure 9.14 Coils for surface hardening by induction heating: (a) cylindrical coil for shaft hardening; (b) pancake coil for heating flat surfaces; (c) hair-pin coil for localised heating; (d) current concentrator

(1) induction heating of a vessel containing a solid or liquid;
(2) induction heating of a pipe through which a liquid flows;
(3) induction heating of a chemical reactor vessel; and
(4) induction heated extruders.

These techniques often employ the same coil arrangements as have been discussed elsewhere, but there are some recent developments using novel induction techniques.

An example of such a novel technique is ROTEK, a rotary kiln heated by an unusual form of induction heating and used for drying or calcining granular solids which flow continuously through the revolving kiln drum. The kiln lining and agitating/lifting flights are made from stainless steel or Inconel and heated by the passage of a very large electric current flowing from end to end; this current is too large (perhaps 20 kA or more) to be fed through sliding contacts and the drum therefore forms a single-turn secondary of a ring transformer with a bar primary which passes through its centre (see *Figure 9.15*). The bar primary is fed from a low-voltage transformer and the efficiency of the device as a heater is very high. Because the ring core can be thermally insulated, the operating temperature can be well above the Curie point and, with an Inconel drum, temperatures exceeding 1000°C can be achieved.

9.6 Metal melting

9.6.1 Introduction

Before considering the types of electric furnaces used today for melting metals, it is useful to review the options available for converting electrical energy into thermal energy contained within the melt. Essentially two approaches are possible. The most direct approach is to employ Joule heating within the metal to be heated. This approach is used in the coreless induction furnace and the channel induction furnace.

The alternative approach is to convert the electrical energy into a heat flux on the metal surface. This can be done in a number of ways. The thermal energy can be derived from an electric arc (either at the metal surface or sufficiently adjacent to it), from an electric plasma (which may or may not be transferred to the surface of the metal) or from the Joule heating of a resistor adjacent to the metal. The resistor may be molten slag on the metal surface, or a crucible containing the metal, or resistance elements adjacent to the metal or its containment crucible. Electron beam and laser heating techniques are not, as yet, widely used for bulk melting operations, although they are finding increasing application on a small scale, for example in the local melting required for fusion welding.

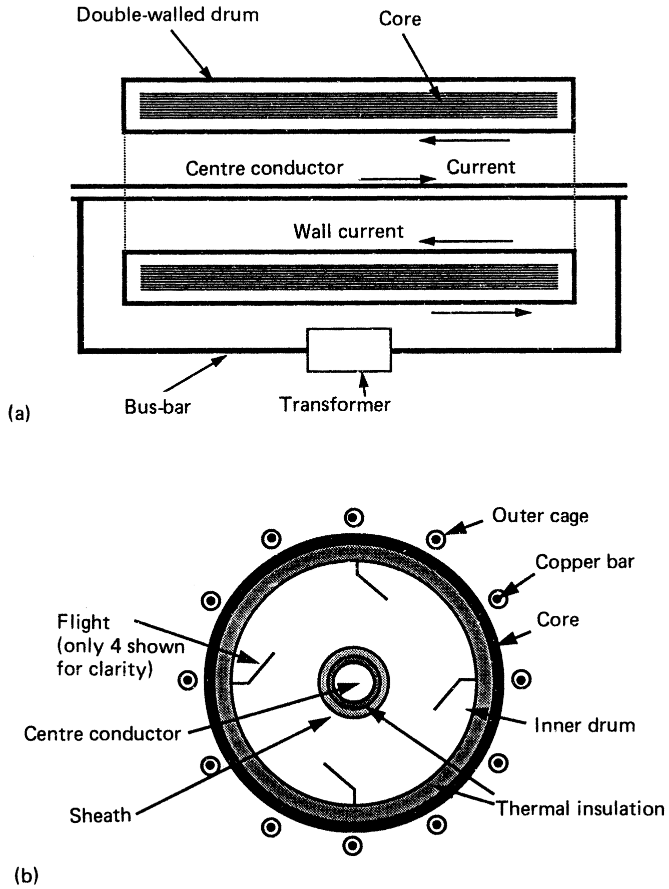


Figure 9.15 ROTEK: (a) operating principle; (b) cross-section of actual machine

In recent years medium frequency coreless induction furnaces have become widely used for a vast number of ferrous and non-ferrous applications. Resistance-heated furnaces have also become firmly established for melting and holding furnaces in the aluminium and zinc casting industries.

9.6.2 Arc furnace

The electric-arc furnace is largely used for the melting of steel scrap to produce liquid steel. Essentially, it consists of a squat cylindrical vessel, refractory lined, with a movable domed refractory roof, as shown in *Figure 9.16*. The furnace is usually charged from a drop bottom bucket. Each of the three graphite electrodes, which pass through holes in the roof, is fed from one phase of the three-phase furnace transformer via flexible cables and bus tubes connecting onto the electrodes. The metal charge forms the star point of the furnace transformer.

Heat-transfer efficiency is high when the arcs are surrounded by solid charge which absorbs the largely radiant energy of the arcs. As the charge melts, the arcs radiate onto the surrounding refractories (which are nowadays largely replaced by water-cooled panels, to reduce overall operating costs), and the thermal efficiency deteriorates because of this. Conversion efficiency of electrical energy into heat in the steel of around 94% is attained.

Electromagnetic stirring may be incorporated by using a non-magnetic steel shell and incorporating a low-frequency stirring coil below the furnace bath. The entire structure, including electrodes, masts, etc., is normally mounted on a hydraulically operated rack and pinion enabling it to be tilted in one direction for pouring and in the reverse direction for slagging. The roof structure is pivoted so that it can be swung aside (with the electrodes raised) for charging. The electrodes can be slipped in the clamp to allow for electrode wear.

The electrode arms can also be raised and lowered individually by hydraulic or winch systems and the electrode height above the melt is controlled by feedback signals derived from the arc voltage and current.

The substation for a large arc furnace is normally adjacent to the furnace itself and contains the furnace transformer, which normally has a star connected primary and an input voltage of 33 kV. The transformer must withstand very large electromechanical forces produced by the high short-circuit currents; it is oil cooled and has terminations brought out to which the flexible cables are connected. The furnace power is varied by on-load tap-changing. Electrical contact to the electrodes is made by a large copper pad contained in the electrode clamp connected in water-cooled bus-bars which rise and fall with the furnace electrodes. Various configurations have been adopted to ensure that the geometry remains as nearly symmetrical as possible, independent of the bus-bars, to minimise out-of-balance currents. The furnace electrodes are normally connected in delta, and where very high currents are used, the delta is closed at the electrode clamp in order to minimise the effects or reactance in the transformer secondary circuit.

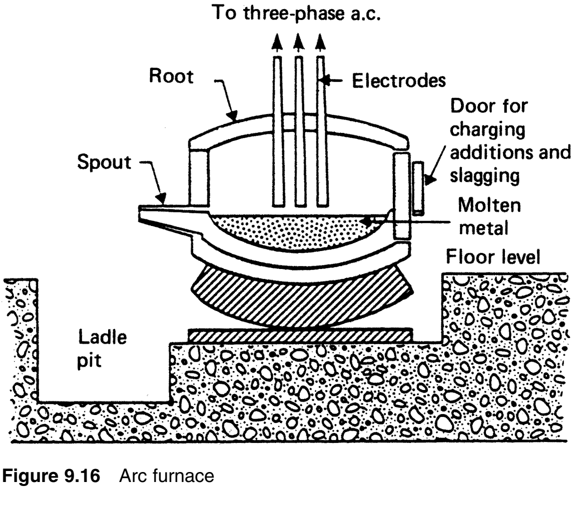


Figure 9.16 Arc furnace

As an electrical load, the arc furnace is less than ideal. The arc can vary between extinction, at zero current, to the short-circuit condition, caused by scrap contacting the electrode. Physical movement of both the solid charge pieces and the melt, when formed, causes variation in the arc length which can fluctuate many times within a second, and superimposed on these effects is arc-length variation caused by the mechanical vibration of the electrode and its supporting structure.

These effects prevent a constant high power level being delivered by each of the phases, and this results in:

(1) reduced melting rate attained for a given transformer capability;
(2) the occurrence of current fluctuations which are reflected back into the supply system, these occurring especially in the early stages of the melting operation;
(3) voltage fluctuations at the point of common coupling to other electricity users; and
(4) acoustic noise generation from the furnace.

A typical large steelmaking arc furnace is of around 100 t melt capacity and powered at perhaps 50 MVA. No other available electric furnace can match the sort of melting capability attainable from such a single unit, and the furnace, therefore, is likely to remain the only electrical alternative for the tonnages currently required by the steel industry.

9.6.2.1 Submerged-arc process

The submerged-arc process is not essentially an arc process as heating occurs also by direct resistance with, perhaps, some limited heating from arcs and sparks during interruption of the current path. The principal applications are for reducing highly endothermic ferroalloys of high melting point, such as ferromanganese, nickel, chrome, silicon, tungsten and molybdenum, which are subsequently remelted in arc furnaces to produce special alloys. Fused oxides may also be produced via the submerged arc route.

The design of a submerged arc furnace depends on its application. In principle, it is a dished vessel (*Figure 9.17*),

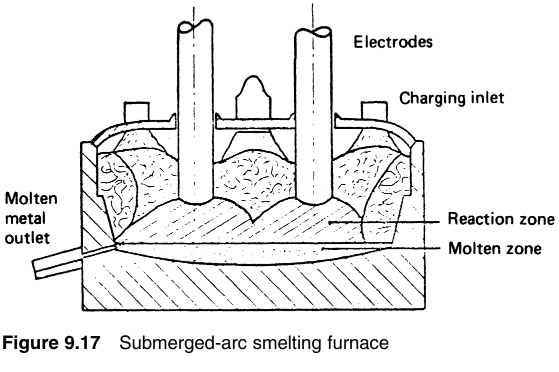


Figure 9.17 Submerged-arc smelting furnace

brick lined, as with the arc furnace. But there the similarity ends: the dish and the roof are axially fixed, although the roof, together with the electrodes, may rotate. The furnace is charged through ports in the roof, and molten metal and slag flow from the furnace continuously. The electrodes are of the Soderberg type, formed *in situ* by pouring a mixture of pitch and tar plus anthracite into a steel tubular shell. The process is carried out several metres above the furnace, and as the electrode is lowered, it bakes, so driving off the volatile binding. By the time it enters the furnace it is a solid mass. Electrodes capable of carrying very high currents up to 120 kA, can be produced in this way.

9.6.2.2 Vacuum arc furnace

The vacuum arc furnace (*Figure 9.18*) is used primarily for remelting metals of very high quality, including titanium, tantalum, niobium, hafnium, molybdenum, tungsten, zirconium, and for refining some steel and nickel alloys. Ingots of up to 100 t can be produced. The furnace operates at low pressure, down to about 0.01 Pa, and very effective degassing of the droplets of molten metal (which have a high surface area) occurs. The ingot forms a molten 'skull' which freezes in contact with the copper mould, thus eliminating contamination from refractory linings and the relatively low volume of liquid present minimises thermal stresses and piping at the ends. Impurities either evaporate or collect in the molten pool on the ingot surface. The electrode is either prefrabricated or melted first in a vacuum induction furnace; the arc is d.c. and operates with a current of 10–25 kA and a voltage of 20–30 V. A low voltage is used to prevent the arc attaching to the walls of the vessel, and an additional field coil, which interacts with any radial component of arc current, tends to help the stabilising effect and produces a strong stirring action.

9.6.2.3 Electroslag refining

Electroslag refining (*Figure 9.19*) is directly competitive with vacuum arc processes for materials not unduly reactive in air. A high degree of refining can be obtained, since the droplets of molten metal penetrate through the molten slag, enabling desulphurising to be carried out and oxide inclusions to be removed. The process, like the vacuum arc furnace, forms a molten pool on the solidifying ingot and has similar advantages.

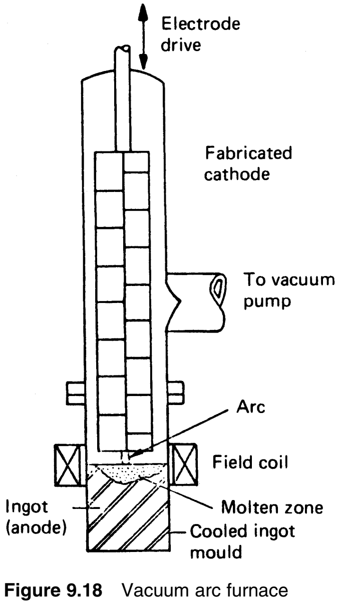


Figure 9.18 Vacuum arc furnace

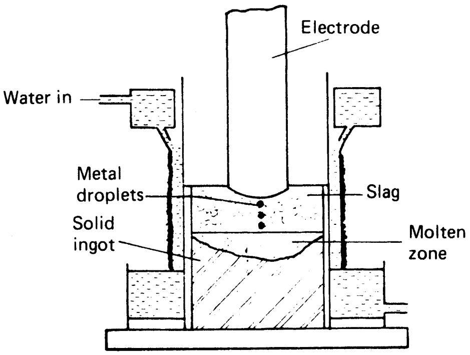


Figure 9.19 Electroslag refining furnace

The electroslag refining process is essentially one of resistance heating, since it relies on electrical conduction in the molten slag. Single- or three-phase operation (using three electrodes over one ingot) is possible. The operating voltage is kept in the range 40–60 V to prevent formation of arcs, and currents of up to 3 kA are used. Operation is controlled by limiting the voltage and keeping the electrode immersed in the electrically conducting slag in order to avoid drawing an arc. The slag depth and composition are carefully controlled to maintain its electrical conductivity and to allow it to refine the molten metal droplets. Cylindrical and slab ingots can be produced: typical ingot diameters are 350 mm (single-phase) and 900 mm (three-phase), with a melting rate of 180–360 kg/h at 360 kVA (single-phase) or up to 6 MVA at 180 V and 18 kA (three-phase). Large ingots can be produced by switching electrodes during refining.

9.6.2.4 Electron-beam furnace

Electron beams are used for welding, melting and the production of evaporated coatings. The beam is obtained from a heated filament or plate and is accelerated in an electron gun by a high electric field produced by one or more annular anodes. Electrons on the axis of the gun pass through the final anode at very high velocities (e.g. 8.5×10^7 m/s at 20 kV). The electron gun and chamber are kept at a low pressure of around 10^{-3} Pa and, as little energy is lost from scatter or production of secondary electrons, practically all the kinetic energy of the beam is converted to heat at the workpiece. Thus the conversion efficiency of electrical energy input to thermal energy in the workpiece is very high. The electron-beam furnace (*Figure 9.20*) utilises a cooled ingot mould in the same way as the vacuum and electroslag furnaces.

Ingots, slabs, tubes, castings, pellets and powder can be produced. One system, shown in *Figure 9.20*, comprises one, two or three guns arranged around a consumable electrode. Individual power ratings up to 400 kW are possible, which enables total power inputs of up to 1.2 MW and melting rates of 500 kg/h to be obtained.

9.6.3 Coreless induction furnace

The coreless furnace essentially consists of a refractory crucible encircled by a solenoid coil excited from a single-phase

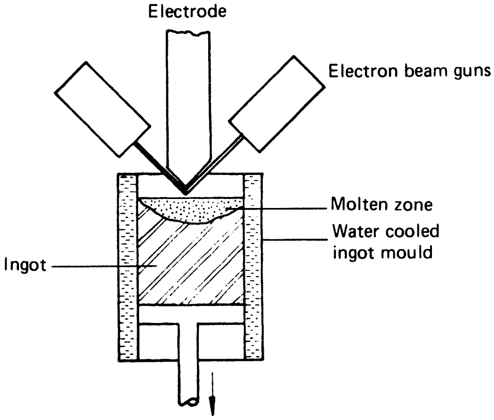


Figure 9.20 Electron-beam melting furnace

a.c. supply (*Figure 9.21*). The fluctuating axial magnetic field linking the charge within the crucible causes I^2R heating within it. The power induced in the charge depends on the physical properties of the material, the magnitude and frequency of the flux linking it and its geometrical shape. For efficient energy transfer to the load, the cross-section of the current paths must be greater than three times the penetration depth of the electromagnetic wave in the material.

This penetration depth δ is given by equation (9.2). Thus, to melt individual components of charge contained in the crucible, the components must have dimensions lying across the coil axis of 3δ or the voltage generated between the individual components must be sufficiently high to establish current paths of sufficiently large diameters within the bulk of the charge.

Depending on the resistivity of the material being melted, the coreless furnace converts electrical energy to heat in the charge at an efficiency of between 60 and 85%. It is useful to note that, for most furnace designs, the total circuit reactance is far greater than the resistance (by as much as an order of magnitude) and the largest reactive component is caused by the separation of the coil from the metal charge. The change in the reactance of the charge causes the total circuit reactance to change by about 30% when melting typical iron scrap charges. In addition, lining wear or build-up of slag or dross on the walls of the crucible typically alter the total circuit reactance by a further 30%. It is also seen that, as the frequency of the supply increases, there is a reduction in the circuit power factor. The furnace efficiency, however, is independent of frequency, although it is improved by a high charge permeability such as when melting ferromagnetic materials, e.g. steel scrap.

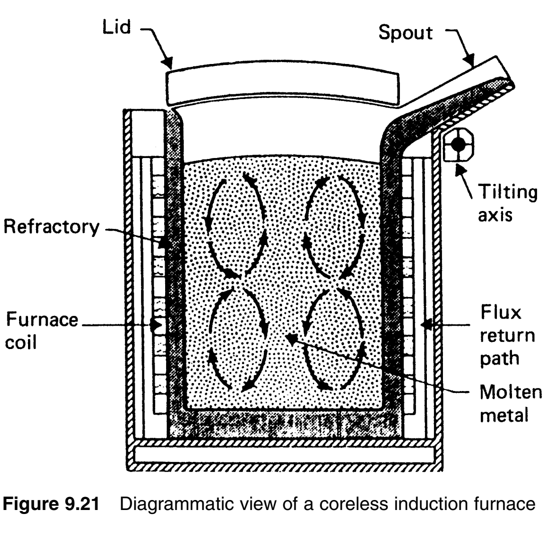


Figure 9.21 Diagrammatic view of a coreless induction furnace

Water cooling of the induction coil is required to remove the I^2R losses from it—the resulting steep thermal gradient through the crucible wall then improves its integrity in maintaining a separation of the melt from the induction coil. Normally, the coil is manufactured from a thick-walled high-conductivity copper tube, and the design is a compromise between mechanical, electrical and hydraulic requirements. The coil, in the most common modern designs is wound with electrical insulation and the turns are separated using studs, fastened to each turn and arranged in vertical rows, held in place by hardwood pillars. This coil assembly is then restrained vertically and radially by the furnace structure.

For small furnaces, the furnace body can be made from high-strength insulation boards or aluminium side-panels interconnected by non-metallic components, but for larger furnaces (say over 1 t iron capacity) the furnace body is made from a steel frame or from rolled-steel plate. In this case, an electromagnetic screen must be interposed between the coil and its supporting structure. This improves the power factor and coil efficiency compared to the unscreened smaller furnace design, but the lamination packs introduce a small iron loss component.

In recent years, the capital cost of coreless melting furnaces has been significantly reduced in real-money terms, for a given melting rate. This has been due to the development of high-power-density furnaces fed at medium frequency via the solid-state inverter.

Inverter power sources are designed to allow the output frequency to alter to maintain tuning to the natural frequency of the coil system. Thus, capacitor switching is eliminated and the power applied to the furnace is dependent solely on the limits of current and voltage within the inverter, these being chosen to allow substantially constant power input to be achieved throughout the melting of a cold charge, the load conditions of which will change during the total melt cycle.

Inverters used are of the a.c. to d.c. type, and conversion efficiencies of mains to medium-frequency power of over 95% are achieved. Two circuit designs compete in the furnace market, both being capable of giving similar melting efficiencies. These are the current-fed inverter (parallel inverter) and the voltage-fed inverter (series inverter).

The current-fed inverter operates at a leading output power factor and follows the load resonance of the circuit, control of power being by phase-angle control of the inverter rectifier stage. The voltage-source inverter operates slightly below the resonant tank frequency, and power control is achieved by moving the operating frequency away from the resonant condition, allowing a fixed rectification input stage to be used.

The load characteristics imposed on the supply are determined by the rectifier input stage, the harmonics generated being determined by the pulse number of the rectifier. For six-pulse inverters, the predominant harmonics are the fifth and seventh, and for 12-pulse the eleventh and thirteenth, respectively.

Six- or 12-pulse input stages are used, depending on the supply requirements. Individual rectifiers are rated up to

2 MW, and for higher ratings units are connected in parallel. The voltage-source inverter presents a fixed input power factor to the supply but the current-fed inverter, which uses a controlled rectifier input stage, has a load-dependent input power factor. Inverter bridges up to 1 MW from four devices are used and, to achieve higher ratings, devices are connected in series or parallel. Circuit protection is important for long-term inverter reliability, the current-source inverter being easier to protect because of the absence of the high stored energy which must be dissipated in the voltage-source inverter.

The magnetic field reacting with the induced currents in the melt causes forces to be applied to the liquid metal in the furnace, resulting in a surface stirring action which can be advantageous in assisting with dissolution of alloy additions or newly added charge materials. The magnetic pressure creates a metal surface which is dome shaped and the height of this dome (also termed the meniscus), ultimately sets a limit to the power density which can be tolerated. The limiting power density is sometimes empirically taken as linearly related to (frequency)$^{1/2}$.

For molten iron, the limit to power density is around 300 kW/t for a mains-frequency furnace, although the meniscus lift can be dampened by using a larger ferrostatic head above the power coil. By comparison, power densities of up to 750 kW/t are acceptable when using medium frequencies, in the range 180 Hz to 10 kHz. However, metal surface movement markedly increases at the lower frequencies, benefiting assimilation of alloys and light materials. Thus, where this is a significant factor, frequencies of around 250 Hz or even lower are preferred. As lower frequencies are chosen, the voltage generated between individual charge pieces becomes less (for a given furnace power and size), since this is related to the voltage per turn of the furnace coil, which is proportional to $f^{3/4}$. Thus, the lower frequency furnaces are more dependent on the characteristics of the charge pieces than furnaces of higher frequency, and hence choice of frequency has to be a compromise, at least for a given furnace rating and size, and should take into account the required surface stirring action.

Since the first application of large, high-power-density medium-frequency coreless furnaces in the UK (around 1980), the iron-foundry trade has chosen this type of furnace for virtually all new electric melting schemes.

In common with the iron-foundry sector, the lower cost, greater flexibility, and higher power density advantages of modern medium-frequency induction furnaces have tended to eliminate mains-frequency furnaces from consideration for the majority of coreless melting applications in the non-ferrous metals sector.

The low heat requirements to melt lead and zinc generally militate against induction melting, whereas the high melting temperatures of copper alloys are not as efficiently met by other fuel types. Undoubtedly aluminium is the metal which is melted in the widest range of furnaces employing the widest range of fuels and melting practices. There has been significant growth of induction furnaces particularly for recycling in furnaces varying from 0.5 to 8 t capacity.

In the foundry sector several variations of the basic coreless furnace can be found. For example, in furnaces of up to 1 t capacity melting copper and its alloys, parallel-sided clay graphite or silicon carbide crucibles may be used in preference to conventionally refractory-lined furnaces, although consideration must be given to the heat generated directly in these graphite-based materials, particularly the more highly conducting carbon bonded silicon carbide types. This is of special importance in free-standing crucible furnaces generally in the form of push-out or, more recently, by drop coil designs. These furnaces have been rapidly accepted by the foundry sector because of the ability to melt small batches of metal very quickly, e.g. 60 kg in 13–14 min. At the end of a melt, crucibles can be removed from the vicinity of the power coil and carried directly to the casting mould. Inter-melt contamination is avoided by using different crucibles for each alloy type.

9.6.4 Channel induction furnace

During the 1970s, the channel furnace became accepted as a melting unit for cast iron, and this acceptance was largely due to the improved refractory technology allowing useful levels of power to be applied. Up until then, in iron foundries, the furnace had mainly found application for metal holding.

Electrically, a mains-frequency solenoid coil encircles one leg of a lamination pack and the resultant alternating magnetic field induces Joule heating in a loop of molten metal which surrounds the coil, this metal loop acting as a single-turn secondary of the transformer (*Figure 9.22*). The close coupling between the induction coil and the metal loop provides a higher natural power factor than occurs with the coreless induction coil, and the metal loop provides a higher natural power factor than occurs with the coreless induction furnace. The efficiency of converting electrical energy into heat in the metal is therefore also higher, being above 95%. The molten-metal loop needs to be continuous, and it is not allowed to solidify in case this continuity is broken during contraction. The loop is contained by refractory, enclosed usually within a rigid-steel inductor box, and this can be separated from the main furnace body to allow refurbishment of the inductor box refractory. To mechanically support the refractory between the loop and the induction coil, a bushing, which may be a water-cooled jacket, is interposed (electrically divided into two). The cooling system is designed to produce a steep thermal gradient in the refractory, so improving its integrity.

For melting cast iron, the furnace normally consists of a vertical cylindrical bath having the inductor box attached to its base. The refractory in the main bath is able to be heavily thermally insulated.

The melting power can be applied continuously during the whole of the charging period, and this provides a very consistent and repeatable melting routine.

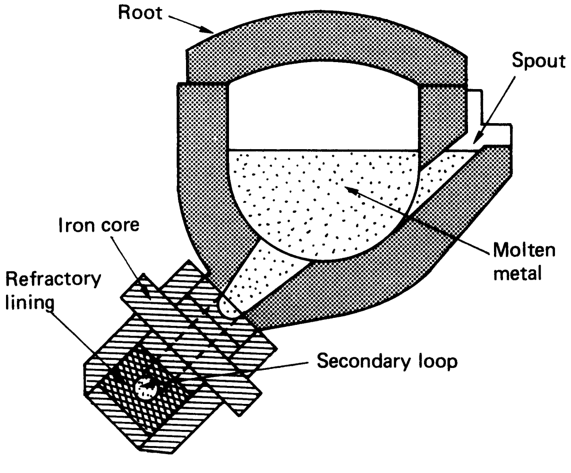


Figure 9.22 Diagrammatic view of a channel induction furnace

The use of channel furnaces for aluminium melting has been constrained by problems of oxide build up in the loop necessitating special designs to allow easy cleaning. In recent years European furnace companies and research workers alike have attempted to produce high-power designs of up to 1.3 MW per inductor with freedom from loop blockage for bulk melting applications, thus utilising the high electrical efficiency of the furnace to minimise melting costs.

9.6.5 Resistance furnaces

Electric radiant-heating techniques for metal melting are most widely employed in the zinc and aluminium foundry sectors with only minor penetration into the higher melting point copper base alloy market. In general, resistance heating is used where metal demand patterns are not appropriate to induction furnaces and hence should not be considered as a competitive technique. For example, furnaces for aluminium range from bale-out-crucible furnaces of 25 kg capacity to well insulated box-like receivers of over 10 t capacity.

Resistor elements may be in the form of wire (e.g. iron–chromium–aluminium alloys) which can be mounted in various configurations, or for higher temperature applications silicon carbide rods either of conventional or spirally cut tube section.

A particular advantage of electric resistance heating is superior metal temperature control, achieved by feedback to the power supply. The simplest power supply comprises direct on-line switching of, for example, a three-phase 415 V mains supply feeding a star-connected element configuration, control being achieved by means of a contactor. Although this is acceptable for wire elements, time-dependent resistance changes of silicon carbide elements (ageing) usually demand voltage regulation which can be effected by a tapped transformer. At power inputs of 50 kW it often becomes more economical to use thyristor circuitry controlling power by means of phase-angle control or burst firing. In the former case, the resulting waveform is not purely sinusoidal and harmonics can be developed at odd integer values of the fundamental (third, fifth, seventh, etc.). A three-phase open delta system (or six-wire load) is usually favoured for silicon carbide element furnaces for two reasons:

(1) to minimise the presence of triple harmonics in the supply lines; and
(2) to allow independent control of each phase load in the event of, for example, partial load failure.

In this configuration, harmonics create r.m.s. current levels in excess of those normally expected for a purely resistive load and, therefore, cable sizes need to be increased.

Burst firing does not generate the type of harmonics described above, but can cause flicker on tungsten filament lamps discernible to the human eye. This is due to particular combinations of voltage fluctuation caused by load ‘steps’ or bursts, and their frequency. Undue voltage fluctuation is minimised by careful selection of supply cable impedance. Thyristor control may also be used in preference to direct on-line switching in situations where close temperature control is required.

Crucible furnace designs comprise a steel shell containing insulating material, heating elements and a carbon based crucible located in the central chamber (*Figure 9.23*). The highest ratings are around 120 kW, giving a maximum melting rate of 230 kg/h (aluminium) in a 600 kg capacity furnace. Crucible furnaces exist in both tilting and bale-out configurations, and great strides have been made to improve performance figures in the light of competition between manufacturers and from other fuels. For example, a typical 180 kg capacity bale-out furnace has a holding requirement of about 4 kW with advantages of automatic start-up, good working environment and close temperature control.

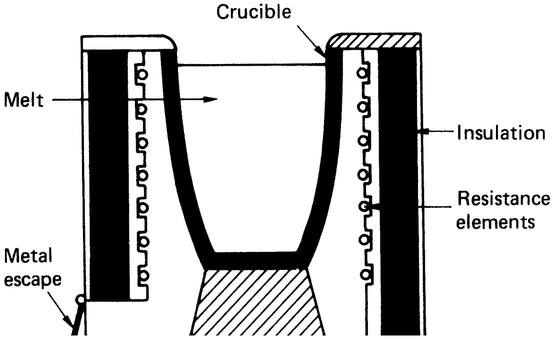


Figure 9.23 Diagrammatic view of a resistance crucible furnace

Bath-type furnaces, using roof-mounted elements radiating onto a refractory lined bath of metal, are available for a wide range of metal melting and, more commonly, metal storage applications. The rate of heat transfer to the melt is limited by the low emissivity of the melt surface, a factor which, conversely, is of considerable advantage in reducing heat losses during normal bale-out applications. Since heat transfer is also proportional to bath surface area, significant melt rates are achieved only with high-capacity furnaces. Typically, the metal content is around eight times the optimum hourly melt rate achieved when these large structures have reached thermal equilibrium.

For bulk storage and dispensing applications, good thermal insulation techniques have led to the development of integrated systems in which low-power mineral-insulated elements operate around 2–3 kW/t of metal capacity.

Immersion heating using silicon carbide or wire elements in silicon carbide based tubes provides an alternative high-efficiency heating mode for lower melting point metals.

9.7 Dielectric heating

Processing of non-metals by the conventional heating techniques of convection, conduction and radiation is often limited by the physical characteristics of the materials such as thermal conductivity and temperature sensitivity. Furthermore, since all these modes of heating depend on the transfer of heat through the product surfaces there is inevitably a temperature gradient between the surface and the centre, a problem which increases as the thickness of the product increases. In many operations this is undesirable, leading as it does to process inefficiencies and product degradation. Dielectric heating which covers both the r.f. and microwave parts of the electromagnetic spectrum can, in many circumstances, alleviate these problems by virtue of its ‘volumetric’ or at least deep penetration form of heat transfer. The industrial applications of dielectric heating are many and varied ranging from drying of textiles and the welding of plastics to the thawing of meat and the heating of rubber extrusions. Dielectric heating depends on a number of polarisation effects, the most commonly described one being dipole orientational polarisation (*Figure 9.24*). This is very important at microwave frequencies, but of relatively little significance at the lower, radiofrequencies.

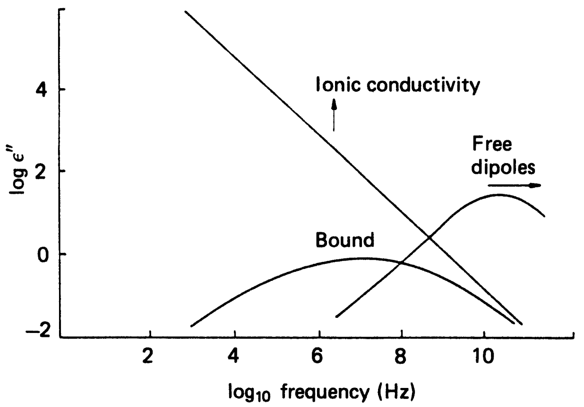


Figure 9.24 Dielectric properties of water: variation with frequency and temperature. Arrows indicate change with increase of temperature

The dominant mode in the r.f. range is space charge orientation, which in turn is dependent on the ionic conductivity of the material being processed. Therefore, for a particular material it is possible in theory to choose the most appropriate frequency from those available in the industrial, scientific and medical (ISM) bands (*Table 9.8*). In reality, many products can be processed by either r.f. or microwave and the choice of which can be made on other considerations such as the engineering required to make a satisfactory enclosure compatible with the process line requirements. The heat transferred per unit volume of product is given by

$$P = 2\pi f \varepsilon_0 \varepsilon''_r E^2 (\mathrm{W/m^3}) \qquad (9.3)\Leftarrow$$

where f is the frequency (MHz), E is the electric field strength (V/m), and ε_s is the loss factor or relative permittivity.

Table 9.8 Frequency allocation for industrial, scientific and medical purposes (existing allocations, 1980)

Frequency (MHz)	*Frequency tolerance* (±)	*Area permitted*
0.007	10 kHz	USSR
13.56	0.05%	World-wide
27.12	0.6%	World-wide
40.68	0.05%	World-wide
42, 49, 56, 61, 66	0.2%	Great Britain
84, 168	0.005%	Great Britain
433.92	0.2%	Austria, Netherlands, Portugal, W. Germany, Yugoslavia, Switzerland
896	10 MHz	Great Britain
915	13 MHz	North and South America
2375	50 MHz	Albania, Bulgaria, Hungary, Romania, Czechoslovakia, USSR
2450	50 MHz	World-wide, except where 2375 MHz is used
3390	0.6%	Netherlands
5800	75 MHz	World-wide
6780	0.6%	Netherlands
24150	125 MHz	World-wide
40680		Great Britain

The loss factor, i.e. the product of the dielectric constant and the loss tangent, varies with a number of parameters including frequency, moisture content and temperature. The relationships are often quite complex as, for example, in drying where as the temperature of the material rises the moisture content will start to fall. However, one of the major applications of dielectric heating, moisture profile correction, takes advantage of the relative values of loss factor between areas of high and low moisture concentration. These are such that preferential heating and drying of the wetter areas takes place with resulting product quality improvements. *Table 9.9* gives an indication of the loss factor of a number of materials for a range of frequencies; some are quoted at a number of different temperatures. Interpretation of such data needs to be undertaken with care. In the case of water the effect of having any level of ionic material present (as in most 'real' water), will substantially increase the loss factor at r.f. but have relatively little effect at microwave frequencies.

In common with all high-voltage equipment, it is necessary to prevent access to live components and in this case it is also necessary to contain the electric field for both operator safety and to avoid interference with other users. Both considerations are dealt with by enclosing the equipment in a metal case, which may have interlocked access doors for batch operations and attenuating ducts for continuous processes.

9.7.1 RF dielectric heating systems

The available systems for producing and transferring RF power to dielectric heating or drying applicators can be divided into two distinct groupings; the more widespread *conventional RF heating equipment*, and the more recent *50 Ω RF heating equipment*. Although conventional RF equipment has been used successfully for many years, the

Table 9.9 Typical loss factor and frequency relation

Material	*Temperature* (°C)	*Frequency* (MHz)			
		1.0	*10*	*100*	*3000*
Ice	—	0.50	0.067	—	0.003
Water	1.5	1.6	0.17	0.61	25
	15	2.5	—	—	16
	65	5.6	—	—	4.9
	95	7.9	0.72	0.17	2.4
Porcelain	25	0.015	0.013	0.016	0.028
Glass					
Borosilicate	25	0.002	0.003	0.004	0.004
Soda-silica	25	0.07	—	0.051	0.066
Nylon (610)	25	0.07	0.06	0.06	0.033
	84	0.76	0.43	0.23	0.10
PVC	20	0.046	0.033	0.023	0.016
QYNA	96	0.24	0.14	0.086	—
VG5904	25	0.60	0.41	0.22	0.10
VU1900	25	0.29	0.17	0.087	0.015
Araldite (E134)	25	0.34	0.41	0.48	0.15
Adhesive	25	0.11	0.12	0.11	0.07
Rubber (natural)	25	0.004	0.008	0.012	0.006
Neoprene (GN)	25	0.54	0.94	0.54	0.14
Wood (fir)	25	0.05	0.06	0.06	0.05
Paper (royal-grey)	25	0.11	0.16	0.18	0.15
	82	0.08	0.14	0.19	0.23
Leather (dry)	25	0.09	0.09	0.12	—
15% water	25	0.78	0.49	0.45	—

ever tightening EMC regulations, and the need for improved process control, is leading to the introduction of RF heating systems based on 50 Ω technology.

9.7.1.1 Conventional RF heating equipment

In a conventional system, the RF applicator (i.e. the system which *applies* the high frequency field to the product) forms part of the secondary circuit of a transformer which has the output circuit of the RF generator as its primary circuit. Consequently, the RF applicator can be considered to be part of the RF generator circuit, and is often used to control the amount of RF power supplied by the generator. In many systems, a component in the applicator circuit (usually the RF applicator plates themselves) is adjusted to keep the power within set limits. Alternatively, the heating system is set up to deliver a certain amount of power into a standard load of known conditions, and then allowed to drift automatically up or down as the condition of the product changes. In virtually all conventional systems, the amount of RF power being delivered is only indicated by the d.c. current flowing through the high power valve (usually a triode) within the generator.

A typical conventional RF heating system is shown schematically in *Figure 9.25*.

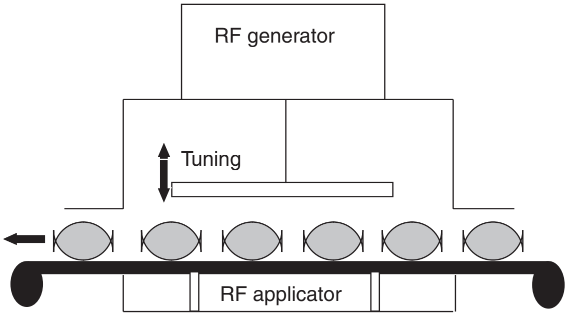


Figure 9.25 Components of a conventional RF heating system

9.7.1.2 50 Ω RF heating equipment

RF heating systems based on 50 Ω equipment are significantly different, and are immediately recognisable by the fact that the RF generator is physically separated from the RF applicator by a high power coaxial cable (*Figure 9.26*).

The operation frequency of a 50 Ω RF generator is controlled by a crystal oscillator and is essentially fixed at exactly 13.56 MHz or 27.12 MHz (40.68 MHz is seldom used). Once the frequency has been fixed, it is relatively straightforward to set the output impedance of the RF generator to a convenient value—50 Ω is chosen so that standard equipment such as high power cable and RF power meters can be used. For this generator to transfer power efficiently, it must be connected to a load which also has an impedance of 50 Ω. Consequently, an impedance matching network has to be included in the system which transforms the impedance of the RF applicator to 50 Ω. In effect, this matching network is a sophisticated tuning system, and the RF applicator plates themselves can be fixed at an optimum position.

The main advantages of this technology over the conventional systems are:

(i) Fixed operation frequency makes it easier to meet onerous international EMC regulations.
(ii) The use of 50 Ω cable allows the RF generator to be sited at a convenient location away from the RF applicator.
(iii) The RF applicator can be designed for optimum performance, and is not itself part of any tuning system.
(iv) The use of a matching network gives the possibility of an advanced process control system. The positions of components in the matching network give on-line information on the condition of the dielectric load (such as its average moisture content). This information can be used to control the RF power, speed of conveyor, temperature of air in applicator etc. as appropriate.

9.7.1.3 RF dielectric heating applicators

Whether conventional or 50 Ω dielectric heating systems are used, the RF applicator has to be designed for the particular product being heated or dried. Although the size and shape of the applicator can vary enormously, they mostly fall into one of three main types.

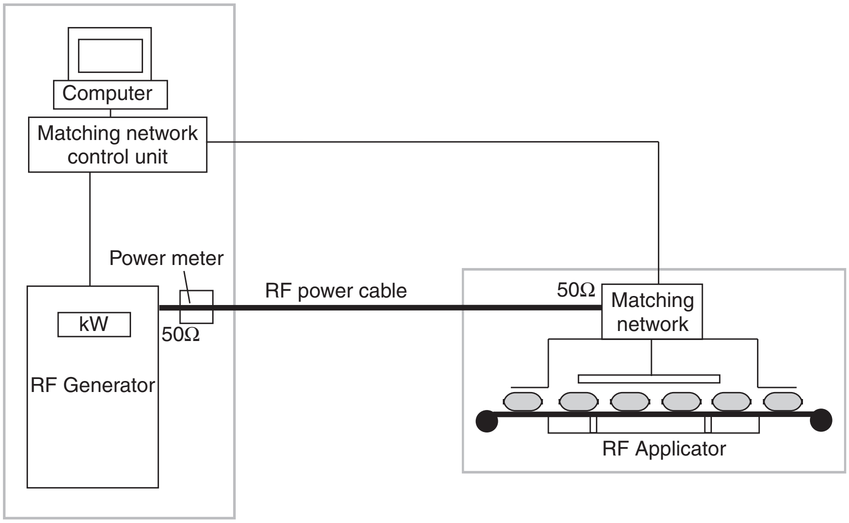


Figure 9.26 The components of a typical 50 Ω dielectric heating system

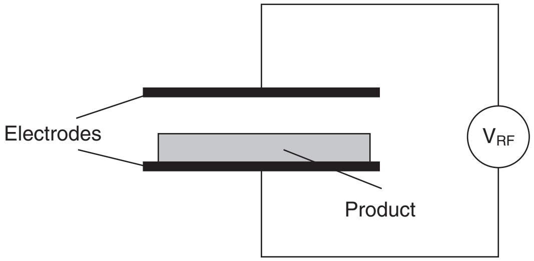


Figure 9.27 Simple through-field RF applicator

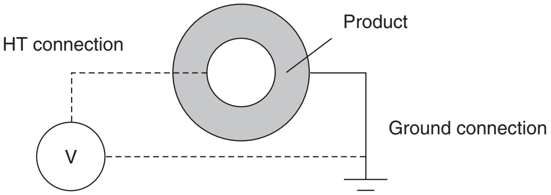


Figure 9.28 Concentric through-field

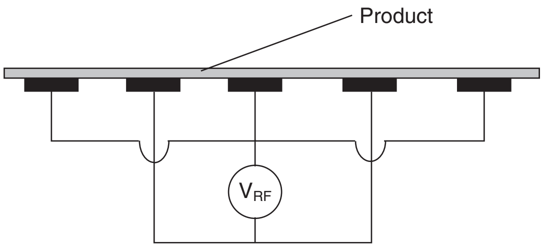


Figure 9.29 Fringe-field applicator

'Through-field' applicator Conceptually, a through-field RF applicator is the simplest, and the most common, design, with the electric field originating from a high frequency voltage applied across the two electrodes which form a parallel plate capacitor (*Figure 9.27*).

This type of applicator is mainly used on relatively thick products, or blocks of material.

A variation on this is the concentric through-field, shown in *Figure 9.28* in which the product to be heated, normally a liquid or fluidised bed, fills the space between the two cylindrical electrodes. The inner electrode is at high voltage, the outer one at earth potential.

'Fringe-field' applicator An alternative RF applicator arrangement, often used in drying applications, is known as the *fringe-field* system. In this case, the product passes over a series of bars, rods or narrow plates which are alternatively connected to either side of the RF voltage supply (*Figure 9.29*).

The major advantage of this configuration is that the product makes a complete contact across the bars, and there is no air gap between the RF applicator and the product. This ensures that there will be a virtually constant electric field in the material between the bars (an important requirement to maximise moisture levelling performance). It also reduces the electric field that has to be applied between the electrodes to generate a given power density within the product. The major disadvantage of this arrangement is that only relatively thin layers of product can be used—otherwise there will be an electric field variation throughout the product thickness.

'Staggered through-field' applicator For intermediate thickness of products, a modified through-field applicator is often used; known as a *staggered through-field* applicator (*Figure 9.30*).

This arrangement reduces the overall capacitance of the applicator which, in turn, makes the overall system tuning easier. It also reduces slightly, the voltage that has to be applied across the electrodes to produce a given RF power density within the product.

9.7.2 Microwave power sources and applicators

For industrial heating applications using microwaves, the usual power source is the magnetron. At the permitted operating frequency of 2450 MHz, the largest output from a magnetron commonly used, is 5 kW, although 10 kW units are just becoming available. For the other frequency band (896/915 MHz) high-efficiency magnetrons of 60 kW output exist. When higher powers are needed, a number of magnetrons may be fed into one applicator. The most common form of industrial heating oven is the multimode oven, essentially an enlarged version of the familiar domestic oven with, in the case of continuous processing, appropriate product ports, to allow the passage of product yet confine the microwaves within the oven. In such applicators the antenna of the magnetron may be mounted directly into the oven, but more often the microwave power is transmitted from the power supply via metallic pipes (waveguides) to the oven cavity where it is launched into the chamber by a variety of means. Other forms of heating applicator which incorporate, for example, launching horns,

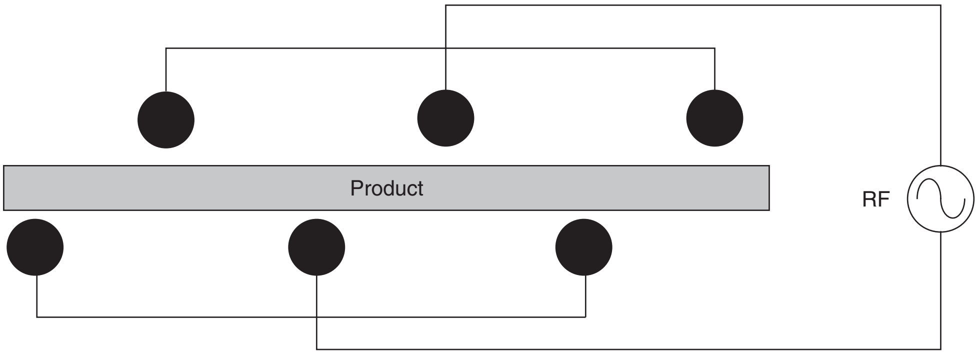


Figure 9.30 Staggered through-field applicator

leaking wave antennas, or single-mode resonant cavities, are possible.

Industrial microwave heating has been used extensively in the rubber industry for curing and preheating prior to moulding. In the food industry it has been used for tempering, melting, cooking and drying. Recently, microwave vacuum dryers have been developed for drying expensive, high-quality temperature-sensitive pharmaceuticals.

9.8 Ultraviolet processes

Inks and surface coatings can be cured at high rates with ultraviolet sources. The coatings are specially formulated, using monomers and photoinitiators, so that very rapid polymerisation is brought about on exposure to ultraviolet radiation. Although this is not strictly a heating process, it has much in common, and is very often in direct competition with infra-red heating for drying or cross-linking; however, the energy usage is normally much lower, since the process requires only the stimulation of an overall exothermic reaction.

The active spectral region covers the range 250–400 nm in the ultraviolet, and visible wavelengths in the range 400–500 nm may also be used. The envelopes of the ultraviolet lamps are usually made from pure silica glass, which has higher transmission than other glass for the shorter ultraviolet wavelengths. Energy radiated from an electrical discharge is not governed by the black-body laws: the distribution can be confined to a few intense narrow bands of wavelength. A typical output spectrum from a medium-pressure (sometimes called high-pressure) mercury vapour discharge lamp is shown in *Figure 9.31*. Additional spectral bands can be generated by incorporating metal halide dopants in the lamp fill. Power ratings of these lamps are in the range 2–20 kW over active lengths of 250–1800 mm.

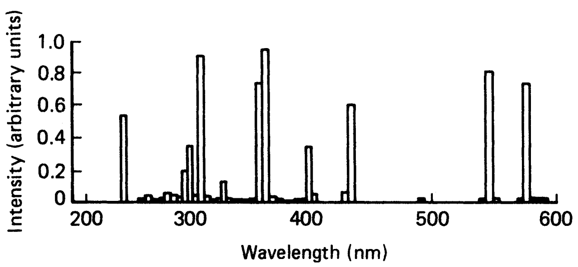


Figure 9.31 Spectral output from a medium-pressure mercury lamp

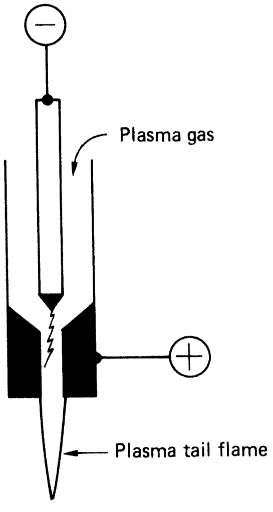


Figure 9.32 Rod and nozzle type of plasma torch construction

Low-pressure actinic lamps are used for curing thick sections of glass-reinforced polyester resins. These lamps operate at much lower power densities than medium-pressure lamps, a typical power rating being 80 W for an active length of 1.5 m. Suitable phosphors on the internal surfaces of the lamp tubes convert the primary radiated output, which is concentrated at 254 nm, to broad-band emission in the region 300–500 nm. The lamp tubes are made of glass, which has good transmission properties down to about 300 nm.

Both medium-pressure and low-pressure mercury lamps are used to disinfect air and water, and to sterilise surfaces. Wavelengths shorter than 300 nm are required, the action spectrum peaking at about 260 nm. In the case of low-pressure, 'germicidal' lamps the primary radiated output at 254 nm is utilised. The effect achieved is, again, not the result of a heating process, but the consequence of the biologically damaging properties of ultraviolet radiation.

9.9 Plasma torches

9.9.1 Types of plasma-torch design

A plasma torch is a device that uses an electric arc discharge to generate a thermal plasma, that is, a partially ionised gas in which the degree of ionisation is linked to the temperature of the gas according to the laws of equilibrium thermodynamics. When the temperature of a gas rises above about $6\times\cancel{10}^{3}$°C the electrical conductivity becomes sufficiently high to make the gas a reasonable conductor of electricity. Temperatures in the core of an arc may reach $2\times\cancel{10}^{4}$°C or higher.

Torches that are used in electroheat applications may be broadly classified into the three families described below.

(1) *Rod and nozzle electrode types (Figure 9.32)*: an arc burns between the end of a rod and the internal surface of a nozzle and gas is blown around the arc and through the nozzle. The current supply is usually d.c., but torches designed for a.c. use are also available.

(2) *Linear coaxial tube types (Figure 9.33)*: the arc burns between the internal surfaces of tube electrodes and gas is blown through the electrodes around the arc. The current supply is usually d.c., but some designs also operate on a.c. The power-handling capacity is improved if arc root motion is induced by the imposition of axial magnetic fields.

(3) *Electrodeless types, particularly induction-coupled types (Figure 9.34) and microwave types*: an induction coupled arc burns in a ring-shaped electric field induced within a coil carrying current typically at a frequency of a few megahertz. The arc plasma is effectively the workpiece in an r.f. induction heater. In microwave torches the arc is maintained by currents driven by microwave fields in a resonant cavity. Peak plasma temperatures in electrodeless torches are typically within the range $7\times\cancel{10}^{3}$ to $1.1\times\cancel{10}^{4}$°C, that is, rather lower than for the electrode types.

If the plasma torch is intended to handle significant power levels, the electrodes must be adequately cooled to dissipate

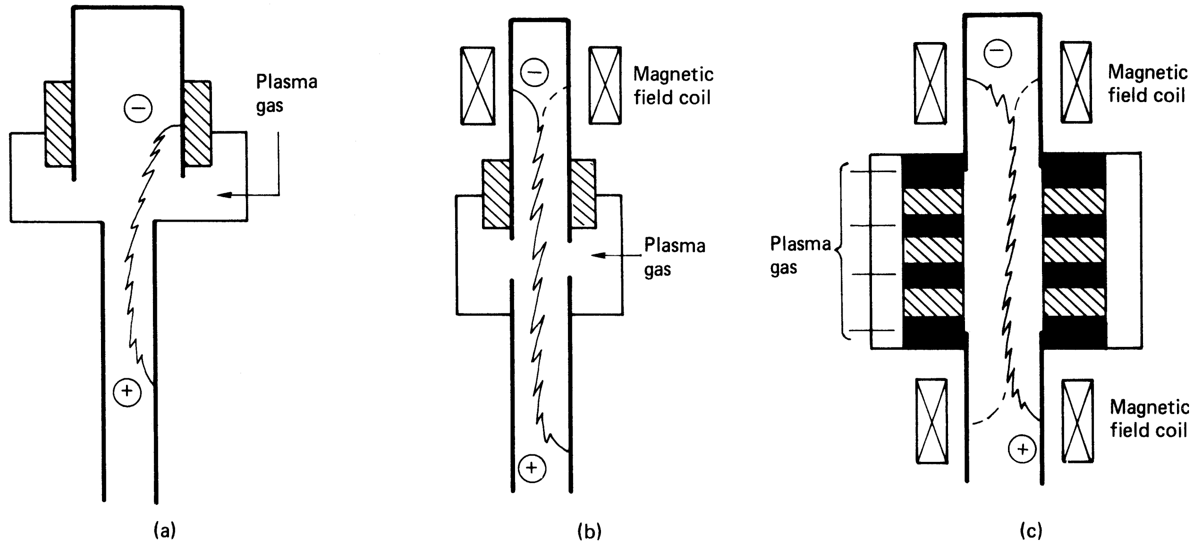


Figure 9.33 Linear coaxial tubular electrode types of plasma torch construction. (a) The Hüls type developed from Schoenherr's original early design; (b) A typical 2 MW non-transferred torch; (c) A typical stretched or segmented torch rated at 6 MW or higher

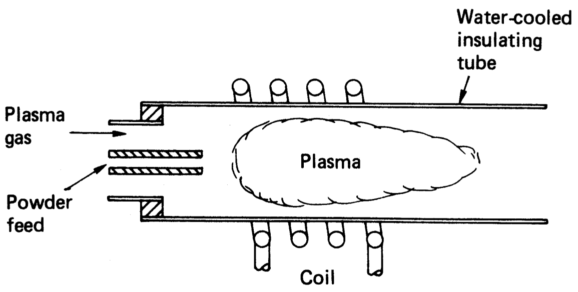


Figure 9.34 An induction-coupled plasma torch

the heat transferred at the arc roots. Water cooling is usual and electrode cooling considerations are critical for torch design. Water cooling may also be necessary for other parts of the torch exposed to the arc plasma.

9.9.2 Electrical connection of plasma torches

Plasma torches are used in various different types of connection (*Figure 9.35*). In the non-transferred mode the torch behaves as an arc-powered gas heater. In the transferred mode the torch behaves as one electrode for an arc while the workpiece is electrically connected as the other electrode.

A third mode of connection, the superimposed mode, is a hybrid connection requiring the use of an auxiliary a.c. or d.c. power source which may be either transferred (as shown) or non-transferred.

9.9.3 Performance

Transferred torches are available for industrial applications at power ratings up to about 7 MW in a single device. State-of-the-art torches are being developed for 12 kA capacity but 6 kA capacity torches are widely available. In most industrial applications voltages up to about 500 V may be accommodated without serious stray arcing but, with careful design, long arcs (1 m or longer) operating at up to about 1 kV are possible. Electrode life-times are dependent on the torch current, the materials of construction and the reactivity of the environment to which they are exposed. Life-times over 100 h are usually necessary for application in major industrial process plant.

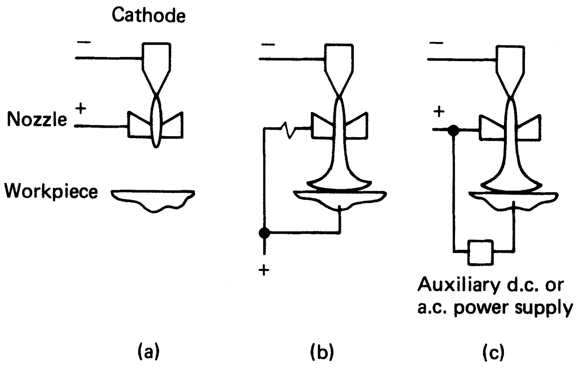


Figure 9.35 Electrical connections of plasma torches. (a) Non-transferred; (b) Transferred; (c) Superimposed mode shown transferred to workpiece

The efficiency of transferred torches is very dependent on the specific application. Electrical to thermal conversion efficiencies can reach 95% but, in a practical furnace, the efficiency of conversion to heat transferred to the hearth may fall as low as 50% unless heat losses from the arc plasma column are usefully intercepted. The mechanism of heat transfer is mainly convection.

Non-transferred torches are available in power ratings up to about 100 MW, but for industrial applications a practical maximum rating of about 10 MW is imposed by electrode life-time considerations. Life-times of 1000 h and longer are possible in air or in hydrocarbon gases, provided that the current may be maintained below about 1.5 kA.

The efficiency of non-transferred torches may reach 90% for coaxial tube electrode types when used with high gas flow rates, corresponding to relatively low gas temperatures of 3×10^3 to 4×10^{3}°C. The efficiencies of rod and nozzle types, which are generally used for lower power ratings up to 100 kW, tend to be lower at about 60% at best.

The major advantage of electrodeless torches is that reactive gases are more easily maintained in a state of high purity than in other torch designs, and that long torch-component lifetimes are achieved. Their efficiency is limited, by the conversion efficiency to high-frequency power, the coupling efficiency into the plasma and heat losses to the containing walls, to the range 40–60% at best.

9.9.4 Plasma furnaces or reactors

The simplest type of furnace or reactor is, in principle, the 'in-flight' reactor in which the input materials are injected directly into the plasma stream, or even into the torch, and the required processes occur with the reactants all suspended in the plasma stream. In-flight reactors generally require the input materials to be gaseous or finely divided because the time available for the process is short, typically of the order of milliseconds.

Transferred plasma torches are commonly used in open bath furnaces. A return electrode is then usually built into the hearth refractories, but three-phase a.c. and bipolar d.c. multiple torch systems have been developed to eliminate the need for the hearth connection. The feed rate of materials to open bath furnaces is usually controlled to prevent the build up of unreacted materials on the surface, but plasma torch operation has also been demonstrated in a submerged mode.

Non-transferred torches are often used in shaft-type furnaces where a hot reaction zone is created close to the point of injection of reactive plasma gases in the base of a packed shaft. Furnaces with rotating shells are also available, particularly for the fusion of pure refractory compounds (*Figure 9.36*) or when a long residence time is required as in the treatment of some waste materials.

The advantages of a plasma furnace are that the heat input is independent of the process chemistry, so that highly oxidising or fully reducing furnace atmospheres are equally permissible, that furnace operation may be made largely independent of the physical characteristics of the input materials, that furnace gas outflow rates are reduced to levels that allow the treatment of fine powders without excessive carry-over, that very high temperatures may be achieved limited only by the containment requirements, and that compact high energy density, high throughput systems, are possible.

Some established applications of plasma technology are listed in *Table 9.10*.

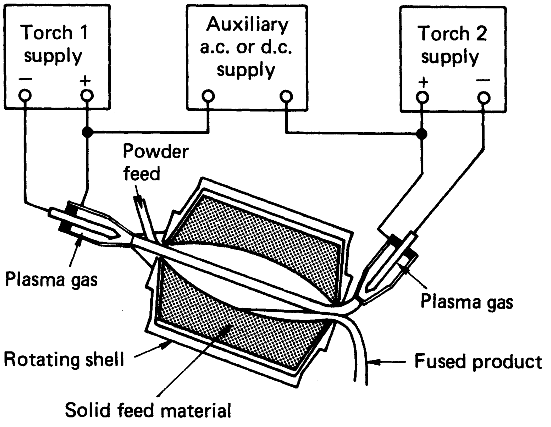


Figure 9.36 Rotating shell furnace for fusing pure refractory powders and using a non-transferred, superimposed mode torch arrangement

Table 9.10 Applications of plasma heating technology

Application	*Plasma system*
Welding, cutting, surface heat treatment of metals	Transferred, typically up to 100 kW
Precision cutting of metal foils, cutting non-metals	Non-transferred, from 0.5 kW
Spray coating of refractory metals and ceramics	Transferred or non-transferred, typically up to 100 kW, in-flight reactors
Platinum group metals recovery from scrap vehicle exhaust catalyst	Transferred, 1–2 MW
Steel tundish heating	Major application, transferred 1–3 MW
Consolidation melting and cold hearth melting of reactive metals (e.g. Ti)	Transferred, torch anodic, typically 800 kW
Blast furnace blast superheating	Non-transferred, up to 8 × 2 MW
Plasma fired cupolas	Non-transferred, 4 MW and 6 × 1.5 MW
Metals recovery from steel-plant dusts	Major application, from transferred 1 MW to 3 × 4.5 MW non-transferred (shaft furnace)
Smelting of metal ore fines	Transferred and non-transferred from 1 MW
Acetylene production, process off-gass reforming	Non-transferred, up to 8.5 MW
Titanium dioxide pigment production	Non-transferred, 2 MW oxygen heater
Waste treatment	Emerging major application

9.10 Semiconductor plasma processing

Plasma processing has been used in the semiconductor industry since the late 1960s when rudimentary processes were introduced for removing photoresist from silicon using an oxygen plasma. This was an extension of an earlier application in the medical field in which an oxygen plasma was used to volatilise organics from biological specimens in order to separate the inorganic residue or 'ash'. The misnomer 'ashing' is still used in the semiconductor industry to describe this resist-stripping process.

The widespread use of plasmas for semiconductor applications stemmed from work carried out in the early 1970s in which fluorocarbon gases were used to generate highly reactive fluorine atoms and C_xF_y radicals. These species were shown to be capable of etching most of the materials of interest in device fabrication, and the relative etch rates could be controlled by manipulation of the concentration of these species.

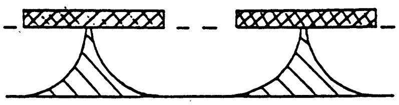
More recently, plasma processes have been used for plasma-enhanced chemical-vapour deposition (PECVD) both of insulating films like silicon nitride and also of semiconducting films such as amorphous silicon (a-Si). The use of PECVD Si_3N_4 for passivation of integrated circuits (ICs) has allowed plastic packaged ICs to have reliability comparable to hermetically packaged devices.

More novel applications of plasma processing include techniques in which material is grown such as *plasma oxidation*—oxidation of existing material or *deposition* of SiO_2 and plasma-assisted epitaxy. As yet these techniques are not in widespread use but offer some significant advantages over their conventional counterparts.

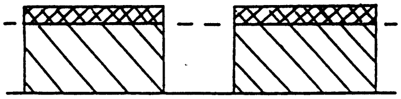
The reasons for the increasing use of plasma processes are three-fold.

(1) *The requirement for low-temperature processing.* Thermal oxidation and chemical vapour deposition are performed at temperatures in the range 800–1000°C, and sometimes higher. These thermal processes consume a considerable amount of energy leading to an increased thermal budget for the fabrication facility. In addition, they introduce unwanted dopant redistribution, thermal warping, dislocations and stacking faults in the semiconductor substrate. Furthermore, when an aluminium layer is used for circuit interconnection, these elevated temperatures preclude the deposition or growth of further insulating layers; as a result the fabrication of multilayer circuits becomes problematic.

(2) *The need to etch small geometry structures.* As the dimensions of circuits and devices are reduced, conventional solution or wet-chemical etching becomes more difficult, because the surface tension of the solutions tends to cause the liquid to bridge the space between adjacent features, thereby preventing further etching of the underlying film. As wet etching proceeds downwards through a film, it also proceeds laterally because of its essentially isotropic characteristic, thereby undercutting and broadening lines and restricting the minimum dimensions of the device. In contrast, plasma etching is anisotropic, resulting in a very high ratio of vertical to lateral etch rates. Very steep edge profiles are achievable for interconnection layers with minimal undercutting (see *Figure 9.37*).



(a) Isotropic (wet chemical)



(b) Anisotropic

Figure 9.37 Etch profiles for isotropic and anisotropic etching. Undercutting beyond the mask edge is minimal in the case of anisotropic plasma etching

(3) *The desirability of a dry process.* Strong acids can be used to strip resist films and to etch semiconductor wafers, but with the associated disadvantages of lack of control and the need for a large amount of high purity chemicals and deionised water. In contrast, plasma etching is a dry process which takes place in a clean vacuum environment. The sources of heavy metal and sodium ion contamination commonly found in solution are eliminated. The process can be monitored by controlling the flow rate of the reactive gases and plasma parameters, and end-point detection of the process is possible.

9.10.1 Basic mechanisms in plasma processing

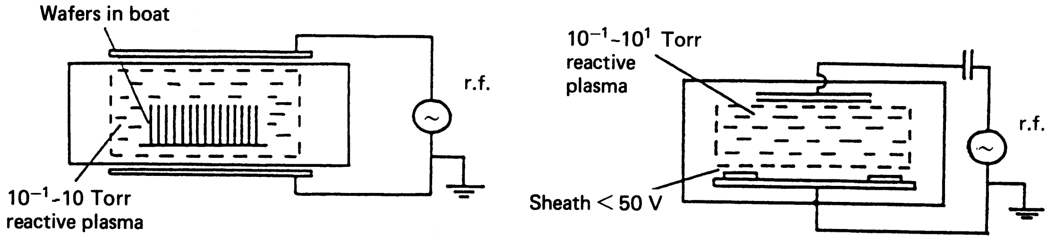
All plasma processes use a low-pressure gas in the form of a plasma to provide the required chemical species, whether for material removal (etching and ashing) or for material deposition. Free electrons within the plasma processing reactor are accelerated by an applied electric field until they gain sufficient energy to excite some of the parent gas molecules into highly reactive states by collisional processes. The resulting glow discharge, although still mainly composed of neutral ground-state molecules, contains significant numbers of free radicals, ions and other excited species. The degree of ionisation may typically be only of the order of 10^{-4}, but the excited species can be highly reactive, to the extent that reactions either within the body of the plasma, or on exposed surfaces can occur at rates which would only be achieved thermally at temperatures in excess of 1000°C.

These plasma-generation and surface-reaction processes, although widely used, are still not well understood and are very difficult to model. In most systems, however, free radicals are believed to be the major reactant species and ions serve primarily to enhance their reaction rates.

The plasma, due to its degree of ionisation may be regarded as an equipotential volume. Surfaces in contact with the plasma, however, can attain potentials which differ from the plasma potential by amounts which depend upon the configuration of the system and its operating conditions (gas composition, pressure, applied power, excitation frequency, etc.). Because of the higher energies and mobilities of electrons compared to positive ions, electrically isolated surfaces develop a floating potential a few volts below that of the plasma in order to maintain a net charge neutrality. The electric field which maintains this flux neutrality causes a reduction in the electron concentration near the surface, resulting in a plasma sheath, or dark space, where the generation of excited species which decay by photo-emission is much lower than in the bulk of the plasma.

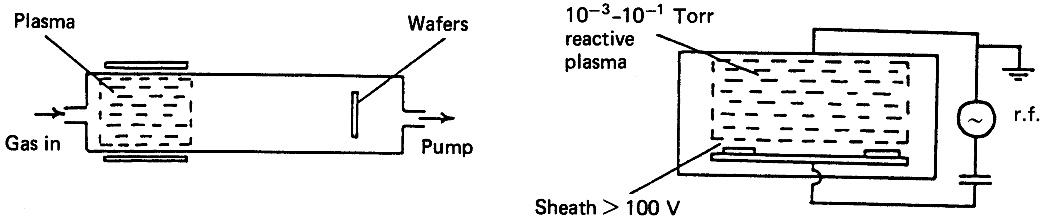
Electrodes in contact with the plasma develop similar plasma sheaths, whose magnitude depends upon the relative areas of the electrodes and the frequency of the RF. Higher frequency = smaller dark space. In most plasma-processing systems the energy is fed into the plasma by capacitive coupling (see *Figure 9.38*). In a capacitively driven plasma with internal electrodes, a large sheath voltage appears at the smaller electrode and a smaller one at the larger electrode. These voltages are typically of the order of tens or hundreds of volts in practical systems.

The electrode sheaths are usually a few millimetres thick, so that a device wafer with topographical flatness of the order of a micrometre, placed on an electrode does not significantly perturb the field directionality. It is this directionality which gives rise to anisotropic etching in ion-enhanced processes, by enhancing the etch rate in the direction normal to the electrode surface.



(a) Barrel etcher

(c) Plasma etcher and capacitively coupled PECVD configuration



(b) Downstream etcher

(d) Reactive ion etcher

Figure 9.38 Typical configurations for plasma-processing equipment

Table 9.11 Characteristics of the most common plasma etching configurations

Configuration	*Etch profile*	*Selectivity*	*Etch rate*	*Type*	*Range* (T)	*Electrode arrangement*	*Wafer position*
Barrel	Isotropic	V. good	Fair	Chemical	10^{-1}–10^{1}	External	Immersed in plasma
Downstream	Isotropic	V. good	Fair	Chemical	10^{-1}–10^{0}	External	Downstream of plasma
Plasma etch	Variable	Fair–good	High	Physico-chemical	10^{-1}–10^{1}	Planar diode	Anode (grounded)
RIE	Anisotropic	Fair	Fair	Physico-chemical	10^{-3}–10^{-1}	Planar diode	Cathode (driven)
MRIE	Anisotropic	Fair	High	Physico-chemical	10^{-3}–10^{-2}	Planar diode	Cathode (driven)
Triode	Variable	Fair	High	Physico-chemical	10^{-3}–10^{-1}	Planar triode	Cathode (driven)
RIBE	Anisotropic	Fair	Fair	Physico-chemical	$<10^{-3}$	Planar triode	Cathode (grounded)

9.10.1.1 Plasma etching

A dry etching process (other than physical sputtering) can be described in terms of four sequential steps: generation of the reactive species, adsorption onto the wafer surface, chemical reaction (may be radiation enhanced), and desorption of the reaction products. Any one of these steps may be the rate-limiting step. It is also necessary that the reaction products be volatile in order that the etch products do not contaminate the surface and prematurely terminate the process. For this reason, fluorine or chlorine containing plasmas are used for virtually all inorganic films used in semiconductor applications and oxygen for organics.

Types of plasma etching processes The most common plasma etching configurations are shown in *Figure 9.38* and their characteristics are listed in *Table 9.11*. In chemical plasma etching active species are generated in a reactor by an externally excited r.f. or microwave discharge and the wafers are either immersed in the plasma at a floating potential (barrel etching) or located downstream of the plasma in a gas flow (downstream etching). Ion bombardment is minimal in the former and absent in the latter. Etching is due to free atoms and excited neutrals and is, therefore, isotropic. With an appropriate choice of etch gas the selectivity can be very high because of the absence of physical enhancement but etch rates are relatively low for the same reason. The lack of anisotropy makes this approach unsuitable for very large scale integration (VLSI) patterning applications, but it is widely used for photoresist stripping (ashing).

By far the most widely used means of achieving directional etching is to use systems in which the wafer is physically bombarded by the ions in the plasma as well as exposed to the chemically reactive gaseous and ionic species. In the plasma etch mode the wafers are placed on the larger (grounded) of two electrodes whereas in reactive ion etching (RIE) they are placed on the smaller (driven) electrode and the system is operated at lower pressure. Under these conditions the ion energies bombarding the wafer are of the order of a few tens of electron-volts in the former and several hundred in the latter. RIE therefore tends to be more anisotropic, but this is gained usually at the expense of reduced selectivity (due to the increased ion bombardment) and reduced etch rate (due to the reduced flux density of reactive species at the lower pressure). However, all of these parameters can be optimised for a given application.

Magnetron-enhanced RIE (MRIE) and triode etching are more recent refinements of this class of operation. In the former case free electrons are magnetically confined in the plasma by a magnetic field orthogonal to the electric field. The collision probability is, therefore, increased and the reactant densities can be enhanced by up to two orders of magnitude. This results in higher etch rates than RIE at similar pressures. The reduced plasma impedance results in lower sheath voltages, but the higher flux of low energy ions is more efficient in promoting physically stimulated processes and so anisotropy is still good, with the added advantage of less radiation-induced damage to the device substrate.

In triode systems, r.f. power is applied to a third electrode in such a way that the cathode sheath voltage is controlled independently of the generation of the reactive species. This allows some degree of decoupling over the physical and chemical components of the process, and thus permits anisotropy and selectivity to be controlled.

Reactive ion beam etching (RIBE) describes the process in which a spatially collimated, usually reasonably mono-energetic, flux of one or more ion species is directed at a substrate leading to etching or erosion of that substrate. Ion energies may be controlled independently of the neutral flux by means of extraction grids.

Choice of etchant For silicon based materials, fluorinated gases have been extensively used, which include CF_4, C_2F_6, CHF_3, SF_6 and, more recently, NF_3. The major criterion in selection is to create volatile products with the constituents of the target at reasonable working temperatures. Etching other semiconducting materials such as III–V compounds, has required a slightly different approach as the group III fluorides are involatile at workable plasma parameters. For this reason chlorinated gases have been widely used, such as Cl_2, CCl_2, F_2, $SiCl_2$ and BCl_3.

9.10.1.2 Plasma enhanced chemical vapour deposition

Chemical vapour deposition (CVD) is a reaction in which two types of gas react at about 1 atm and at high temperature to form a solid phase and a gas phase. Plasma enhanced CVD (PECVD) may be defined as a gas-phase reaction in a low-temperature plasma that forms a similar thin solid film by precipitation onto a substrate. The catalytic effect of the plasma is to accelerate the reaction between the dissociated molecules of the reactant gases. Techniques such as mass spectroscopy and emission spectroscopy allow study of the various reactions within the PECVD process.

The composition and properties of the films vary significantly with substrate temperature and also with the method of generating the plasma. In general, PECVD films are less dense, usually due to the incorporation of hydrogen within the structure. In the case of semiconductors produced by this process, e.g. a-Si, the presence of hydrogen significantly affects both the physical and electronic properties.

Both inductively coupled and capacitively coupled high-frequency power supplies have been used for PECVD, at frequencies ranging from 50 kHz to 13.56 MHz. Of great importance is the power density distribution within the plasma since the life-times of the reactive species are short and must be continually produced in the deposition zone. For this reason the most successful PECVD systems are capacitively coupled and planar in construction—many having developed from the designs of plasma etchers. The introduction of a dilutant gas into the plasma, such as helium, has proved beneficial to film properties. This is thought to be due to the fact that the helium effectively cools the plasma by reducing the electron temperature and hence reduces damage. Uniformity of deposition in such systems is now better than 3% over 8 in. wafers and the step coverage is very good.

9.10.2 Power supplies for plasma production

Four basic types of power supply are used to excite the plasma in most semiconductor plasma-processing equipment. These are: d.c. supplies, high frequency (h.f.) supplies for the range 20 kHz–1 MHz, radiofrequency (r.f.) supplies for the range 1–100 MHz and microwave supplies. The h.f. and r.f. supplies are the most commonly used.

9.10.2.1 H.f. and r.f. power supplies

R.f. power supplies provide an output current and voltage which are alternating with equal but opposite amplitude at a given frequency. The connection to a plasma system is complicated by the fact that the 'plasma load' or r.f. connection to the system cannot be considered as a simple resistive load, but a load with a complex impedance dependent mainly on pressure, applied field and type of gas. The value of plasma impedance can give rise to substantial phase shifts of applied voltage and current resulting in an energy loss in the plasma reactor due to unwanted dissipation.

The output of a generator has a specified termination impedance or output impedance (most commonly 50 Ω resistive). To achieve maximum power transfer to the plasma the load (plasma) impedance must be equal to the output impedance and this is usually achieved using a π- or L-type impedance matching network. These networks effectively transform the plasma load impedance to that of the output impedance and can be automatically tuned to suit the particular plasma.

For the h.f. case (<1 MHz) the connection of the load to the generator is generally simpler than for r.f. systems since the plasma impedance is more stable, stray capacitances are small and circuit losses insignificant. Such power supplies tend to be used in the inductively coupled mode for PECVD systems. For the r.f. case the most common frequency is 13.56 MHz which is an international allocation for industrial use. For these systems particular care must be taken with the design of components and circuitry. At these frequencies and above, r.f. currents travel in a conductor through its surface and, generally, the penetration (skin depth) decreases as the frequency increases. This means that coils and capacitors, etc., should have highly conducting and large surface areas to reduce losses.

The most widely used r.f. plasma processing system is the parallel-plate configuration with electrodes varying in size from 100 mm to about 1 m, depending on the type and use of the system. It is normal for the base of the chamber to form the substrate holder which is often part of the chamber or a separate earthed electrode. The top (r.f. powered electrode) is usually of a similar diameter, depending on the application, and parallel to the lower electrode. The generator will normally give some visual indication of forward and reflected power and thus allow some degree of control of the generation process.

9.10.2.2 D.c. and microwave power supplies

The d.c. low voltage power supply operates at voltages of 100–800 V and uses a magnetron for plasma confinement. The target or cathode is thus encased in a high density magnetic field, usually supplied by permanent magnets. This type of supply requires rugged construction as rapid and repeated short circuits can occur from the powered electrode in the plasma system to the positively connected earths at the start of the process. The output current of the supply (typically around 10 A) must also be regulated using closed-loop control, as small changes in the output voltage can create large changes in operating current. A common application of this type of supply is in sputtering equipment.

Plasma excitation by microwaves is also used for semiconductor plasma processing. The frequency of operation is usually 2.45 GHz and most power supplies use a magnetron oscillator with the energy introduced through an isolator into the wave cavity. The microwave circuit is monitored by measuring the power of the incident and reflected waves using a directional coupler. This form of discharge is known as an electrodeless discharge and contamination of the substrate is low. Such supplies are used in etchers, oxidation systems and, more recently, in plasma stream transport systems in which the substrate is placed downstream of the discharge and the plasma transported to it by magnetic flux tubes.

9.10.3 Current trends and future developments

It is certain that semiconductor plasma processing will continue to find increasing application throughout the microelectronics industry. In the field of etching most of the current activity in high fidelity pattern transfer is in the area of physico-chemical processes, and particularly in RIE and its refinements. Submicrometre features can already be defined with vertical or bevelled profiles in most of the materials of interest for VLSI applications, but there are still shortcomings in several respects. The enhanced etch rates and lower plasma sheath voltages obtainable with MRIE and triode systems are likely to provide a spur for further developments in these areas. RIBE, because of its flexibility and control over bombardment energy, looks attractive for many applications.

Much research and development is continuing in the field of PECVD of amorphous silicon. Applications of a-Si ICs include solar cells and thin-film transistors for active-matrix liquid-crystal displays. Small flat-screen colour displays to rival the conventional cathode-ray tube are currently available and their use is expected to become widespread in the future.

9.11 Lasers

9.11.1 Introduction

Laser[1] is a compound word which means 'light amplification by stimulated emission of radiation'.[2] Lasers (or laser devices) basically consist of: (a) a light amplification (or laser) medium; (b) an optical cavity which contains the laser medium; and (c) a pumping source which gives energy to excite atoms or molecules in the laser medium to higher energy levels.

The basic principles involved in producing laser light are shown schematically in *Figure 9.39*. The pumped laser medium has a state of inversion population where a specific upper level has a larger population than the lower level compared to the expected Boltzmann distribution of thermodynamic equilibrium.

Light that passes through such a laser medium is amplified by induced transitions from upper to lower energy levels. Light that is confined in an optical cavity increases its intensity by multiple reflection between two mirrors and multipassage through the laser medium.

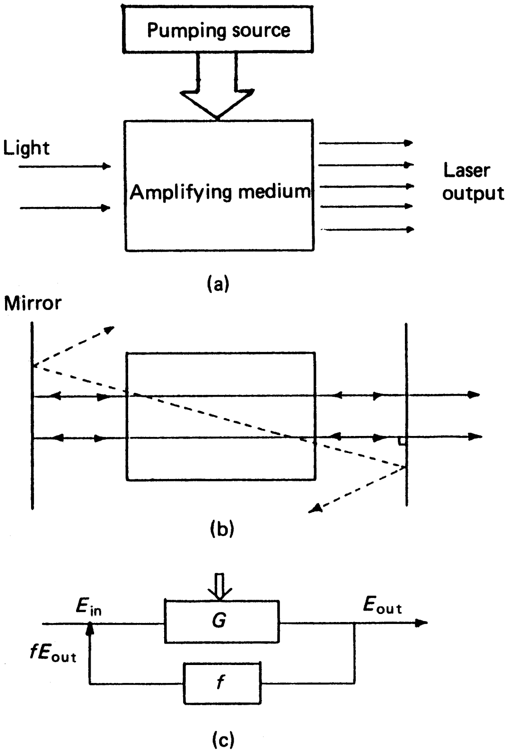


Figure 9.39 Principles of lasers. (a) Light amplification; (b) Optical cavity; (c) Oscillation by positive feedback

The laser medium may be solid, liquid, or gaseous in the form of a plasma (Section 9.10). Depending on the energy levels of the laser transition, the wavelength of laser light may be in the range from X-rays to millimetre waves.

The properties of laser light, compared with ordinary thermal radiation, can be summarised as follows.

(1) Laser light is an electromagnetic wave which has a high spatial and temporal coherence.
(2) Laser light has a high spectral purity which is determined by the resonant curve of the optical cavity and the energy level of the laser transition.
(3) A beam of laser light has a low divergence, the lower limit of which is determined by diffraction as the divergence angle of λ/D, where λ is wavelength and D is the beam diameter.
(4) The high focusability of laser light gives a small focal spot of the order of a wavelength. This gives a high energy or power concentration in space.
(5) A very short light pulse can be generated by a laser. The pulse width can be compressed to the order of femtoseconds.
(6) Well-defined polarised light can be generated by introducing a polariser into the cavity.

These characteristic features of lasers have wide application in modern science and industry.[3] Various kinds of lasers are listed in *Figure 9.40* along with their specific wavelengths and performance properties.

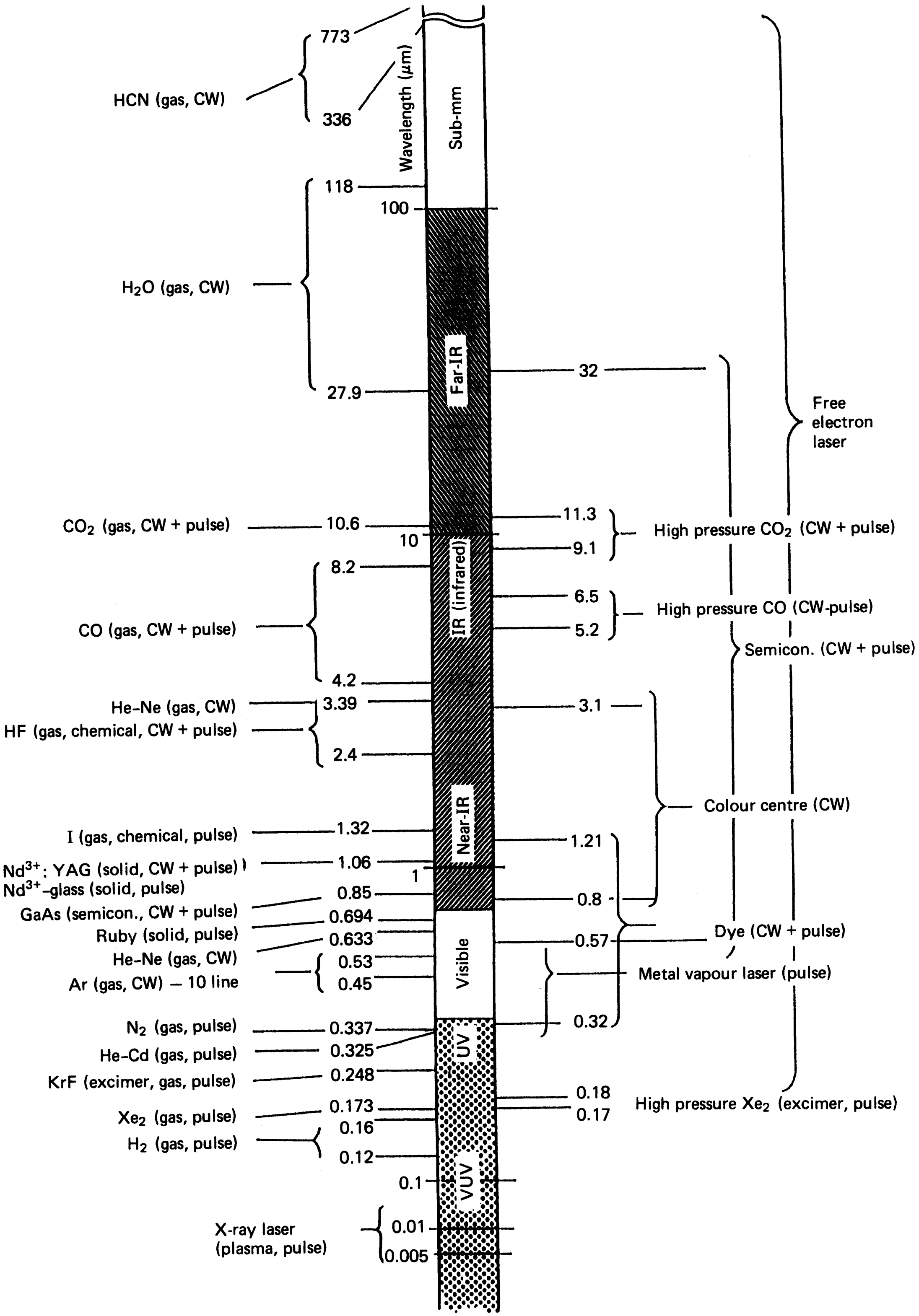


Figure 9.40 Typical lasers and wavelengths. CW, continuous wave; IR, infra-red

Table 9.12 Features and applications of excimer lasers

Excimer	*Wavelength* (nm)	*Output*	*Efficiency* (%)	*Application*
ArF	193	5–25 ns pulse 1–1000 Hz, ~500 mJ	≤1	Spectroscopy, photochemistry
KrF	248	2–50 ns pulse 1–500 Hz, ~1 J	≤2	Medical, lithography, dye laser pumping
XeCl	308	1–80 ns pulse 1–500 Hz, ~1.5 J	≤2.5	
XeF	351	1–30 ns pulse 1–500 Hz, ~500 mJ	≤2	

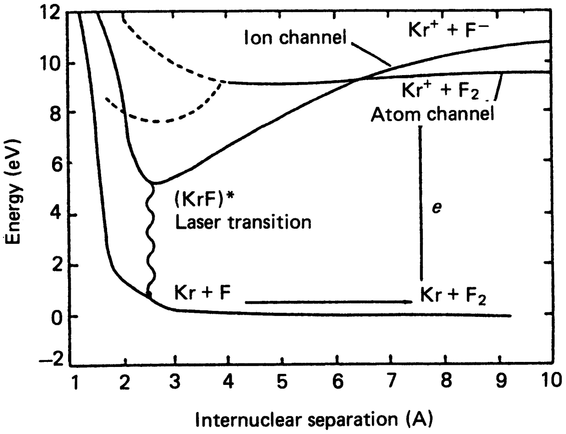


Figure 9.41 Energy-level diagram for the KrF excimer laser

9.11.2 Gas lasers

9.11.2.1 Excimer lasers

Typical excimer lasers that are commercially available are shown in *Table 9.12* along with their wavelengths and performance characteristics. An excimer is an excited molecule $(AB)^*$ which is stable only in its excited state. When the electronic state of A or B is excited or ionised in a mixture of both, they combine to form the excimer $(AB)^*$. A population inversion between $(AB)^*$ and (AB) can be formed because the ground state (AB) is not stable and dissociates to A and B as shown in *Figure 9.41* with the potential curve of $(AB)^*$ and (AB), as in the example of KrF.

Due to the short life-time of the upper state, excimer lasers operate in a pulsed mode with short-pulse duration. The wide bandwidth gain also enables pulse widths of pico- and femto-seconds to be generated, which is a useful feature for the investigation of high time-resolution phenomena. Repetitive operation of up to 1000 Hz provides a high average power of the order of 100 W.

The high photon energy of the short-wavelength light from an excimer laser, along with its high power capability, provides new application possibilities for photochemical material processing, medicine and as a light source for lithography of large scale integrated circuits.

The KrF laser has the potential for use in nuclear fusion,[4] which requires an output energy of megajoules per pulse and megawatts average power with high efficiency of 5–10%. A 10 kJ KrF laser has been constructed in the USA and a 100 kJ KrF laser is being planned and designed in Europe for application to fundamental science and industrial technology.

9.11.2.2 CO_2 lasers

A discharge of CO_2, N_2 and He excites a specific vibrational level of the CO_2 molecule and produces population inversion for laser action. The wavelength of the CO_2 laser extends from 9 to 11 µm, corresponding to the different vibrational-rotational levels. The efficiency of the CO_2 laser is typically 5–15%. The CO_2 laser is the laser with the capability of delivering the highest output power in continuous wave (CW) operation at the present state of technology and a 100 kW laser is commercially available with a high gas flow circulating system in conjunction with a stabilised discharge.

The CO_2 laser is one of the most widely used industrial lasers. The fields of application are material processing, such as cutting, welding, hardening and surface treatment, medical processing, diagnostics and heating of plasmas, and as a pumping source for various infra-red lasers.

9.11.2.3 Other gas lasers

The He–Ne laser is one of the most popular lasers and has been used extensively for diagnostics, data processing and alignment. The ion laser operates at short wavelengths (350–800 nm), and relatively high power (CW 20 W). This laser is used for spectroscopy, dye laser pumping for tunable-laser applications, and laser colour displays. The N_2 laser can generate short pulses of 0.3–10 ns duration at 337.1 nm wavelength (ultraviolet radiation) and with high peak powers of a few tens of megawatts. This laser is used for non-linear spectroscopy and dye laser pumping.

Metal-vapour lasers, especially copper vapour lasers, are under development as pumping sources for dye lasers used for laser isotope separation. Output powers of 100 W to 1 kW at 510 and 578 nm (green light) have been achieved. The He–Cd laser is a three-colour laser emitting at 441.6, 537.8 and 636.0 nm, with a relatively high CW power of 100 mW.

9.11.3 Solid-state lasers

9.11.3.1 Neodymium (Nd): YAG lasers

A single crystal of Nd doped yttrium–aluminium–garnet (YAG) produces laser radiation at a wavelength of 1.064 µm. A xenon flashlamp is used for pumping. The efficiency is relatively high at a few per cent. High average power of CW or repetitive pulse operation of 100 W to 12 kW has been obtained.

Frequency conversion of the fundamental output at 1.064 µm has been obtained to give third (355 nm blue light) and fourth (266 nm UV light) harmonics using nonlinear optical crystals with high efficiency. This feature of multicolour output leads to applications ranging from thermal to photochemical processing. The technology of crystal growth for manufacturing large laser elements is progressing rapidly and a single crystal of about 10 cm diameter without a central core is already being successfully manufactured. Laser diode (LD) pumping techniques are also under development and an LD pumped YAG laser with high efficiency of 30–40% has been demonstrated.

9.11.3.2 Neodymium:glass lasers

Neodymium doped glass emits radiation at 1.053 or 1.062 µm when *phosphate* glass or *silicate* glass is used, respectively, as a host material. The thermal conductivity of glass is low, and the cooling rate to remove the residual heat should not be too high. Therefore, the repetitive operation of a glass laser is limited. On the other hand, large laser elements with high optical uniformity can easily be manufactured.

Glass lasers are therefore used for laser fusion experiments where high peak power (terawatt) with short pulse (nanoseconds) is required even without repetitive operation and high efficiency. The largest glass laser amplifier is 45 cm in diameter with segmented elements or 35 cm in diameter with a monolithic element. The largest laser system is the NOVA glass laser at Lawrence Livermore National Laboratory, USA, which can deliver 120 kJ/ns at 1.053 µm fundamental wavelength and 70 kJ in third harmonics of blue light. The GEKKO XII glass laser at the Institute of Laser Engineering, Osaka University, which is used for international collaborative research, has output powers of 20 kJ/ns at 1.053 µm, 15 kJ as green and 12 kJ as blue light.

Glass lasers of higher power are planned and are being designed to generate 100 kJ to megajoules output energy as blue light for thermonuclear fusion ignition experiments and also for basic high-energy-physics research.

9.11.3.3 Laser diode and laser-diode-pumped solid-state lasers[5]

Semiconductor diode lasers have progressed in terms of performance and application achievements. One of the most important fields of application is in optical communication with fibre transmission. A wide tunable range with narrow spectral width is required for applications in coherent optical communications with frequency division multiplexing. Laser disks and laser scanners also provide a demand for LDs. High power and long life are the basic requirements for commercial use.

Typical LD performance specifications that have been individually achieved include:

(1) long life (>10 years),
(2) high power (>1 W (CW, one junction), > 115W (CW, 1 cm array), >200 W (pulse, 1 cm array)),
(3) efficiency (>50%),
(4) spectral purity (<1 MHz (20 mW)), and
(5) tunability ($\Delta\lambda = 2$–6 nm).

One of the most important energy applications of LDs is as a pumping source for solid-state lasers. *Figure 9.42* shows the energy levels and absorption spectrum of an Nd laser compared with the emission spectra of a laser diode and an Xe flash lamp. The output emission of the LD can be tuned to the most efficient absorption line for population inversion. Thus the LD laser pump has the potential for high

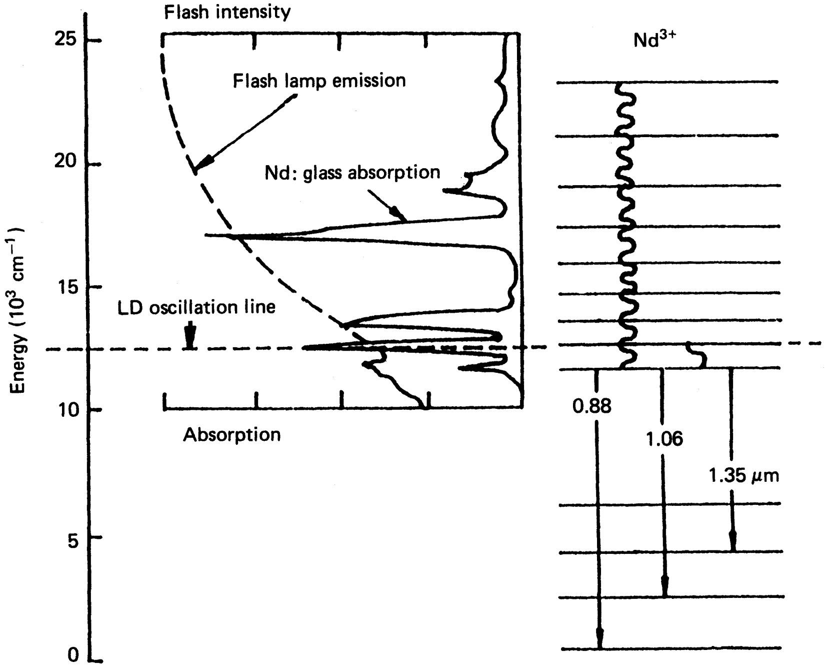


Figure 9.42 Spectrum of Neodymium: glass absorption, flash lamp emission and LD oscillation line. The transition and energy state of the neodymium laser is shown

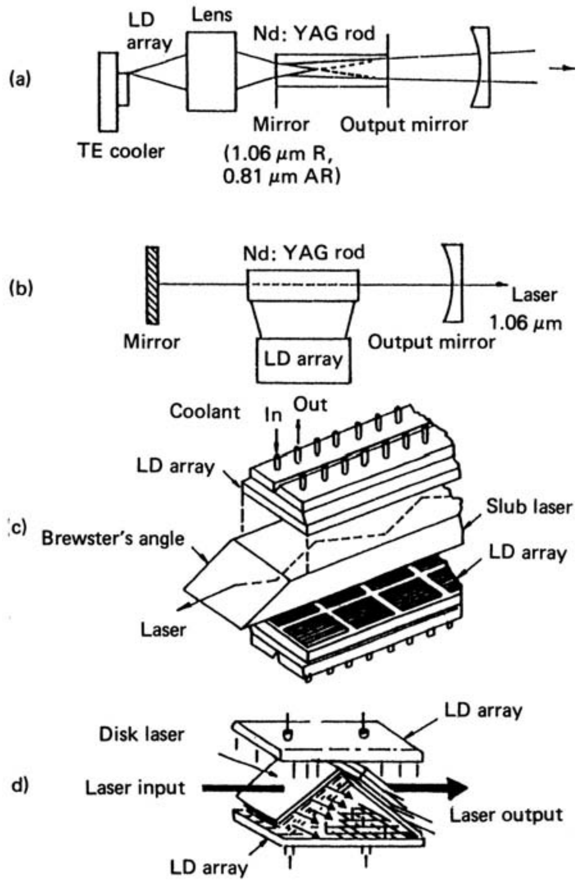


Figure 9.43 LD-pumped solid-state lasers and various layouts. (a) End pump; (b–d) Side pump

efficiency, low residual heat and solarisation for high repetition operation and long life. Solid-state lasers can be assembled as compact and easily manipulated devices by using LD pumping. *Figure 9.43* shows a schematic of an LD pumped solid-state laser. An end-on pumping scheme is suitable for micro-tip lasers with single-mode operation. A side-on pumping scheme is more flexible for assembly with high power lasers. The record for output power is 3.8 W CW with 10.9 W CW pumping or 4.9 W CW with 14 W CW pumping, with end-on pumping and 5 kW with side-on pumping.

Research trends for high-efficiency frequency-conversion technology (to green (2ω), blue (3ω) and UV (4ω) light) and for the development of new solid-state laser materials which have tunability or long fluorescence life-times are opening new fields of laser technology. Together with LD pumping techniques and in conjunction with gas, liquid or solid-state lasers, such trends are leading to new applications of laser technology in science, industry and medicine.

9.11.3.4 Tunable solid-state laser and new materials

The development of new solid-state laser materials is now proceeding rapidly, with the aim of achieving tunability and higher performance characteristics. Many materials capable of emitting at wavelengths between 680 and 2500 nm are under investigation. These are neodymium or chromium doped garium–scandium–gadrinium–garnet (GSGG), neodymium, erbium or holmium doped yttrium–lithium–frolide (YLF), neodymium doped YVO_4, etc. One of the well developed solid-state tunable lasers is the Ti : Sapphire laser which has a wide tunable range of 660∼1180 nm with a high power output.

9.11.4 Application of high-power lasers

9.11.4.1 Laser processing and machining

Laser interaction with matter The interaction of laser beams with matter may be divided into three categories with regard to both physical processes and applications: thermal processes, photochemical processes, and plasma processes.

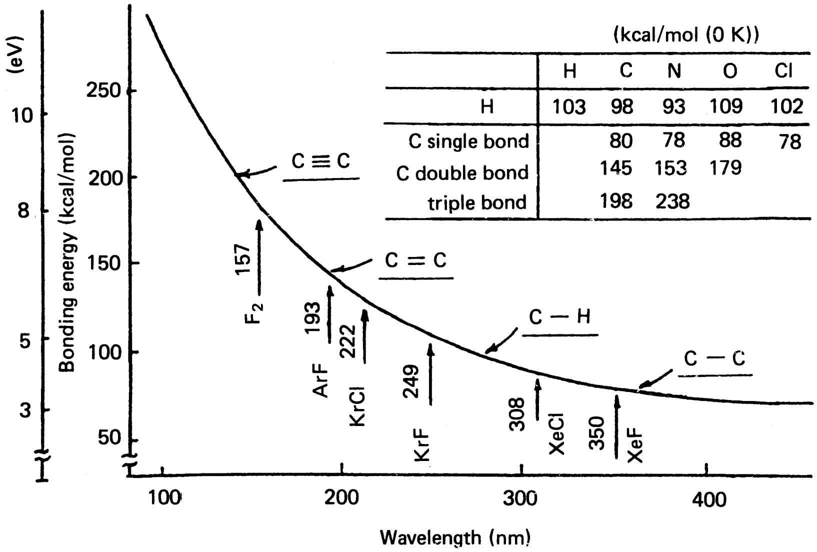
When a material is irradiated by laser light it absorbs the photon energy at the surface. The efficiency depends on the reflectivity of the surface and hence on such material properties as electrical conductivity or the existence of an appropriate absorption band. The heat on the surface then transfers to the substrate. This process can be calculated from thermal models when the thermal conductivity and heat capacity are known. At higher intensities, melting and evaporation occur. Laser-induced thermal effects provide a wide variety of laser machining technologies.

Another characteristic laser-induced process is the laser photochemical effect or laser-induced chemical reaction. The photochemical effect may be divided into three fundamental physical processes, excitation of molecular vibrational/rotational and electronic states, dissociation of molecules, and photoionisation. These processes may induce subsequent chemical reactions if appropriate reaction materials are introduced. Because laser light has a high degree of monochromaticity, a specific energy state can be selectively excited. This effect provides the means of controlling the chemical reaction by using a tunable laser. The correspondence of photon energy and wavelength produced by a number of typical excimer lasers with the chemical bonding energy of some organic radicals is shown in *Figure 9.44*. By using a short-wavelength laser, direct dissociation can be initiated.

At ultra-high laser emission intensities, plasma effects become dominant. In this region of interaction, temperatures of millions of degrees can be realised. Strong UV and X-ray radiation and the formation of multiply ionised atoms are observed.

Laser machining[6] Laser machining is dominantly based on the thermal processes induced by the interaction of the laser beam with matter. The characteristic features of laser machining are non-contact, rapid scan, fine spot and high energy density. Laser machining is one of the most widely distributed technologies currently used in industry. Its uses include drilling, cutting, joining or welding, material removal for trimming, marking and scribing, surface modification by transformation hardening, shock hardening, cladding and alloying. By controlling the heating and cooling rate, the crystalline structure of a surface can be manipulated to be amorphous, polycrystalline or non-equilibrium alloy.

Neodymium: YAG lasers or CO_2 gas lasers are widely used for these applications with CW operation or repetitive pulse operation to give a high average power.



(kcal/mol (0 K))

	H	C	N	O	Cl
H	103	98	93	109	102
C single bond		80	78	88	78
C double bond		145	153	179	
triple bond		198	238		

Figure 9.44 Chemical bonding energy and related wavelengths of excimer lasers

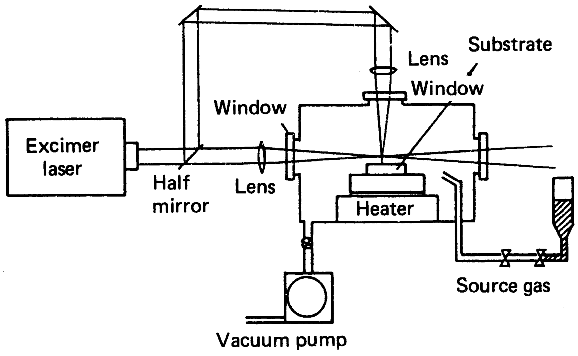


Figure 9.45 Laser CVD

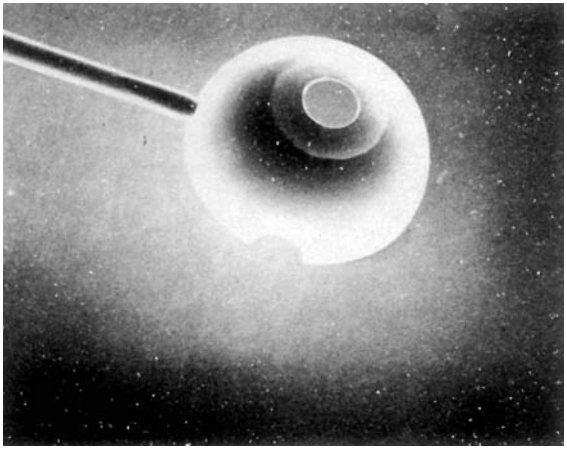


Figure 9.46 Excimer laser etching on a hemisphere of thin polyimide film

Laser photochemical processing Excimer lasers are now being used because of their short wavelengths for the initiation of photochemical processes, relying on their wavelengths being coincident with the bonding of the chemical compounds involved (*Figure 9.44*).

Laser chemical vapour deposition (CVD): a schematic diagram of a laser CVD system is shown in *Figure 9.45*. The source gas (for example Si_2H_6) is dissociated by a parallel laser beam (ArF excimer laser) and silicon is deposited on the surface of the substrate. In the case of *Figure 9.45*, the perpendicular beam is used to write some patterns on the substrate by the selective evaporation of the deposited thin layer.

Thin metal and dielectric layers as well as oxide may be formed by laser CVD through a combination of various source gases and laser wavelengths.

Thus laser CVD provides a means for a wide variety of processing techniques with low intrinsic damage to the substrate, and with the capability of maskless patterning.

Laser etching: laser-induced photochemical etching can be categorised into two different schemes, reactive photoetching and direct dissociative etching. Semiconductors or metals can be etched by irradiation with a laser in an atmosphere of reactive gas.

Plastics can be etched by direct dissociative etching without thermal effect by using short-wavelength lasers, the photon energy of which is higher than the bonding energy of the molecules.

Laser etching can be a low intrinsic damage process and maskless patterning is possible. *Figure 9.46* shows an example of patterning by direct dissociative etching of plastics.

Lithography: the process of microlithography provides a means of microprocessing integrated circuits. Shorter wavelengths are desirable and excimer lasers are being developed as the light sources. X-ray lithography is also under development. Plasmas produced by lasers are expected to be feasible as intense X-ray sources because of a high conversion efficiency from laser to X-ray (more than

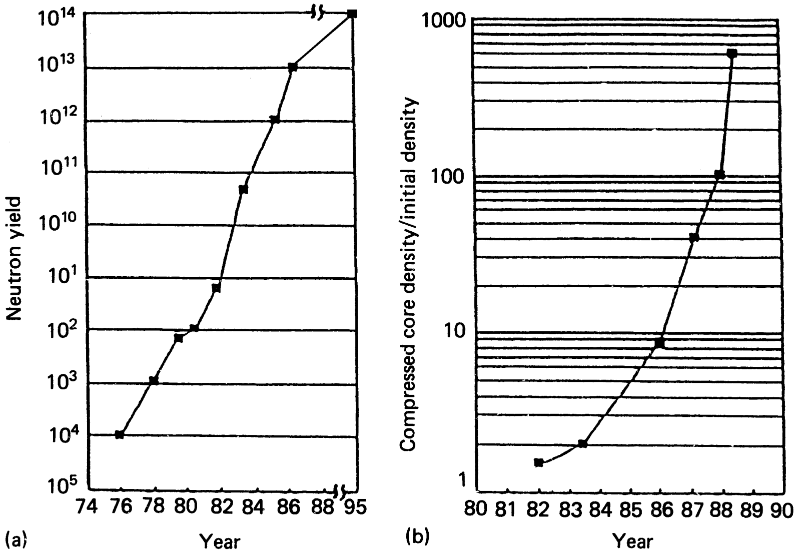


Figure 9.47 Progress towards laser fusion. (a) Neutron yield; (b) Compressed density

50%), a small point source of the order of the focal spot of the laser light, a bright source of 10^{12}–10^{13} W/cm^2 steradian, soft X-rays which are suitable for lithography, and a short X-ray pulse of the order of the laser pulse width.

These features of laser produced plasma X-ray sources can also be useful for X-ray microscopy and X-ray microanalysis.

9.11.4.2 Laser fusion[4,7]

Laser fusion is one of the most interesting applications of power lasers, where the many properties of a laser are fully utilised. The extreme concentration of energy in time and space which produces a high-density compression of fusion fuel has only become possible via laser implosion.

Focused laser beams are irradiated uniformly onto the surface of a fuel pellet and produce high-temperature plasma on the surface. The outward expansion of the plasma generates inward dynamic pressures which compress the hydrogen isotope fusion fuel. Adiabatic compression increases the density and also the temperature of the fuel and initiates the nuclear fusion. The progress which has been made towards realising laser fusion is illustrated in *Figure 9.47* via the enormous increases in compressed core density and neutron yield which have been achieved. The temperature and density which have been realised are 10^8 K and 1000 times the solid hydrogen density, respectively.

Glass lasers are widely used for present laser fusion experiments.

Table 9.13 Glass laser systems world-wide

Country	*Laboratory*	*Facility*	*Beam No.*	*Wavelength* (μm)	*Output* (kJ)	*Booster amplifier diameter* (cm)	*Final beam diameter* (cm)
USA	LLNL	NIF (FY2008-)	192	1.05/0.35	2000/1800	Disk 40 × 40	38 (square)
		Nova(1986–1999)	10	1.05/0.53/0.35	125/80/55	Disk 46	74
	LLE.Rochester	Omega up-grade	60	1.05/0.35	60/30	Disk 20	20
Russia	RFNC, IEP	Iskra-6 (proporsal)	128	1.05/0.35	500/300	Disk 20 × 20	20 (square)
	Lebedev	Delfin (-1996)	108	1.05	10	Rod 5	5
Japan	ILE.Osaka	Gekko XII	12	1.05/0.53/0.35	20/15/12	Disk 20	35
		Gekko PW (FY2000-)	1	1.05	1 (1PW)	Disk 35	35
		Gekko MII	2	1.05/0.53/0.35	2/1.5/1	Disk 20	20
France	CEA	LMJ (FY2008-)	240	1.05/0.53/0.35	2700/2000	Disk 40 × 40	38 (square)
	CELV	Phebus	2	1.05/0.53/0.35	20/10	Disk 46	74
	Ecole Polytech.	LULLI	6	1.05/0.53/0.25	0.3	Rod 8	8
UK	AWE	Helen (proporsal)	32	1.05/0.35	200/150	Disk 40 × 40	38 (square)
	Rutherford	Vulcan	8	1.05/0.53	2.5	Disk 10	10
China	SIOFM, Shanghai	Shen-Gang II	8	1.05/0.53	8	Disk 20	20
	Chengdu	Xingguang II	2	1.05/0.53/0.35	0.3/0.2	Disk 15	20
		Xingguang III (proporsal)	64	1.05/0.35	90/60	Disk 30 × 30	25 (square)
Italy	Frascatti	ABC	1	1.05/0.53	0.1	Rod 8	10

Laser systems which are now becoming operational are listed in *Table 9.13* along with several parameters of each system.

The next stage in the research and development of laser fusion is a demonstration of ignition and energy gain. The required laser energy is estimated to be 100 kJ to megajoules of green or blue laser light with a pulse width of the order of nanoseconds.

For energy production in a commercial laser fusion reactor, repetitive operation of 1–10 Hz with a megajoule pulse energy and with 5–10% laser efficiency is required. This corresponds to the need for a megawatt short-wavelength laser with high efficiency and high focusability, which would also have a wide range of industrial applications other than laser fusion.

9.11.4.3 Other applications of high-power lasers

The fields of diagnosis and measurement are where lasers have been primarily used since the early days of their evolution. Measurements of physical quantities such as position, velocity, acceleration, direction, surface configuration, and three-dimensional configuration, utilise the coherence and directionality of laser light. Scattering measurements of Thomson (electron), Rayleigh (atom or molecule), Mie (microparticle), Raman and Brillouin provide a wide variety of methods for diagnosing physical and chemical parameters of gases, liquids, solids and plasma. Monochromaticity of laser light is widely used in these applications.

Medical applications use the thermal effect of focused laser beams. Photochemical effects are being developed for medical treatment with the progress that is being made in realising short-wavelength high-power lasers.

9.11.5 Laser pumping methods and electric power supplies

The methods of pumping a laser with their associated power supplies depend on the type of laser. They may be classified as follows.

D.c. discharge through low pressure laser gas for HeNe, HeCd, low-pressure CO_2, and ion lasers. The characteristics of the power source depend on the type and design of laser through parameters such as gas pressure, length and bore radius of the discharge tube, mixing ratio of the laser gases and the buffer or energy transfer gases. The output required from the power source may vary from milliamperes to kiloamperes, and a few volts to kilovolts.

Pulse discharge with pre-ionisations for excimer lasers, TEA CO_2 lasers (transversely excited atmospheric pressure). To obtain a uniform discharge through the high-pressure laser gas, pre-ionisation by ultraviolet irradiation, or by surface corona discharge on the cathode is used. For the fast discharge needed for the pre-ionisation, a fast circuit with small capacitor and low stray inductance is needed in close proximity to the laser chamber. The main circuit for pumping discharge is composed of a fast capacitor bank (for the CO_2 laser), pulse forming network (PFN) with lumped capacitors and inductors or pulse forming line (PFL) with distributed transmission line. To obtain a faster rise of source power, a magnetic switch with a saturable core is used. A newly developed circuit for powering high-power excimer lasers is shown in *Figure 9.48*; this has spiker and sustainer circuits. A high repetition frequency of 1 kHz and an average output power of 2 kW can be obtained.

Electron-beam-controlled discharge for high-pressure, high-power CO_2 lasers. One of the most advanced electrical discharge concepts is electron-beam-controlled discharge. A uniform discharge through a high pressure gas up to a few tens of atmospheres and with an optimised E/p ratio for laser pumping can be achieved, where E is the electric-field strength and p is pressure. A uniform electron beam is introduced into the discharge region through a thin foil that separates the high-pressure region from a vacuum region where the electron beam gun is situated. Rectangular high voltage pulses of a few microseconds duration and a few hundred kilovolts amplitude are generated by a PFN with the capacitor elements of a Marx type generator.

Optical pumping of solid-state lasers with d.c. or pulse operation. For d.c. laser operations, a krypton arc lamp is widely used. The specifications for the power source depend

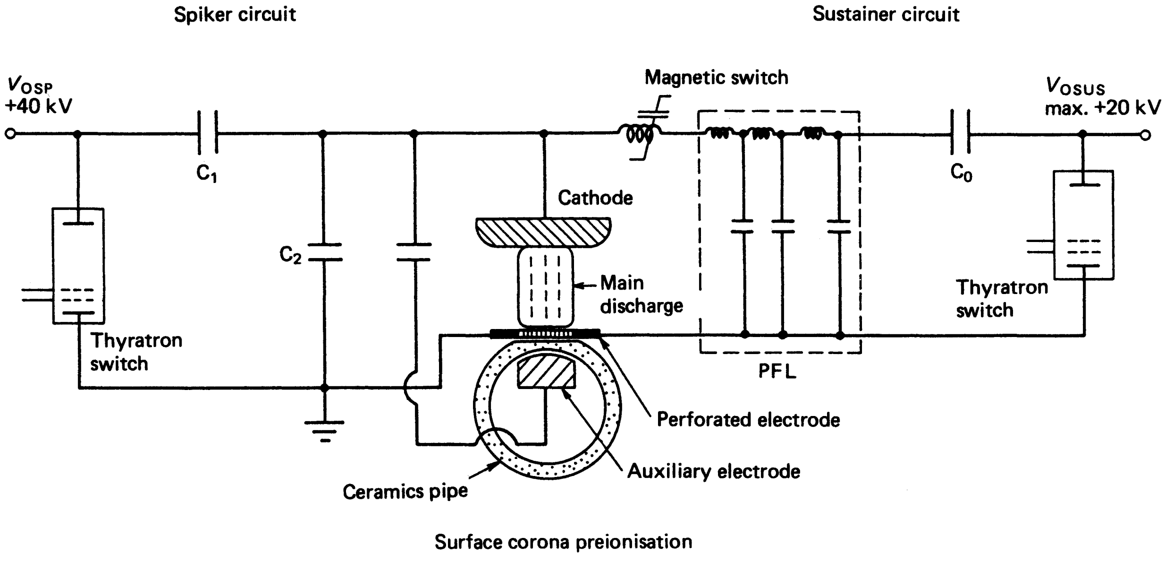


Figure 9.48 Combination of a surface corona pre-ionised discharge electrode with a spiker-sustainer pulse circuit (repetitive operation)

on the current density, bore radius, arc length, and gas pressure of the lamp. Typical discharge parameters are 100–200 V, 10–40 A, for a krypton arc lamp of 6 mm inner diameter, 150 mm arc length, with 4 atm gas pressure. A constant-current d.c. power supply with feedback stabilisation and a few kilovolts trigger circuit is used.

For pulse operations, a xenon flash lamp is widely used. The power source is usually PFN with the lamp elements, the length of current pulse is designed to match the fluorescent life-time of the laser material which is between 100 μs and 1 ms.

Recently, LD pumping has become a feasible technology for high-efficiency, d.c. or high repetitive-pulse operation of solid-state lasers. The most important feature of LD pumping is the matching of the LD emission spectrum to the absorption of the laser materials. The specification of the power source for an LD is 2–3 V and a few amperes for one element. LDs with an output of 1–150 W (quasi-CW) are commercially available and 420 kW with 10 Hz, 100∼400 μs pulsewidth with arrayed LD has been reported. LD-pumped solid-state lasers, together with frequency conversion, will become one of the key technologies of high-power laser systems of the future.

Acknowledgements

Peter Jones would like to thank his former colleagues M. Copsey, A. Dexter, J. Driscoll, J. Griffith, I. Harvey, P. Hayes, J. Hutchison, M. Lees, R. Perkin, D. Simpson, L. Smith and R. Townend for their contributions. Section 9.6 was based on an article originally published in *Power Engineering Journal* and subsequently in *Cast Metal Journal.*

References

1 *Laser Handbook,* The Laser Society of Japan (1982) (in Japanese)

2 BROMBERG, J. K., *Phys. Today*, **26** (October 1988)

3 NAKAI, S. ed. 'Power Laser Technology', Ohmsha (1999) (in Japanese)

4 'Energy from Inertial Fusion', International Atomic Energy Agency (IAEA), Vienna (1995)

5 CLEO (Conference on Lasers and Electro-Optics), is held every year and provides wide and up to date information on lasers

6 *Proceedings of International Conference on Laser Advanced Materials Processing—Science and Applications*, 21–23 May 1987, High Temperature Society of Japan, Osaka (1987)

7 NAKAI, S., 'Inertial confinement', *Nuclear Fusion*, **30**(9), 1779, 1963 (1990)

10

Welding and Soldering

W Lucas PhD, DSc, CEng, FIM, FWeldI, EWE
The Welding Institute, Cambridge

S Westgate BSc(Hons)
The Welding Institute, Cambridge

Contents

10.1 Arc welding 10/3
  10.1.1 Power sources for arc welding 10/3
  10.1.2 Manual metal arc welding 10/9
  10.1.3 Metal inert gas welding 10/12
  10.1.4 Flux cored arc welding 10/16
  10.1.5 Submered arc welding 10/17
  10.1.6 Tungsten inert gas welding 10/20
  10.1.7 Plasma welding 10/24
  10.1.8 Electroslag and electrogas welding 10/28
  10.1.9 Metal cutting and gouging 10/29

10.2 Resistance welding 10/29
  10.2.1 Welding equipment 10/29
  10.2.2 Welding process 10/32
  10.2.3 Projection welding 10/34
  10.2.4 Seam welding 10/34
  10.2.5 Resistance butt and flash welding 10/34
  10.2.6 Safety aspects of resistance welding 10/36

10.3 Fuses 10/36
  10.3.1 Fuse technology 10/36
  10.3.2 Element materials 10/37
  10.3.3 Filling materials 10/37
  10.3.4 Fuse-links with short operating times 10/37
  10.3.5 The M effect 10/38
  10.3.6 Composite or dual-element fuses 10/38

10.4 Contacts 10/38
  10.4.1 Low-voltage, low-current contacts 10/38
  10.4.2 Low-voltage, high-current contacts 10/38
  10.4.3 Contact design 10/41

10.5 Special alloys 10/42
  10.5.1 Heating alloys 10/42
  10.5.2 Resistance alloys 10/43
  10.5.3 Controlled-expansion alloys 10/43
  10.5.4 Heat-resisting alloys 10/43

10.6 Solders 10/44
  10.6.1 Fluxes 10/44
  10.6.2 Solder types 10/45

10.7 Rare and precious metals 10/45

10.8 Temperature-sensitive bimetals 10/45

10.9 Nuclear-reactor materials 10/46
  10.9.1 Introduction 10/46
  10.9.2 Fuels 10/46
  10.9.3 Fuel cladding 10/47
  10.9.4 Coolant 10/48
  10.9.5 Moderator 10/48
  10.9.6 Pressure vessel 10/49
  10.9.7 Shield 10/49

10.10 Amorphous materials 10/49
  10.10.1 Power-transformer materials 10/50
  10.10.2 High-frequency-device materials 10/50

10.1 Arc welding

The arc is an ideal means of generating intense heat of sufficient power to melt most materials, especially metals. The arc is formed between an electrode and the workpiece. The electrode can be of a refractory metal rod, such as tungsten, or a metal rod or wire, as shown in *Figure 10.1*. In the tungsten arc process, the arc is formed between the pointed tip of the electrode and the abutting faces of the two components being welded. The faces are melted and the bridge or weld is formed between the components. In the case of a metal electrode, the electrode fulfils the same function as the tungsten electrode in forming a point heat source, but in this case, the electrode is consumed in the course of welding and the molten metal from the electrode helps to bridge the gap between the components being welded.

The welding current can be either d.c. or a.c., but in either case the voltage, in particular, must be sufficient so that the arc discharge or plasma can be sustained. With the tungsten electrode, the arc formed is in a gas flow of argon or helium gas, whilst in the metal arc processes, the plasma is formed either from CO_2 or an argon based gas with a small amount of oxygen or CO_2. The plasma reaches a temperature of the order of 15 000 K, at least in the core, but it is noteworthy that most of the heat is associated with the arc root mechanisms. The voltage distribution across the arc shows three distinct regions, the anode drop, plasma column drop and the cathode drop, *Figure 10.2*. The anode and cathode voltage drops are determined by the electrode material and the ionised gas compositions, but for welding applications the plasma drop is a function of the arc length.

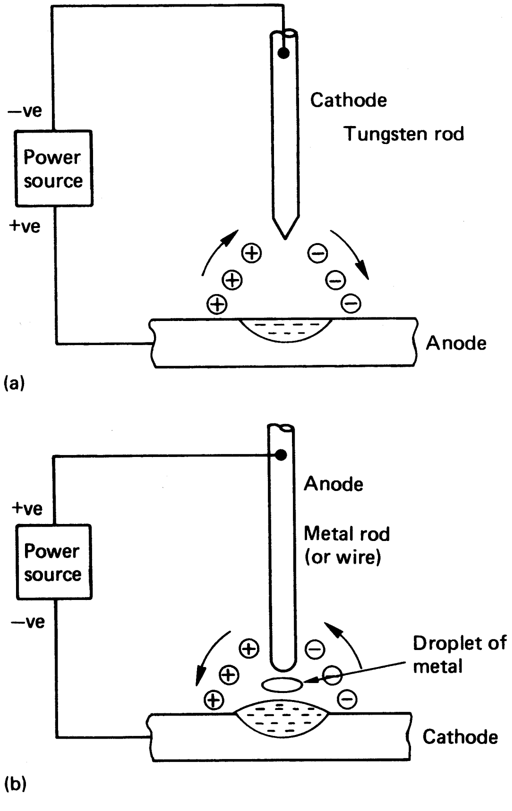


Figure 10.1 Principles of operation for the tungsten electrode and metal electrode welding processes. (a) Tungsten electrode; (b) metal rod (or wire) electrode

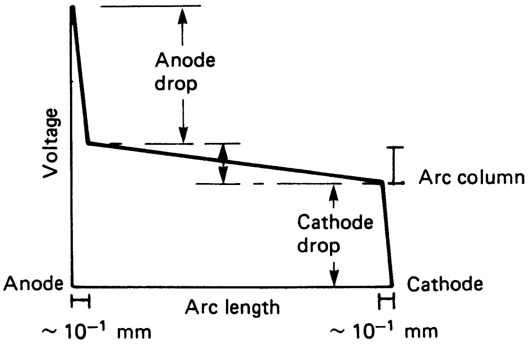


Figure 10.2 Voltage distribution across a tungsten arc

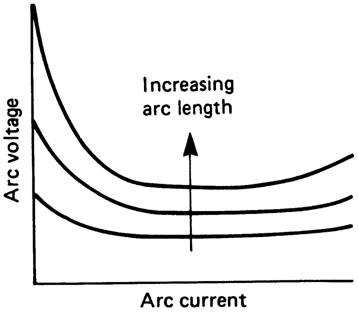


Figure 10.3 Typical voltage–current arc characteristics

The arc voltage–current characteristics are shown in *Figure 10.3*. There are three regions, depending on the current level. At low currents, the voltage–current relationship has a negative slope (as the current increases, the voltage decreases), the flat portion at intermediate current levels is essentially a constant voltage irrespective of current, and at high currents the slope of the voltage–current relationship is positive.

10.1.1 Power sources for arc welding

The basic function of the arc welding power supply is to provide sufficient power to melt the joint between the parts and, if required, the additional weld metal provided by a filler wire. As most arc welding processes require a high current, (50–500 A) but at relatively low arc voltages (10–50 V), the high-voltage mains supply (240 or 440 V) must be reduced. Thus, in its basic form the arc welding power source comprises a transformer to reduce the mains voltage and increase the current. For d.c. a rectifier is placed on the secondary side of the transformer (*Figure 10.4*).

The conventional d.c. power source operates on either a single-phase or a three-phase a.c. supply. The disadvantage of the former is that, even with full wave rectification, the output waveform contains a high degree of ripple, as shown schematically in *Figure 10.5*; a high level of ripple can result in poor arc stability at low current levels. A much smoother d.c. output is provided by a three-phase input and a bridge rectifier which produces full rectification, as shown schematically in *Figure 10.6*. Nevertheless, even with the three-phase output, the final current waveform will contain a

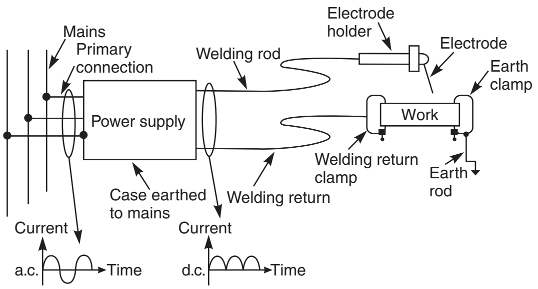

Mains
Primary connection
Welding rod
Electrode holder
Electrode
Earth clamp
Power supply
Work
Welding return clamp
Earth rod
Case earthed to mains
Welding return
Current
a.c.
Time
Current
d.c.
Time

Figure 10.4 Reducing high voltage using a transformer

significant amount of 300 cycle ripple (as shown in *Figure 10.6*), which can noticeably affect the shape of the arc and its stability. The more advanced, so-called 'electronic', control systems have been designed to provide an accurate d.c. output, with high response and sophisticated control of the output.

10.1.1.1 D.c. power-source control

Traditional power-source control uses a variable reactor, moving coil or moving iron transformer, or a magnetic amplifier to control the welding current. Such equipment has the highly desirable features of simple operation, reliability and robustness, making it ideally suited for application in aggressive industrial environments. The disadvantages are relatively high material cost, large size, and limited accuracy with slow response. The latter feature, in particular, can limit the performance of the power source, especially when sophisticated control of the output is required.

In recent years, electronic power sources employing high power semiconductors, have become available which do not suffer from these disadvantages. The various types of electronic power source are:

(1) thyristor phase control;
(2) transistor, series regulator;
(3) transistor, secondary switched (chopper); and
(4) a.c. line, or primary, rectifier plus inverter (SCR or transistor).

The advantages and disadvantages of these types of power source compared with conventional variable reactor or magnetic amplifier power sources are listed in *Table 10.1*. The main features of the different power sources are described below.

Thyristor-phase control Of the power source designs listed in *Table 10.1*, thyristor control represents an excellent compromise in terms of performance and cost. In this power-source design, the output is controlled by the phase angle of the a.c. voltage, at which the thyristors (SCRs) are switched on. Whilst high current sources have a three-phase input, low current sources normally have a single-phase input with current control on the primary side of the

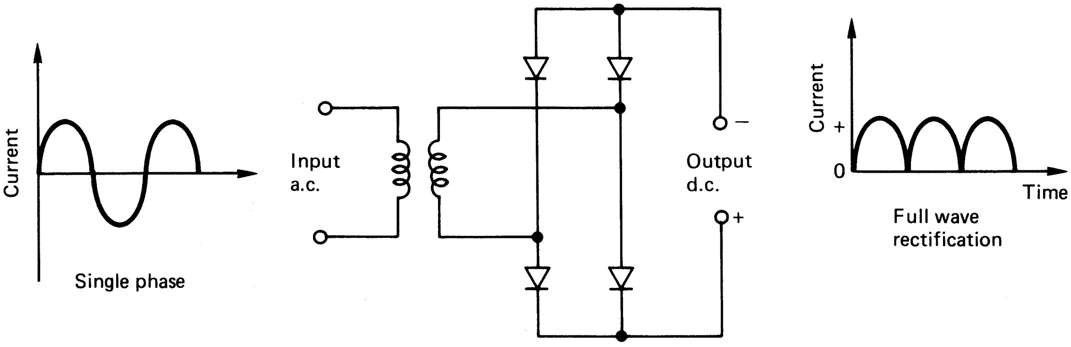

Current
Single phase
Input
a.c.
Output
d.c.
−
+
Current
+
0
Time
Full wave rectification

Figure 10.5 D.c. power source with single-phase supply

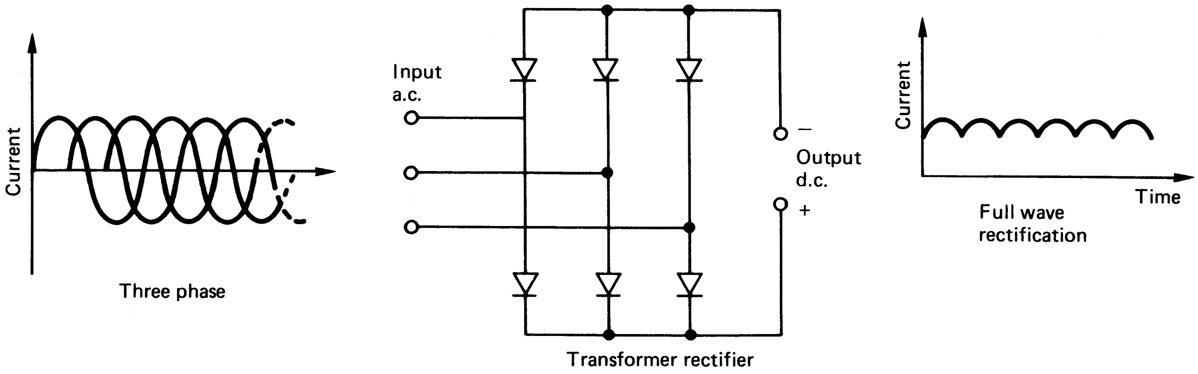

Current
Three phase
Input
a.c.
−
Output
d.c.
+
Transformer rectifier
Current
Time
Full wave rectification

Figure 10.6 D.c. power source with three-phase a.c. supply

Table 10.1 Major operational features of electronic power sources compared with conventional variable reactor or magnetic amplifier power sources

Control type	*Method of control*	*Advantages*	*Disadvantages*
Thyristor, phase	Thyristors replace diodes on secondary output of the transformer. Alternatively, triacs or inverse parallel thyristors are used in the primary of the transformer	Better accuracy of current and time settings. Can be used to produce square-wave a.c. waveform. Can be used for pulsed operation	High ripple unless large amount of inductance is placed in series with output. Pulsed response normally limited to 100 Hz
Transistor, series regulator	Power transistors in parallel; analogue control from low current input signal	Very stable and accurate control of current level—better than 1% of set level. Pulsing over wide range of frequencies (up to 10 kHz), and pulse shape can be varied	Poor electrical efficiency. D.c. supply only
Transistor, switched	Transistor, high frequency switching (20 kHz) of d.c. supply	Accuracy and control similar to series controller. Less wasteful of energy compared with series regulator. Greater arc stiffness can be exploited for low current operation	Although the output is similar to that of a series regulator, pulse frequency and wave shaping is less flexible
A.c. line rectifier, plus inverter	Mains supply rectified to high voltage d.c. and then converted by transistors or thyristors to a.c. operating at 2–50 kHz. Final output produced by small mains transformer and rectified to d.c.	Because the transformer operates at high frequency, the size and weight of the mains transformer can be greatly reduced. Because of its small size, the cost of raw materials is significantly reduced. Accuracy of control is better than with a magnetic amplifier	Rapid switching can cause an unpleasant whining noise from the arc

transformer. The output waveform suffers from inherent ripple, the frequency being a simple multiple of the mains supply, i.e. 100, 150 or 300 Hz. Whilst the ripple is not normally a problem at medium or high welding currents, excessive ripple can cause problems in arc starting and in operating at currents below 20 A.

Transistor-series regulator The transistor based control systems offer highly desirable features for precision welding, namely accuracy and reproducibility of the welding parameters. The analogue type control (*Figure 10.7*(*a*)) where the current flowing through all the transistors is regulated, provides a ripple-free and easily controlled output. The output, under feedback control, can be pulsed at a frequency within the kilohertz range and the pulse waveform and the overall operating sequence can be precisely shaped. However, the regulator stage tends to be wasteful of electrical energy as the excess power in the transistors must be dissipated by water cooling; as the difference in the voltage between the arc and the open circuit appears across the transistors, this excess power (voltage difference times the current) is dissipated by the transistors which are usually cooled by mounting on water-cooled copper heat sinks.

Transistor-switched Switching transistor control, in which each transistor operates in either an 'on' and 'off' state, (*Figure 10.7*(*b*)) is an attractive alternative to analogue control. The output is determined by the ratio of the 'on' time to the 'off' time. As power is dissipated at the moment of switching, this design is more efficient which obviates the

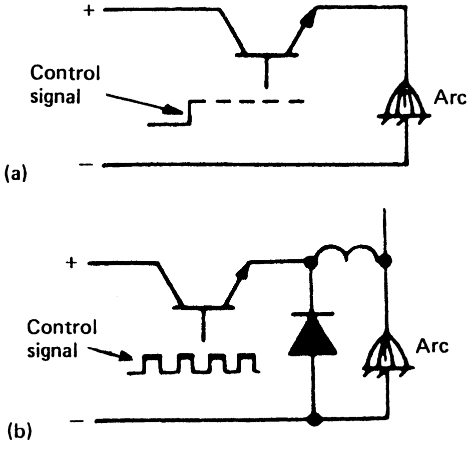


Figure 10.7 Transistor controlled power sources. (a) Series regulator transistor control; (b) Switching on/off transistor controlled power source

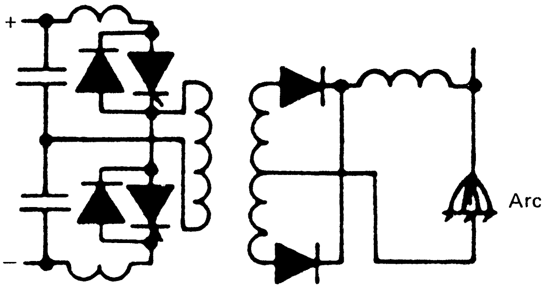


Figure 10.8 High frequency inverter power source

need for water cooling. The output is 'chopped' at a frequency usually in excess of 10 kHz, which has the beneficial effect of constricting the arc shape, thereby making it more directional. Thus, switching transistors have enabled the advantageous features of transistor control (high accuracy and reproducibility) to be exploited without the limitations of the analogue control system, particularly its high energy consumption. Various models of this type of power source have also been designed for specific welding operations, e.g. for welding thin sheets at low current levels, or for high-speed welding at higher current levels.

A.c. line rectifier plus inverter The a.c. line rectifier, plus inverter, type of control (*Figure 10.8*) represents a basically different approach to power source design. Compared with the normal mains frequency transformer/rectifier, the mains supply is first rectified and stored by a capacitor. The medium-voltage d.c. is then converted to a.c. at a frequency in excess of 500 Hz by means of switching transistors or thyristors. The high-frequency a.c. is finally transformed down to a voltage suitable for welding and then rectified to d.c. to provide the normal range of welding current levels. As the transformer operating within the range 20–100 kHz can be substantially reduced in size, compared with conventional transformers for use at mains frequency, higher electrical efficiency and power factor and small size have been achieved. The main limitation at this stage of development concerns the performance of the switching devices which must withstand a high hold-off potential in excess of approximately 1.5 kV. With regard to the operation of the power source itself, the relatively high level of arc noise can be an annoying factor; the noise is related to the switching frequency and power sources operating at frequencies above 20 kHz do not suffer from this problem.

10.1.1.2 A.c. power source control

The simplest a.c. power source is the single-phase type, with moving core control. *A.c./d.c.* power sources have moving core or thyristor control with the thyristors (phase angle control) acting on either the primary or the secondary side. The problem of operating an a.c. welding arc is that a high open-circuit voltage is required to ensure that the arc re-ignites on polarity reversal. For example, when welding with a metal electrode, a covering which contains easily ionised elements is required to sustain the arc, and sine wave power sources are normally operated with an open-circuit voltage of 80 V. When using a tungsten electrode which does not have a flux coating, high voltage or high-frequency oscillation is also applied continuously to ensure arc re-ignition (*Figure 10.9*).

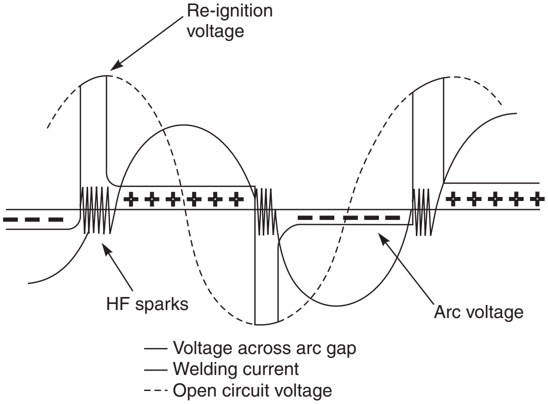


Figure 10.9 AC waveform showing voltage required to re-ignite the arc

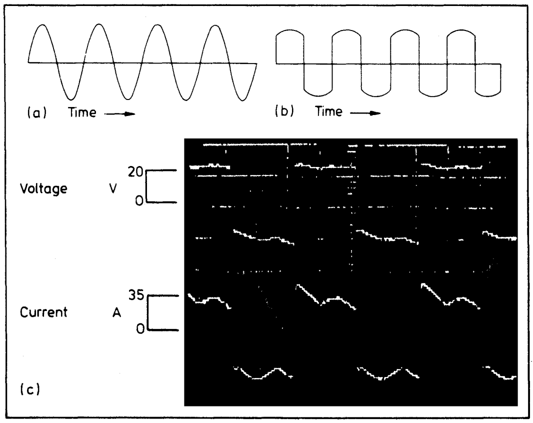


Figure 10.10 Characteristic re-ignition waveforms for: (a) sine wave supply at an open-circuit voltage of 100 V; (b) switched d.c. supply at an open-circuit voltage of 75 V; (c) output waveform of a square wave a.c. power source

In an alternative form of power source the output current assumes a more square waveform as compared with the conventional sine wave (*Figure 10.10*). Two types of power source are available: the squared sine wave a.c., and the switched d.c. power source. The squared a.c. wave is generated by using inverted a.c. (*Figure 10.11*(*a*)), whilst the more truly square waveform is produced by directly inverting a d.c. supply (*Figure 10.11*(*b*)). The rates of circuit current rise for the two types of power source on polarity reversal are typically 110 A in 0.1 ms and 160 A in 0.02 ms for the square wave a.c. and the switched d.c. power source, respectively. In comparison, the rate of change of current for conventional sine wave a.c. can be as much as two orders of magnitude less than the switched d.c. power source; a time of approximately 3 ms is required to achieve a circuit current of 110 A from zero with 50 V across the arc gap.

The advantages of rapid current build up on polarity reversal are:

(1) a reduction in the open-circuit voltage;
(2) easier arc re-ignition; and
(3) reduction in electrode heating.

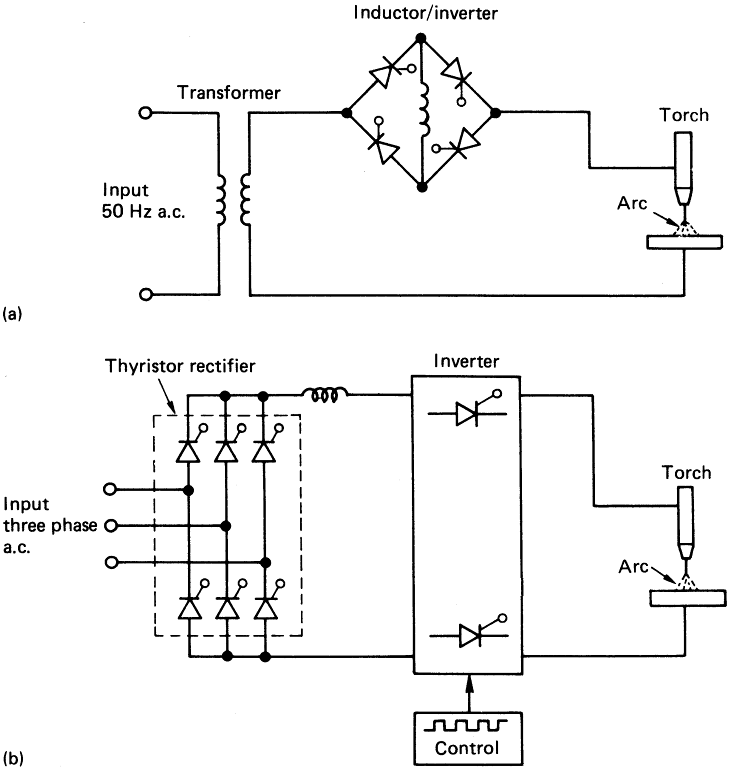


Figure 10.11 Square wave d.c. power source designs: (a) 'squared' a.c.; (b) switched d.c.

These advantageous features of square wave power sources are discussed when considering specific arc welding processes (see Sections 10.1.3 to 10.1.8).

10.1.1.3 Intelligent control

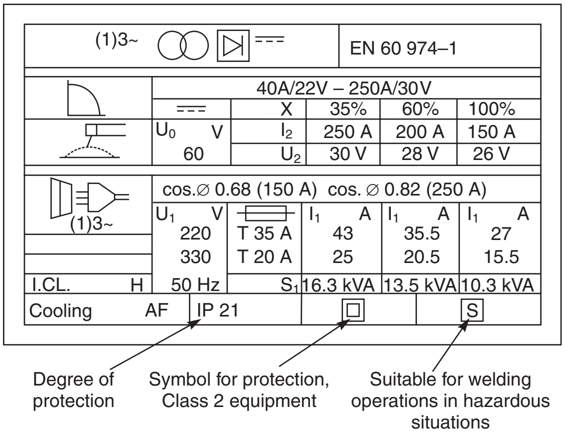
Whilst the earliest welding power systems implemented control using electronic analogue feedback, today a number of power sources on the market employ digital control concepts. Digital control systems have the advantage of being cheaper to develop, less susceptible to the degradation of the electronic components, and easier to update and modify.

Employing software to control the power source output enables intelligence to be included in the control system. For example, power sources have the capacity to store welding parameters and welding programmes tailored to the application i.e. type of welding process, material etc. As an indication of the increase in the intelligence of power sources, the memory requirements have increased typically from 50 kb for the early inverter power sources to 500 kb for the latest designs.

Power sources are also available which implement the control strategy using fuzzy logic. This approach makes it easier to develop sophisticated and intuitive control strategies than would otherwise be possible using traditional coding techniques. Fuzzy logic is being applied in the dynamic control of the welding arc.

10.1.1.4 Rating plate/power source specification

In selecting a suitable power source design for a welding process, it is important to check that the specification of the power source will be suitable for the intended job. Information about a new power source can be obtained direct from the manufacturer or from the product literature. For an existing power source, the handbook can be consulted but often the rating plate information is all that is available. Any power source controlling the welding parameters made to a National, European or International standard should have a rating plate similar to the one shown in *Figure 10.12*. This plate provides information about the



(1)3~			EN 60 974-1		
	40A/22V – 250A/30V				
	⎓	X	35%	60%	100%
	U_0 V 60	I_2	250 A	200 A	150 A
		U_2	30 V	28 V	26 V
(1)3~	cos.∅ 0.68 (150 A) cos. ∅ 0.82 (250 A)				
	U_1 V		I_1 A	I_1 A	I_1 A
	220	T 35 A	43	35.5	27
	330	T 20 A	25	20.5	15.5
I.CL. H	50 Hz	S_1	16.3 kVA	13.5 kVA	10.3 kVA
Cooling AF	IP 21		☐		S

Degree of protection
Symbol for protection, Class 2 equipment
Suitable for welding operations in hazardous situations

Figure 10.12 Rating plate for arc welding power sources

manufacturer, supply requirements, performance and suitability for use. For equipment built in the last few years the rating plate should conform to IEC 60974-1 (EN 60974-1).

The following symbols are used on the manufacturer's rating plate to indicate the type of protection:

⧈	Insulation protection, class 2 equipment.
S	Power source suitable for supplying power to welding operations carried out in an environment with increased hazard of electric shock.
IP XX	Degree of protection of the enclosure.

The symbol S refers to the suitability of the power source for operating in environments with increased hazard of electric shock. The rated no-load voltage of such power sources shall not exceed:

D.c.	113 V peak
A.c.	68 V peak and 48 V r.m.s.

The symbol S is also often displayed on the front of the power source. The code for the degree of protection which refers to risk of electric shock in normal service from direct contact is as follows:

IP2X	The 2 refers to protection to solid objects larger than 12 mm. e.g. fingers.
IPX1	The 1 refers to protection from vertically falling drops of water.
IPX3	The 3 refers to protection from drops of rain water up to 60° from the vertical.

The minimum degree of protection for welding power sources is IP 21. However, power sources specifically designed for outdoor use shall have a minimum degree of protection of IP 23.

10.1.1.5 Welding installations

Typical arc welding installations for both single and multi-welder operations are described in National guidelines e.g. the UK HSE Guideline No 118, 'Electrical Safety in Arc Welding'. In connecting the welding circuit, the following best practice should be adopted:

- the connection between the welding power source and the workpiece should be as direct as practicable;
- use insulated cables and connection devices of adequate current carrying capacity;
- extraneous conductive parts should not be used as part of the welding return circuit unless part of the workpiece itself;
- the current return clamp should be as near to the welding arc as possible.

When attaching the welding current and current return cables, it is essential that an efficient electrical contact is achieved between the connection device and the workpiece to prevent overheating and arcing. For example, the current and return clamps must be securely attached to 'bright' metal i.e. any rust or primer coatings should be locally removed.

Power source and earthing The normal practice in the UK has been to provide a separate earth connection to the workpiece. The reason is that, in the unlikely event of an insulation breakdown between the primary and the secondary circuit, the fuses will blow. However, the separate earth connection increases the risk of stray currents which may cause damage to other equipment and conductors.

As modern power sources have been designed to have a much higher level of insulation (termed double or reinforced insulation), a separate earth connection is not recommended.

There is a potential problem in that both power source designs can often be found in the same welding shop. The newer (double or reinforced insulation) power source can be identified by reference to the power source's rating plate (*Figure 10.12*) which will indicate that it has been manufactured to current standards e.g. EN 60974-1 or IEC 60974-1.

In older power source designs, the welding circuit was sometimes connected internally to the power source enclosure. However, the danger is that even with the welding return lead disconnected, and a separate earth connection, welding is possible with the current flowing through the earth. Because of the risk of damaging protective earth and other connectors, this type of power source is considered to be obsolete and should not be used.

Supply requirements Power sources will require either a single- or three-phase supply at 230 V a.c. or 400 V a.c. In many parts of Europe the 230 V supply is 16 A, but in the UK the standard ring main is 13 A. Therefore, the relatively low power output of the 230 V system is further reduced if a 13 A plug is fitted and a dedicated circuit may be required.

Three-phase supplies may be limited to 30 A, but higher power welding equipment may require a 45 A or even 60 A supply. The effective current is displayed on the rating plate. This value should be used to determine the cable size and fusing requirements.

Apart from the obvious hazards associated with overloading a supply, e.g. overheating and blowing fuses, problems with other equipment may be caused. If the supply has a high impedance (commonly known as soft) as may be the case in rural areas supplied by overhead cables, a high current draw may cause the voltage of the supply to fall below levels which may cause problems with other equipment.

Engines and motor drives For equipment not supplied from the mains, such as generators, information is provided about the characteristics of the motor, load and idle speed and power consumption. While this type of equipment may be ideal for welding outside in remote locations, it tends to generate a high level of acoustic noise. The noise levels are limited by an EU directive but can be as high as 97 dBA.

Environment All power sources should be marked with an IP rating which provides information about the degree to which the equipment is protected against water. Normally equipment will be IP 21 or IP 23. The first number means that it should not be possible to touch live parts and the second number describes protection against water ingress. A power source marked IP 21 is protected from vertical drops of water, as for example may occur if the roof leaks. If the equipment is marked IP 23, this means that the equipment is protected from water at up to 60° from the vertical and is thus suitable for outdoor use. Additional letters to the IP number indicate tests with or without the fan running and increased mechanical protection from electrical hazards.

Areas of increased risk are in wet or humid conditions, confined spaces or when the welder is exposed to large areas of bare metal. For use in this type of environment, it is important to use a power source with an S mark which means that the no-load voltage is below 48 V r.m.s. a.c. or 113 V peak d.c. If the power source has a higher no-load voltage, a voltage reduction device should be used which will limit the voltage at the holder to approximately 25 V.

10.1.1.6 Relevant standards

The relevant standards for arc welding power sources, equipment and accessories are:

EN 60974-1:1998, IEC 60974-1:1998, 'Arc welding equipment—Part 1: Welding power sources'

EN 60974-7:2000, 'Arc welding equipment, Part 7, Torches'

EN 60974-11:1995, 'Arc welding power equipment, Part 11, Electrode holders'

EN 60974-12:1995, 'Arc welding equipment, Part 12, Coupling devices for welding cables'

EN 169:1992, 'Specification for filters for personal eye protection equipment used in welding and similar operations'

EN 60529:1992, 'Specification for degree of protection provided by enclosures' (IP codes)

EN 470-1:1995, 'Protective clothing for use in welding and allied processes—general requirements'

EN 50199:1996, 'Electromagnetic compatibility (EMC)—Product standard for arc welding equipment'

EN 50060:1989 'Power sources for manual metal arc welding with limited duty'

10.1.2 Manual metal arc welding

10.1.2.1 Principles of operation

The manual metal arc (MMA) process is the most versatile welding method. It can be used to weld most materials in a wide range of thicknesses and in all welding positions. The basic principles of the process are shown in *Figure 10.13* and the essential features are a central metal rod surrounded by a flux covering. The function of the flux is as follows:

(1) The vaporised flux contains easily ionisable elements which help to stabilise the arc.
(2) The molten flux surrounds the molten droplets to prevent oxidation.
(3) The presence of slag on the weld pool and the gas shroud of mainly carbon dioxide, prevent oxidation of the weld pool.
(4) The composition of the flux also influences the chemical composition of the weld metal and hence its mechanical properties.
(5) The properties of the molten flux and the flux residue or slag, particularly its fluidity and rate of freezing, determine the so-called 'handleability' of the electrodes and their ability to weld in positions other than flat; a viscous fast-freezing slag is preferred for welding in vertical and overhead positions.
(6) The slag also has an influence on the final shape of the weld bead—for example, a fluid slag/weld pool will 'wet' more smoothly into the parent metal producing a flat weld bead, whilst a less fluid slag will produce a more convex weld bead profile.

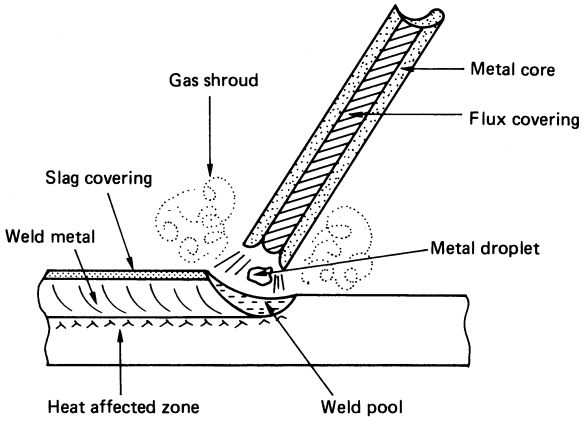


Figure 10.13 Principles of operation of the manual metal arc process

10.1.2.2 Electrode

The MMA electrode which has a diameter of typically 3.25, 4.0 or 5.0 mm, is classified into one of three main types according to the flux covering (cellulosic, rutile or basic). Iron powder may also be added to the covering to increase the deposition rate and to improve the arc stability and smoothness of operation.

Cellulosic electrodes contain a high proportion of cellulose in the covering and are characterised by a deeply penetrating arc with a thin slag covering on the weld pool and weld bead. The low slag volume makes these electrodes highly suitable for vertical down and pipe welding and the resulting mechanical properties of the weld metal are reasonably good. The main disadvantage is the high level of hydrogen generated which increases the risk of hydrogen induced cracking in the heat affected zone. Cracking in this zone is preventable by preheating the material prior to welding which allows the hydrogen to diffuse away from the hard crack susceptible heat affected zone. The preheat temperature is determined by the thickness of material, the carbon equivalent (as determined by the chemical composition), the amount of hydrogen generated in the weld metal and the heat input.

Rutile electrode coverings contain a high proportion of titanium oxide (rutile). The presence of titanium oxide promotes easy arc ignition and smooth arc operation and low spatter. Iron powder can be added to improve metal deposition, and recovery rates above 150% can be achieved without deterioration of the arcing characteristics. The lower weld metal mechanical properties and high hydrogen levels (ca. 20 ml/100 g) make these electrodes less suitable for welding the higher strength steels or thicker materials.

Basic electrodes contain a high proportion of calcium carbonate and calcium fluoride in the form of limestone and fluorspar in their coverings. The electrodes are used for high quality (high strength/good toughness) and low hydrogen weld metal. Low hydrogen electrodes can produce deposits with less than 5 ml/100 g hydrogen. The slag from the covering melts at a lower temperature than that from rutile electrodes and has a higher surface tension. The welding characteristics are termed 'fast freezing', which promotes welding in the vertical position and permits higher welding currents and faster welding speeds. The main disadvantages are the convex weld bead surface profile, and the slag can be more difficult to remove from the weld surface.

10.1.2.3 Power source

As the metal rod electrode is consumed under the action of the arc with the droplets forming the weld bead, the function of the power source can be viewed as simply providing a current for melting. However, the power supply must be capable of providing sufficient voltage to sustain the arc, a rapid rate of rise to a high current level to initiate the arc and to clear the periodic short circuiting of the electrode to the workpiece, and constant current during the welding operation itself. With regard to accommodating short circuits, it must be noted that small diameter molten droplets of metal are not projected across the arc from the end of the rod, i.e. in free flight; large droplets form on the tip of the electrode which then transfer as large globules or during short-circuit bridging of the arc gap. The power supply must provide sufficient current to clear the short circuit, but at a controlled rate of rise to prevent explosive rupturing of the metal bridge.

The operation of the power source can be described in terms of the arc characteristics and the static and dynamic characteristics of the power source. The arc characteristic is nominally flat or slightly rising over the range of current levels (as shown in *Figure 10.3*). The power source static characteristic (*Figure 10.14*) is drooping so that there are well-defined current levels according to the arc length, and the current is essentially constant for a given current setting. Also shown in *Figure 10.14* is the open-circuit voltage which is available both to strike the arc and to ensure arc resignation in the operation. The short-circuit current is the maximum current available for clearing short circuits. Because of the steeply drooping characteristic, the current will only vary by approximately ±10% for changes in the arc length.

The welding current can be adjusted by varying the secondary connections of the transformer which produces a series of voltage–current characteristics. In the more usual design of power sources, the inductance can be adjusted which has the advantage of maintaining a high open-circuit voltage (*Figure 10.14*).

The dynamic characteristics of the power source relate to the way in which the power source reacts to variation in the load, e.g. during arc ignition, a short circuit, a momentary arc outage, or fluctuations in the arc length.

In the UK, the specification for the design and construction of welding power sources is EN60974 *Arc welding equipment Part 1 Welding power sources*; hobby transformers are covered by EN 50060: 1989. The load voltages for air-cooled power sources can be calculated using the equation

$$U_2 = (20 + 0.04 I_2)\ V \text{ (up to 44 V at 600 A and above)}$$

where U_2 is the conventional load voltage, and I_2 is the conventional welding current.

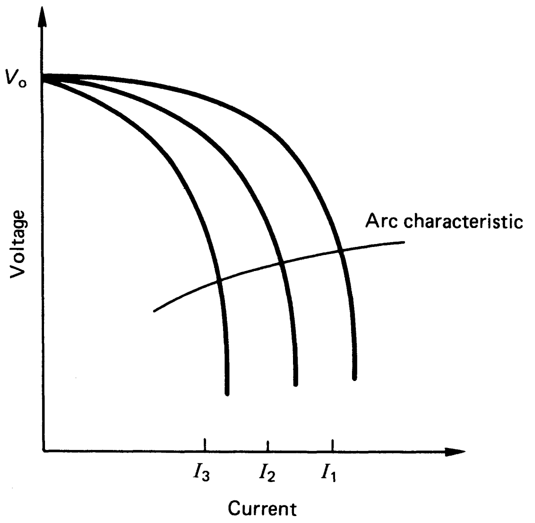


Figure 10.14 Variation in static output characteristics of the power source to adjust the welding current. V_o, open-circuit voltage

The standard also specifies the following maximum open-circuit voltage for oil- and air-cooled power supplies:

a.c. equipment	80 V (r.m.s.)
d.c. > 10% ripple	80 V (r.m.s.)
d.c. < 10% ripple	100 V (r.m.s.)

EN 609774-1:1998 does not specify a ripple level, but the maximum no-load voltage is 113 V peak for d.c. and 68 V peak and 48 V r.m.s. for a.c.

The maximum permissible no-load voltage (open-circuit voltage) for normal a.c. welding has been agreed by the European Committee for Electrotechnical Standardisation (CENELEC) to be 80 V. However, in confined spaces or wet conditions, the open-circuit voltage can be further restricted to 42 V, and 12 V in certain European countries. As 12 V, in particular, is insufficient to operate the process, a voltage-reduction device is incorporated in the equipment which allows the full open-circuit voltage to be obtained when the electrode is short circuited to the workpiece.

The minimum open-circuit voltage and electrode polarity requirements for the various types of electrode are given in BS EN 499:1995—*Welding Consumables—Covered Electrodes for Manual Metal Arc Welding of Non Alloy and Fine Grain Steels—Classification*. The requirements can be related to the various electrode types. For example, most cellulosic electrodes require an open-circuit voltage of > 50 V, but some electrodes can be used with either d.c. electrode-positive (DCEP) or d.c. electrode-negative (DCEN) and require a low open-circuit voltage of only 50 V for a.c. operation. Basic electrodes are the more demanding for a.c. operation and require a higher open-circuit voltage (>60 V). D.c. operation is the more normal electrode polarity.

10.1.2.4 Single operator transformer equipment

The most common form of single-operator equipment comprises a transformer wound on the primary side for normal low-voltage (1 V) mains and on the secondary side for 60–100 V, together with a tapped inductor (*Figure 10.15*) or moving-core inductor. It is preferable to connect the primary across the two lines of a three-phase supply rather than between line and neutral. Sets are produced with T or open-delta connection to three-phase supplies, but as the single-phase load cannot actually be balanced over three phases no great advantage is obtained. The voltage-reduction device is placed in the secondary circuit of the transformer, i.e. between the output side of the welding equipment and the electrode holder.

10.1.2.5 Two-operator equipment

Transformer welding sets are also available for two welders who can work simultaneously. These may consist of two single-phase units included in a single tank (*Figure 10.16*) or comprise a single transformer to which two variable inductors are connected in parallel on the secondary side. Sets of both types have the advantage that the two secondary circuits can be connected together to a single arc to give a high current for large electrodes or heavy work.

10.1.2.6 Power-factor correction

Because the load taken by a welding transformer is highly inductive, the power-factor will necessarily be

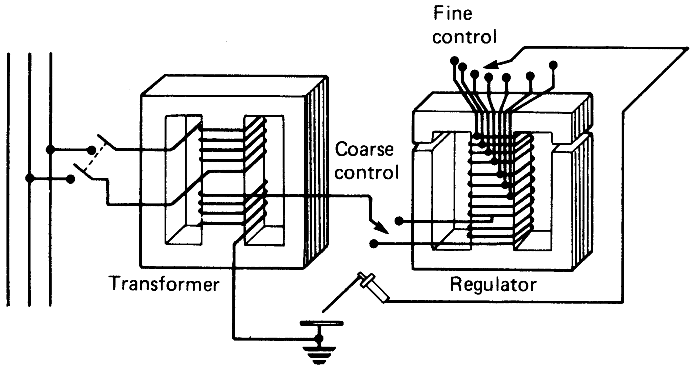


Figure 10.15 Single-phase transformer and tapped inductor

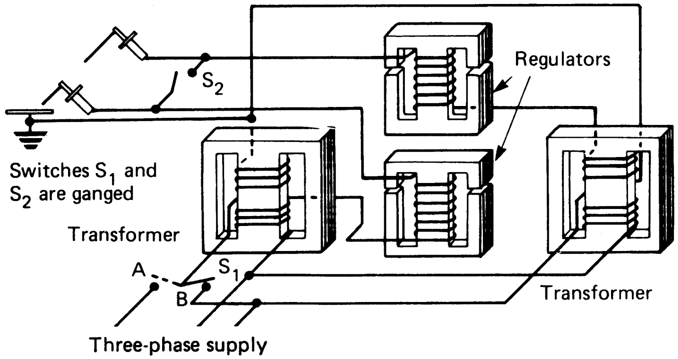


Figure 10.16 Double-operator welding equipment

low. For single-operator equipment it is of the order of 0.3–0.5 lagging, depending on the design of the set and the type of electrode used, arc length, etc. It is well known that the greater the inductance of the welding circuit, the better the conditions. A welding set which is not provided with a capacitor or other means of power-factor correction cannot; therefore, operate at a high power-factor.

The simple, single-phase transformer welding source lends itself readily to power-factor correction by the connection of a capacitor on the primary side. Power sources are available in which the capacitor is incorporated in or attached to the welding transformer.

Where many single-operator sets are connected, care must be taken not be overcorrect the power-factor. It must be borne in mind that a welding load is intermittent, and probably for 50% of the time for which the set is connected no welding will be in progress. If, therefore, each welding equipment is provided with a separate capacitor, there is the danger of a leading load being taken if the diversity factor of the welding load is small. This difficulty may be overcome by:

(1) automatic switching on of the capacitors when the load is applied;
(2) a central bank of capacitors to correct the load of the whole work based on an average observed load; and
(3) careful instructions to operatives to switch off capacitors as soon as they stop welding.

10.1.2.7 Multi-operator transformer equipment

Multi-operator transformer equipment consists of a single transformer to the secondary windings of which a number of arcs can be connected. Each welding circuit must be provided with a current regulator which can be a variable resistor or inductor. The latter gives higher efficiency and improved arc stability.

The most common form of installation is a three-phase transformer having a delta-connected primary winding and an interconnected-star secondary (*Figure 10.17*). It will be noted that the inductive current regulators are connected to the line terminals and the material to be welded is connected to the star or neutral which should be connected to earth.

The number of reactors and welders' circuits on the three-phase system must be a multiple of 3, so that they can be distributed evenly on the three secondary windings.

This type of plant is economical for installations where work is concentrated in one shop, and also for outside construction work where it is desirable to safeguard welders from possible shock from the primary voltage of the supply. This precaution is particularly necessary in the case of shipbuilding and bridge and storage-tank construction.

EN 60974-1:1998 IEC 60974-1:1998 indicates a method of connecting the welding equipment and specifies special plug and socket connections and distribution boxes. The leading dimensions of these are laid down, so components from different manufacturers are interchangeable.

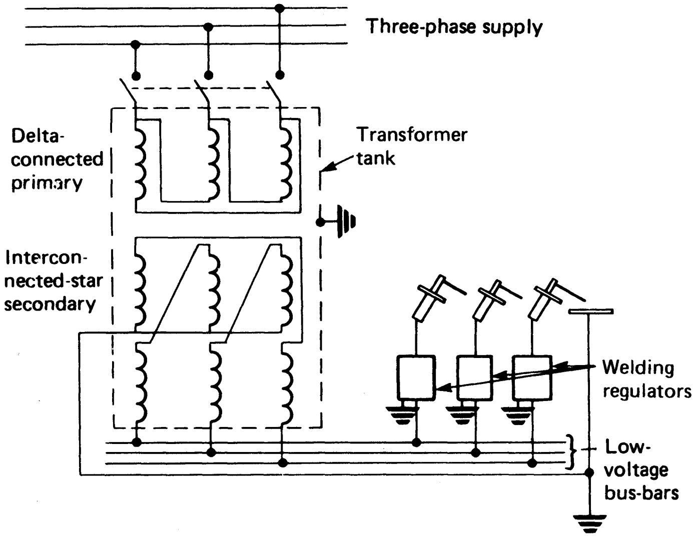


Figure 10.17 Three-phase multi-operator welding equipment

The sizes of transformers are limited to four (57, 95, 128 and 160 kV-A continuous rating) which are capable of providing one, two, three or four welding operators per phase, each at a maximum continuous hand-welding current of 350 A, or a lesser number at a higher current. The three standard sizes of current regulators are designed to give a maximum welding current of 350, 450 or 600 A, at 90 V.

10.1.2.8 Application

There are no special power source requirements for any of the particular group of electrodes (cellulosic, rutile and basic) although some power source manufacturers claim superior running characteristics with cellulosic electrodes. If recommended by the manufacturer, particular brands of electrodes can be operated with a.c. or d.c. power sources. Note, not all d.c. electrodes can be operated on a.c. but a.c. electrodes can usually be used on d.c. For a.c. operation, the no-load voltage is important and some electrodes, particularly the basic types, may require 70 to 80 V (a.c.).

A.c. electrodes are frequently operated with the simple, single-phase transformer whilst for d.c. electrodes, thyristor and transistor controlled power sources are used. The small size and weight of the inverter type power source offers the advantage of being easily transported and, therefore, ideal for moving around the workshop or site work.

The welding current level is determined by the size of the electrode and typical operating ranges for a range of electrode sizes are 50 to 400 A. As a rule of thumb, when choosing a suitable power source of adequate capacity, an electrode will require approximately 40 A per millimetre (diameter). As an example, for a commonly used electrode size of 4 mm, with an operating range of 140 to 180 A, the power source should have a current capacity of at least 200 A at an appropriate duty cycle i.e. 20%, 60% or 100%, depending on the type of work.

With small hobby type MMA transformers a duty cycle figure is not given. Instead, the rating is based on the number of electrodes that can be burnt in one hour, both when welding from cold and when the equipment has warmed up.

10.1.3 Metal inert gas welding

10.1.3.1 Principles of operation

The generic name 'metal inert gas (MIG) welding' has been given to those welding processes which use a continuously fed small-diameter, solid-wire electrode and a gas shroud. The principal operating features are shown schematically in *Figure 10.18*. The wire diameter is normally 0.8, 1.0, 1.2 or 1.6 mm and the shielding gas is usually CO_2 or argon with 2–5% O_2, or 5–25% CO_2. Argon–helium mixtures (with O_2 or CO_2 additions) can be used for special applications, e.g. for welding stainless steel.

MIG welding is slowly replacing MMA for manual welding, especially in the welding of thin section ferrous and non-ferrous materials, and when continuous operation or high deposition rates are required. Because of the continuous feeding of the wire electrode, MIG welding has found special application in mechanised and robotic welding. The process can also be used at high current levels, almost

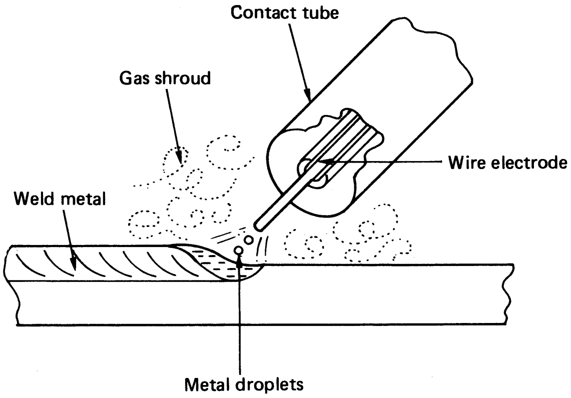


Figure 10.18 Principles of operation for the MIG process

twice that of MMA which can be exploited for welding thicker materials at higher speeds, but only in a mechanised operation.

Because the process is often used to increase the deposition rate/welding speed, care must be taken to avoid introducing weld defects such as lack of fusion. For this reason the process requires careful tuning of the power source, and welder training, when high quality is required.

10.1.3.2 D.c. metal transfer modes

There are three main metal transfer modes:

(1) short circuiting or dip;
(2) spray; and
(3) pulsed.

Short circuiting is a low-current mode of metal transfer, whilst the spray mode only occurs at high current levels. The pulsed mode is a technique used to obtain spray-type metal transfer at low current levels.

Short circuiting metal transfer The short-circuiting mode is used at low current levels, typically less than 250 A with a 1.2 mm diameter wire at which levels it is impossible to operate with an open arc to give free flight (spray) metal transfer. The transition from the short-circuiting mode to the spray-mode is shown schematically in *Figure 10.19.* The mode is established by setting a low open-circuit voltage so that the molten metal forming at the tip of the electrode wire transfers by short circuiting with the weld pool as shown in *Figure 10.20.*

If a free-flight technique, i.e. a spray-type mode of metal transfer, were to be attempted below the threshold current level, the low arc forces would be insufficient to prevent large globules forming at the tip of the wire which would then transfer erratically across the arc under normal gravitational forces. Thus, in operating the short circuiting transfer mode, the initial selection of the welding parameters (open-circuit voltage, inductance and wire feed speed) and the dynamic characteristics of the power source are crucial in stabilising the arc and metal transfer.

Spray metal transfer The spray metal transfer mode operates above the threshold current level (see *Figure 10.19*) and in this mode the droplets transfer to the workpiece in free flight. The voltage must be set at a higher current than for the short-circuiting mode in order to ensure that the tip of the wire does not bridge the arc gap which is typically set at 6 mm. The droplet diameter at current levels slightly above the threshold level approximates to the wire diameter.

The threshold current level is determined by the wire diameter and the shielding gas composition. Typical values for 1.6 mm diameter, low carbon steel wire in a range of argon–helium gas mixtures are shown in *Figure 10.21.* It can be seen that the threshold value increases significantly in high helium gas mixtures, indicating that the spray transfer is more difficult to obtain.

The droplets are detached by a combination of an electromagnetic effect, a plasma jet and the formation of gas bubbles in the molten droplet. The droplets have velocity when they are detached from the tip of wire but then accelerate across the arc gap.

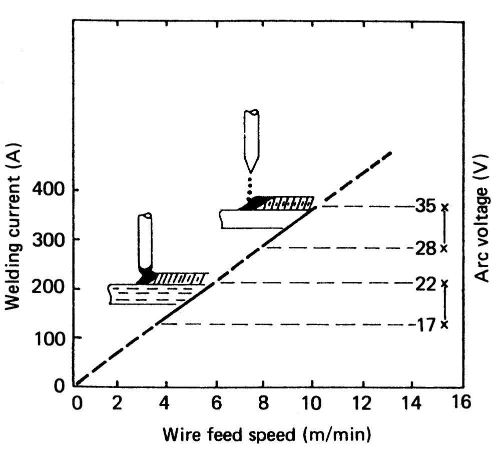


Figure 10.19 Normal operating ranges in MIG welding. Note that the voltage must be reduced at low current levels to induce short circuiting metal transfer

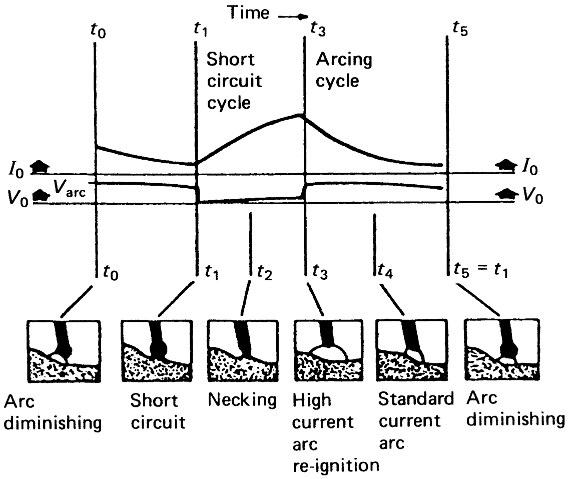


Figure 10.20 Short circuiting metal transfer in MIG welding. Note that a fluctuation in the welding current waveform occurs whenever the electrode short circuits to the workpiece

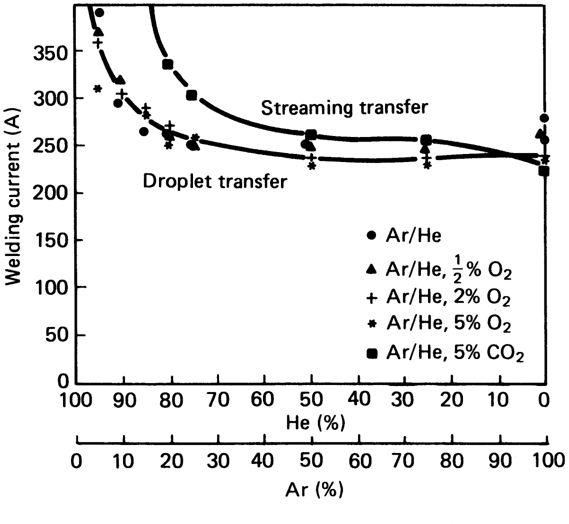


Figure 10.21 Threshold values for spray metal transfer

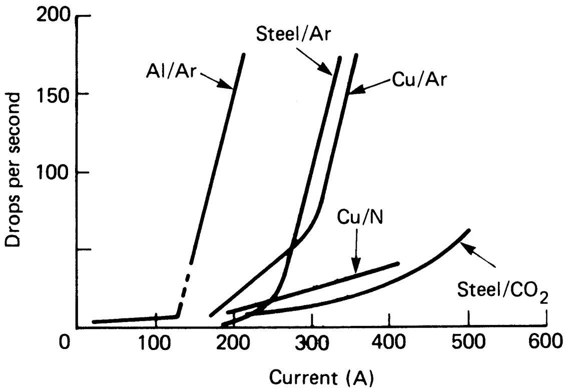


Figure 10.22 Droplet frequency for various material–gas compositions of 1.6 mm diameter wire

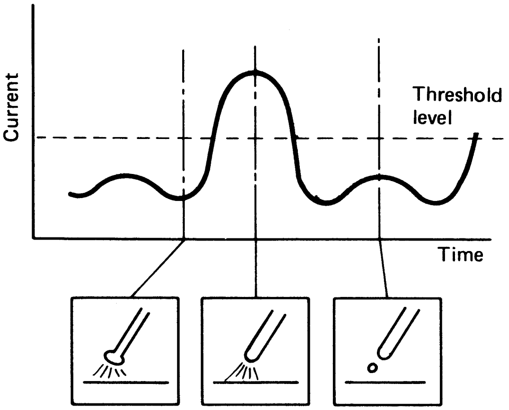


Figure 10.23 Pulsed metal transfer in MIG welding

The frequency at which the droplets transfer increases as the current level increases; the relationship between droplet frequency and the current for various material combinations is shown in *Figure 10.22.* The droplet size is about the same as the wire diameter at the threshold level, but decreases significantly as the welding current increases. At current levels well above the threshold level the droplets transfer in a stream of fine droplets and the wire tip assumes a pencil-point shape.

At very high current levels, e.g. above 450 A for 1.0 mm diameter, low carbon steel wire, the end of the wire which is softened by resistive heating can rotate. Rotation of the wire is promoted by extending the electrode stick-out, i.e. the distance the wires extend from the contact tip to the arc. The principal benefits of extended stick-out are increased deposition rate and an improved weld bead penetration profile (bowl shaped).

Conventional pulsed metal transfer The pulsed transfer mode is a free-flight technique, i.e. non-short-circuiting, which can be operated at welding currents below the threshold level. For this technique, high current pulses at a frequency of typically 50 Hz and 100 Hz are applied. The magnitude of the pulses must be sufficient to detach the droplets and project them across the arc. Droplet transfer to the weld pool is often completed during the background period; the process of metal transfer is shown schematically in *Figure 10.23* where it can be seen that, although significant melting occurs during the background period, the high current pulse is required for droplet detachment.

The pulsed MIG technique is an attractive alternative to the short-circuiting mode at low current levels because the open arc/free flight metal transfer greatly reduces spatter and provides greater tolerance to variations in the gun-to-workpiece distance. Until recently, the technique was not widely practised because of the difficulty in setting up and maintaining suitable welding parameters at a fixed pulse frequency. For example, as the mean current level demanded by the wire feed speed is set by adjusting the pulse and background current levels, the welding parameters are likewise a compromise for arc and metal transfer stability and weld bead penetration. The effect of the compromise can be seen in the metal transfer where the droplet size and frequency of transfer can vary depending on the number of pulses used to detach each droplet. Thus, power sources which can generate pulses over a range of frequencies are significantly more flexible, as the pulse frequency can be better matched to the wire feed speed.

Synergic pulsed MIG In the synergic process, the pulse parameters are varied according to the operating conditions. The basic concept is that a ‘unit’ pulse can be derived which, when applied, will detach one droplet of metal. A typical unit pulse is shown in *Figure 10.24*. There is range of unit pulses which will detach droplets of specific volume; to a first approximation the droplet volume increases proportionally with pulse duration.

Process stability can be readily achieved by relating the pulse-repeat frequency to the wire feed speed, i.e. the number of droplets (or increments of wire feed) to be detached from the wire. Under open-loop control; the pulse current and pulse time are set to detach one droplet of the desired volume and the background time and level are then adjusted to generate the required number of pulses to maintain the current burn-off rate. In synergic control, the unit-pulse principle has been used to produce a fully automated control system. The unit pulse (pulse current level and time) is uniquely determined by the material composition and the wire diameter. However, once the unit pulses have been determined for the desired range of materials, the power source is preprogrammed and the operator is merely required to select the appropriate programme for the wire being used. On setting the desired wire feed speed, the

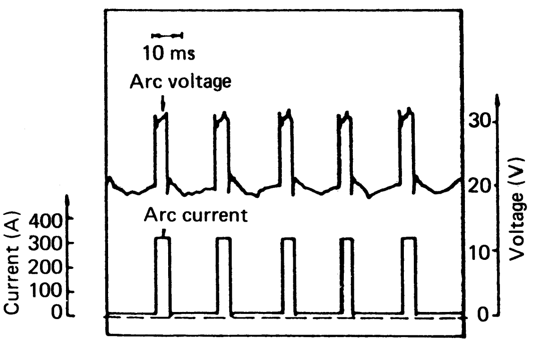


Figure 10.24 Typical oscillogram of arc current and voltage for 2 m/min wire feed speed, 50 Hz pulse frequency and 4 ms pulse duration

power source delivers the correct number of pulses to maintain arc stability.

10.1.3.3 D.c. power source

Short circuiting and spray metal transfer The basic power source for MIG welding has a nominally flat or constant voltage output characteristic as shown in *Figure 10.25*. The welding current is principally determined by the wire feed speed and the arc length by the power source voltage level (open-circuit voltage). Automatic adjustment of the wire burn-off rate is used to compensate for minor variations in the torch-to-workpiece distance, wire feed speed or current pick-up in the contact tip. For example, if the torch is moved away from the workpiece the arc length is momentarily increased, but this results in an increase in the voltage and a reduction in the level. Consequently, as the wire speed is constant, the reduction in the burn-off rate forces the arc length to return to the set arc length. If the torch is moved closer to the workpiece, the converse occurs, i.e. the burn-off rate is momentarily increased to maintain a constant arc length. Alternatively, a low inductance produces a rapid rise in current which assists arc starting.

In setting the power source for the two main operating modes (spray and short-circuiting metal transfer), the open-circuit voltage must be set to produce the required arc length for a given wire feed speed. For the short-circuiting mode, the inductance must also be selected by the operator to give a smooth rupturing of the metal bridge formed between the wire tip and the weld pool during the short circuit. Failure to set these parameters gives an erratic arc, as indicated by momentary arc outages and wire stubbing which results in excessive ejection of molten droplets from the weld pool (termed 'spatter') and a poor weld-bead profile. For the spray metal transfer mode, circuit inductance is not essential, but it helps to smooth the process, especially when operating with a short arc length.

The modern electronic, high response power sources have been designed to overcome some of the inherent limitations of operating the MIG process with more conventional power sources, namely the sensitivity to variations in gun-to-workpiece distance, wire-feed fluctuations, and long-term variations in the main supply voltage or through 'warm-up' drift in the control system. These limitations can make it difficult to sustain a high level of weld quality under low spatter conditions, particularly in mechanised or robotic applications. Transistor controlled power sources, in particular, can have sophisticated control systems such as:

(1) programmable parameter selection—the optimum welding parameters can be generated from the input of information on the wire being used (composition, diameter), shielding gas composition and the wire feed speed; or
(2) arc voltage as the setting up parameter—the operator is only required to set the desired voltage level and the wire feed speed is automatically adjusted by the system control.
(3) automatic setting of circuit inductance.

The trend in electronic power sources is towards 'one-knob' control with the wire feed speed level, which is under feedback control, being held within much closer limits.

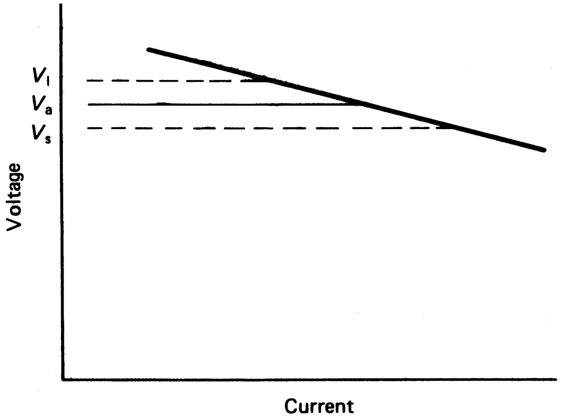


Figure 10.25 Nominally constant voltage output characteristic of the power source for MIG welding and self-adjustment of the arc length. V_a, set arc length; V_s, momentary shortening of the arc length; V_l, momentary lengthening of the arc length

Pulsed MIG In conventional pulsed MIG power sources, current pulses are generated through half-wave rectification of the mains supply to give sinusoidal pulses at a frequency of 50 Hz. The background current is usually supplied by a separate and independent supply. Full-wave and three-phase half-wave rectification are used for 100 and 50 Hz, respectively, whilst submultiples of these frequencies give sinusoidal pulses at repeat frequencies of 75, $37\frac{1}{2}$, $33\frac{1}{3}$ and 225 Hz.

In synergic control, high response, electronic (transistor or inverter) power sources are required in order to generate the 'unit' pulses; typical pulse durations are 2–5 ms and the pulse frequency lies within 25–500 Hz for current levels up to 300 A. The first type of synergic control modified the output of the power source (pulse frequency) by monitoring the wire feed speed. More recently, commercial power sources began to employ a range of synergic control systems to set the pulse frequency and the wire feed speed. The arc length is controlled by monitoring the arc voltage, and the wire feed speed or the pulse frequency may be adjusted to compensate for changes in the gun-to-workpiece distance.

Unlike conventional (constant current) MIG power sources, the first synergic power sources employed a constant-current characteristic with adaptive control of the arc length; the output of a tachnogenerator mounted on the wire feed was used to vary the pulse frequency to maintain arc stability. However, the more modern power sources use a combination of constant-current and constant-voltage characteristics. A common technique is to employ constant-current characteristics during the background period but constant-voltage characteristics during the pulse period for self-adjustment of the arc length. A limitation of this technique, however, is the very high peak current that can be produced. An alternative design is to use a constant-current supply with instantaneous voltage feedback acting on each pulse.

10.1.3.4 Control of short circuiting

In the conventional short circuiting process, metal transfers from the tip of the wire by periodically dipping into the weld pool (*Figure 10.20*). The process is controlled by setting the arc voltage to control the arc length and the circuit inductance to control the rate of rise of the current during the short circuit and the peak short circuit current.

The short circuiting process is characterised by spatter and low heat input to the material. The spatter originates from the explosion of the wire tip as it ruptures under the high short circuit current in a similar manner to the breaking of a fuse. The intermittent periods of arcing and arc outages during the short circuits increase the risk of cold laps.

As power sources have become more intelligent (see Section 10.1.1.3), control strategies can be used to improve process performance. For example, the Surface Tension Transfer (STT) process aims to control spatter and increase the heat input by improving the control of the current waveform and the sequence of events is as follows:

(1) At the onset of the short circuit, the current is reduced immediately and the molten metal on the tip of the wire is allowed to transfer under surface tension forces.
(2) A high 'pinch' current level is applied to speed up transfer of the molten metal.
(3) When the wire tip approaches 'necking', the current is quickly reduced to produce a smooth rupture of the wire.
(4) As the arc will form at a low current level, the risk of eject molten metal from the weld pool is greatly reduced.
(5) During the arcing period a pulse of current is applied to broaden the arc and reduce the risk of weld laps.

The advantages claimed for the STT process over conventional short circuiting metal transfer MIG are a reduction in parent metal dilution and lower labour costs through the higher travel speeds, reduction in repairs and less clean-up work.

10.1.3.5 A.c pulsed MIG

The a.c. arc has been recently stabilised for MIG welding by using a special power source with two inverters. The primary inverter controls the output current whilst the secondary inverter produces rapid switching between electrode positive and electrode negative polarities. A high voltage pulse is applied at current reversal to ensure that the arc is re-ignited. A pulse of current is also applied during the electrode positive half-cycle to ensure that the molten metal droplets are smoothly detached from the end of the wire. The pulse parameters (pulse current level and duration) are set to give one droplet per pulse.

Compared with conventional d.c. MIG, in the a.c. MIG process the wire burn-off rate is increased for a given current level which produces a 'colder' arc and a shallower weld pool penetration. These process features provide greater tolerance to joint gap as demonstrated for lap joints in thin sheet.

10.1.3.6 Twin wire MIG

As a means of increasing the deposition rate, two in-line wires, i.e. one behind each other, are fed simultaneously into the weld pool, *Figure 10.26*. The wires are powered by two separate power sources and as the contact tips are electrically isolated, they can be operated independently i.e. with different wire diameters, current levels or operating modes (continuous or pulsing). In practice, current pulsing is normally used but synchronisation of the power source outputs is necessary to minimise arc interaction. The current pulses are out of phase to avoid arc instability which would be caused by the interaction of the strong magnetic fields generated by the high current pulses. Metal transfers from one wire during the pulse period whilst the other wire has a low background current to avoid droplets being ejected from the arc. To control the process, the power source supplying current to the lead wire is designated as the lead (master). The current from trailing power source (slave) is then synchronised to the current supplied to the lead wire.

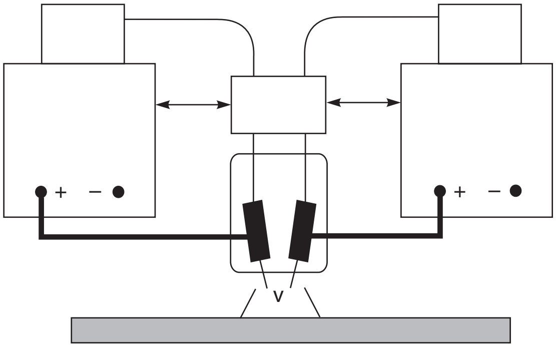


Figure 10.26 Electrical connections for twin wire with separate contact tips

The process requires a special torch but is sufficiently compact for welding the root in a V joint. In a fully automatic operation, very high welding speeds and high deposition rates can be achieved. Welding speeds are up to 3–4 times faster than those achieved using a single wire. For example, for 6 mm leg length fillet welds, the welding speed using the tandem wire process was 1.1 m/min compared with 300 mm/min for the single wire MIG process. Deposition rates can be as high as 20 Kg/hr which is about three times greater than is normally possible with a single wire.

Typical welding parameters for welding thin sheets components are:

Wire feed speed:	1st wire	12 m/min
	2nd wire	10 m/min
Pulse parameters:	pulse current	380 A
	pulse voltage	38 V
	background current	70 A
	frequency	280 Hz

The process uses the pulsed mode on each wire with a deposition rate of 10 kg/hr and a travel speed of 5 m/min.

10.1.4 Flux cored arc welding

10.1.4.1 Principles of operation

Flux cored arc (FCA) welding is an alternative to MIG and MMA welding. The wire consists of a hollow tube which is filled with flux and metal powder. The flux filling is similar in nature to the MMA covering, e.g. basic or rutile. The wire can be used without a separate gas shield (self-shielded welding), and the gas which is required to protect the weld pool is generated from the burning of the flux. In gas-shielded (FCA) welding, a gas shield of CO_2 or an argon–CO_2 mixture is used in the same manner as in solid-wire MIG welding.

10.1.4.2 Wire electrodes

As the first generation of wires were of large diameter (3.2 mm and above), the FCA process was initially used for welding at high deposition rates: the welding current levels were greater than 400 A, which gave deposition rates in excess of 24 kg/h. The availability of smaller diameter wires gives a wide range of deposition rates and the capacity to

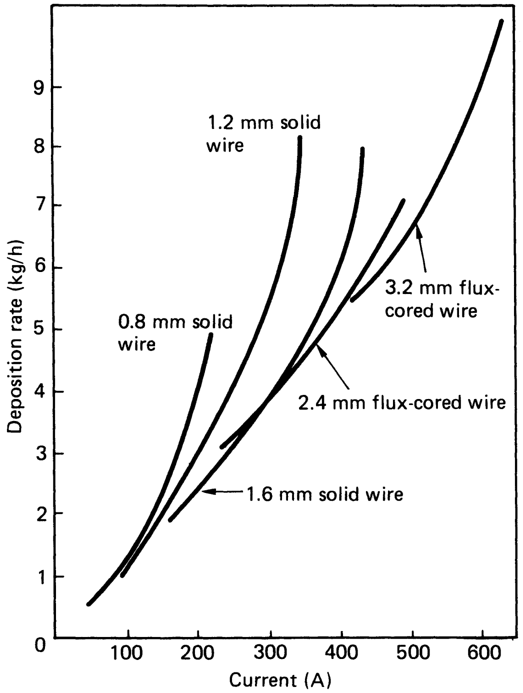


Figure 10.27 Deposition rates for various steel wires and welding processes

weld in the vertical and overhead positions; a comparison of deposition rates for the MIG solid wire, and is given in *Figure 10.27*. However, the past disadvantages of the quality of the wire (e.g. consistency of the core filling) and the cost compared with solid wire, have restricted widespread application of FCA welding. Advances in flux formulations and manufacturing technology to produce an attractive range of smaller diameter wires (down to 0.9 mm) with all-position welding capability and good mechanical properties will make the FCA process more competitive than the alternative MIG and MMA processes.

Cored wires are normally classified according to the American Welding Society Standard AWS-A5.20 which defines the type of wire in terms of the core composition, the performance and the mechanical properties of the weld metal. The more commonly used wires are listed below.

Type T1 These wires are usually rutile based with a CO_2 or argon–CO_2 shielding gas. As the wires have a spray transfer mode with a fast freezing slag, they can be used in all welding positions.

Type T5 These wires have a basic flux core and use a CO_2 or argon–CO_2 gas shield. The transfer mode is globular which produces a convex bead shape, generates more spatter and is not suited to positional welding operations.

Type T7 These wires are self-shielded and have a globular type metal transfer. The slag is designed for high current/high deposition rates. However, the resulting welds have good mechanical (impact) properties.

Type T8 These are self-shielded wires which can be used in all welding positions and with good low-temperature impact properties. Small diameter wires, typically 1.2 mm, have all-position welding capability.

Recent developments in cored wires include barium free E71T-8 wires and E7OT-1 all-position rutile wires.

10.1.4.3 Power source

The power source designed for normal MIG welding is suitable for operating with cored wires. However, as the wire diameter can be significantly larger, and because higher deposition rates are possible in the flat position, the capacity of the power source is greater than is normally specified for MIG welding. For example, a 2.4 mm diameter wire can be used at current levels up to 500 A and a 3.2 mm diameter wire can require a power source with over 600 A capacity.

The wires are normally operated with d.c., but both electrode-positive and electrode-negative polarities are used. The usual practice is to use electrode-positive for a rutile wire and electrode-negative for basic and metal cored wires. As both the short circuiting and spray transfer modes are used, the power sources require voltage range and inductance settings to control the process. The pulsed mode of transfer is not normally used, but claims have been made that the technique can improve metal transfer, especially for larger diameter wires at low current levels.

10.1.5 Submerged arc welding

10.1.5.1 Principles of operation

The submerged arc process is a high current, bare wire electrode process in which the arc operates below a bed of powdered flux. The principles of the process are shown in *Figure 10.28*. The flux has a similar function as in MMA and FCA welding in generating gases to protect the arc and the weld pool from the atmosphere, and providing alloying elements to achieve the desired mechanical properties in the weld metal. However, unlike the MMA and FCA processes, excess flux is laid down which must be recycled via a hopper. The slag, as is formed in MMA, must be removed after welding and discarded.

The electrode is normally wire (1.6–6.0 mm diameter), but for cladding large surface areas a strip electrode of up to 130 mm in width, 0.5 mm thickness can be used. The normal welding current range is 200–2000 A and the corresponding arc voltage level is 25–45 V.

The weld metal recovery rate is almost 100%, as very little metal is lost through spatter. The flux consumed in forming the slag, is approximately equal to the weight of wire deposited. As heat losses from the arc are exceptionally low, due to the insulating effect of the flux bed, the thermal

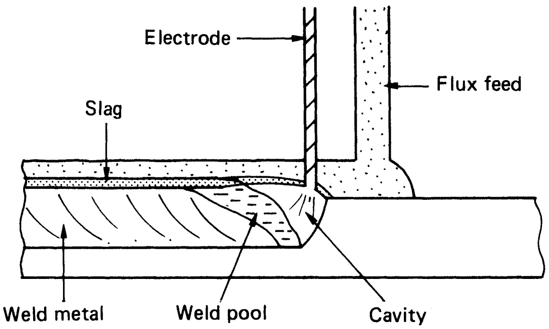


Figure 10.28 The principles of the submerged arc process

efficiency of the process can be as high as 60%, compared with about 25% for MMA welding.

10.1.5.2 Power source

The power sources can be a.c. or d.c. but, because the process is used for large diameter wire or strip electrodes giving high deposition rates, the capacity can be high as 2000 A.

Single wire For a.c. operation, the power source has a constant-current characteristic (drooping *V*/*I* output) or the open-circuit voltage should be at least 80 V to ensure reliable arc reignition on reversed polarity. The d.c. power sources have either constant-current (drooping output) or constant-voltage (flat) characteristics. With the constant-current power source, the arc length must be set by balancing the wire feed for the given current setting, i.e. to match the burn-off rate. The arc length can be adjusted automatically by using the arc voltage to vary the wire feed speed. In the constant-voltage power source, the normal self-adjustment of the arc length is obtained using the arc voltage as the control parameter for a given wire feed speed/welding current setting.

The constant-voltage power source is normally used for thin sheet welding, but the constant-current power source is preferred for welding thicker plates. The constant-current arc is more stable for deep penetration applications where the self-adjustment effect of the constant-voltage power source can lead to arc instability.

10.1.5.3 Electrode polarity

The electrode polarity can be either d.c. (electrode-positive or negative) or a.c. D.c. electrode-positive gives the deepest penetration, d.c. electrode-negative has the greater deposition rate, and a.c. has an intermediate characteristic. The effect of electrode polarity on deposition rate is shown in *Figure 10.29*. The d.c. electrode-positive polarity will produce approximately 20–25% increase in penetration compared with d.c. electrode-negative. For this reason, d.c. electrode-positive polarity is normally used for the root run in welds to ensure that adequate penetration is achieved. However, for surfacing applications where low penetration and parent metal dilution is required, d.c. electrode-negative is preferred.

The wire burn-off rate can be increased significantly by increasing the electrode extension (the distance between the contact tip and the arc). The effect is to preheat the wire according to the relationship

$$\text{Heat generated} = \Leftarrow \frac{I^2 LR}{d}$$

where I is the current, L is the electrode stick-out, R is the resistivity of the wire, and d is the diameter of the wire.

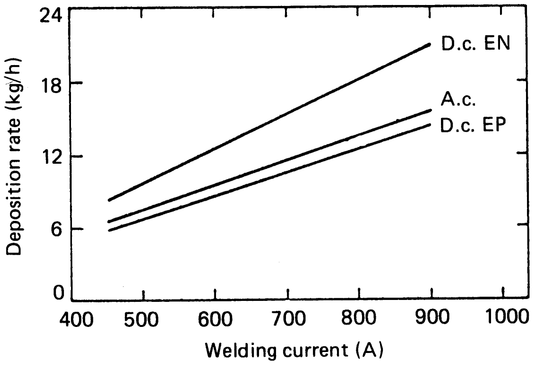


Figure 10.29 Effect of electrode polarity on deposition rate (standard electrode extension 32 mm; 4 mm diameter wire; 32 V). EP, electrode-positive; EN, electrode-negative

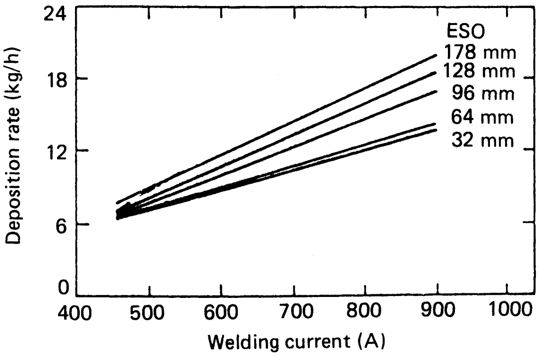


Figure 10.30 Effect of stick-out on deposition rate (d.c. electrode-positive; 4.75 mm diameter wire). ESO, extended stick-out

Preheating of the wire can be used to increase significantly the deposition rate (see *Figure 10.30*). As the wire will soften through resistance heating, the stick-out must be limited or supported to avoid arc wander under the high electromagnetic forces. It is usual practice to support the wire with an insulated guide tube to a distance of 25 mm above the workpiece surface. Even with a guide tube, it is normally recommended that the maximum stick-out be limited to, typically, 76 mm for 3.2 mm diameter wire, but this can be increased to, typically, 130 mm for a 4 mm diameter wire.

10.1.5.4 Hot wire welding

The resistance heating of a wire can be used to increase the weld metal deposition rate by feeding a separate (I^2R heated) wire into the weld pool. The wire is heated using a low voltage of typically 8–15 V from a separate a.c. power source. The wire, which is typically 1.6 mm in diameter, enters the weld pool at a temperature just below its melting point and no arcing occurs. Deposition rates are in excess of those achieved with single-wire d.c. electrode-negative polarity, even with an extended electrode stick-out.

10.1.5.5 Series arc welding—single power source

As series arc welding is used for high deposition rate welding applications, specialised techniques have been designed specifically to increase the deposition rate and the welding speed. As the maximum welding current on a single wire is limited by a deterioration in weld quality through excessive arc forces, the techniques are based on the use of multi-arcs.

The simplest arrangement is two wires connected to the same power source to give d.c. electrode-positive and d.c. electrode-negative arcs or, alternatively, two a.c. arcs (*Figure 10.31*). When used in tandem (d.c. electrode-positive leading), a substantial increase in welding speed, typically 1.5 times the single-wire process, can be achieved with no deterioration in the weld-bead shape.

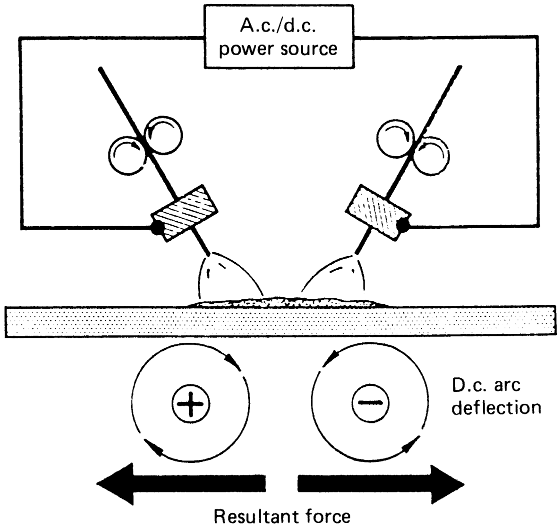


Figure 10.31 Series arc welding

The two arcs operate into a single weld pool and, because of this close proximity, there is significant arc interaction. With d.c., the arcs will diverge, i.e. be repelled by each other, whilst two a.c. arcs will be largely unaffected.

10.1.5.6 Series arc—multipower

Multipower systems use separate wire feed systems and power sources. The increased flexibility of electrode polarity and number of arcs permits a substantial increase in the deposition rate and welding speed and an improvement in the weld-bead shape. Of the possible combinations, multiple d.c. arcs are normally avoided to reduce the risk of arc blow which is caused by the large magnetic fields associated with the high current levels. High current, multipower systems are, therefore, normally a combination of d.c. and a.c. arcs.

The most commonly used arrangement is a twin-wire system with a d.c. electrode-positive leading arc and an a.c. trailing arc (*Figure 10.32*). When operated at high current and low voltage, the d.c. electrode-positive arc will give deep penetration, whilst the lower power (lower current/higher voltage) a.c. arc will provide joint filling with minimum weld pool disturbance and a smooth final weld bead surface profile. The normal spacing between the arcs is 30–50 mm. The power sources are normally a d.c. thyristor controlled rectifier and a single-phase transformer for the a.c. arc.

For three-wire systems, the leading arc is d.c. electrode-positive, but the middle and trailing arcs are normally a.c. as shown schematically in *Figure 10.33*. The phasing of the two a.c. arc voltages is important to minimise arc interaction. In order to produce a phase difference, separate a.c. power sources are preferable, but a Scott-type transformer will give a 90°⇐difference in phase shift between the two outputs. As the deposition rates can be substantially increased with multi-wire welding processes, the techniques are used in industries such as shipbuilding, where large flat plates must be welded. Typical welding speeds for various plate thicknesses are given in *Figure 10.34*, where it can be seen that the joint completion rate for a 20 mm thick plate

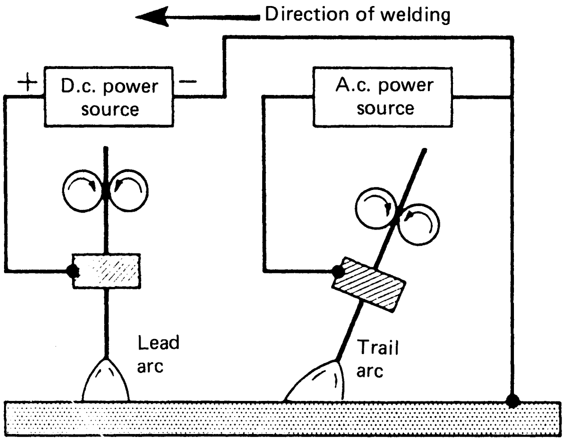


Figure 10.32 Two-wire welding

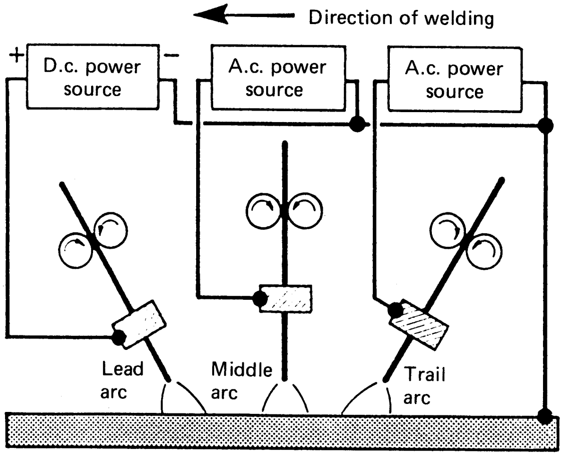


Figure 10.33 Three-wire multiple-electrode welding

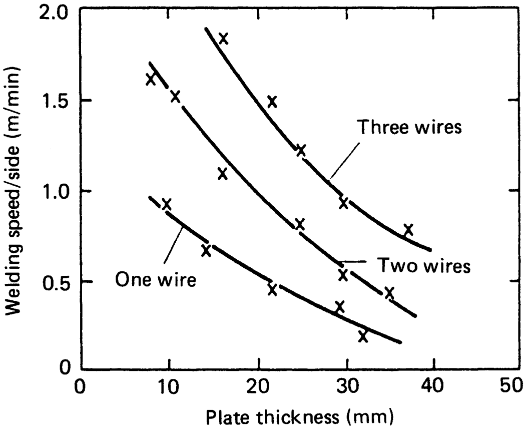


Figure 10.34 The effect of the number of electrodes on joint completion rate (speed/side)

is doubled for a twin-wire system and trebled for the three-wire system.

10.1.6 Tungsten inert gas welding

10.1.6.1 Principles of operation

In the tungsten inert gas (TIG) welding process, the arc is formed between a pointed tungsten electrode and the workpiece in an atmosphere of argon or helium (see *Figure 10.35*). The small intense arc provided by the pointed electrode is ideally suited to precision and controlled melting of the workpiece. Since the electrode is not consumed during welding as with the MMA and MIG processes, autogenous welding can be practised without the continual compromise between the heat input from the arc and the deposition of filler metal. When filler metal is required, this must be fed separately into the weld pool using a wire or rod feed system.

The process can be operated with d.c or a.c. In d.c welding the electrode polarity is always negative so that its electron thermionic emission properties reduce the risk of overheating which would occur with electrode-positive polarity. The arc heat is distributed approximately one-third into the electrode and two-thirds into the workpiece. However, the alternative of d.c. electrode-positive would have advantages in that the cathodic action on the workpiece surface would clean the surface of the oxide coatings. For this reason, a.c. is the preferred polarity for welding materials such as aluminium which have a tenacious oxide film.

10.1.6.2 Electrode

The electrode for d.c. welding is pure tungsten with 1, 2 or 4% thoria, the thoria being added to improve electron emission which facilitates arc ignition. Alternative additives to lower the electron work function are lanthanum oxide or cerium oxide, which it is claimed have improved starting characteristics, lower electrode consumption and, compared with thoria, are non-radioactive. When using thoriated electrodes it is recommended that precautions are taken to avoid contact with the grinding dust and smoke.

The electrode tip angle for the d.c. arc must be tapered to a fine point to concentrate and stabilise the arc. The general rule is that the lower the welding current, the smaller the electrode diameter and the tip angle. Recommended electrode diameters and tip angles for d.c. and a.c. arcs are given in *Table 10.2*.

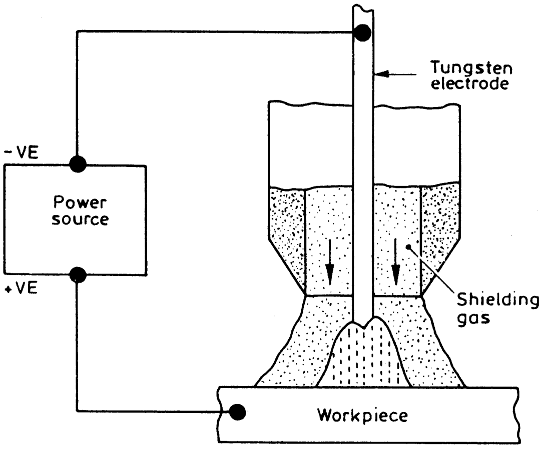


Figure 10.35 Principles and features of the plasma arc process

In a.c. welding, the electrode must operate at a much higher temperature due to the heat generated during the positive half-cycle. As the rate of tungsten loss is somewhat less than with thoriated electrodes, pure tungsten or tungsten with zirconia electrode is preferred. Furthermore, because of the greater heating of the electrode, it is difficult to maintain a pointed tip, and the end of the electrode assumes a spherical or 'balled' profile.

10.1.6.3 Shielding gas

The shielding gas must be inert or slightly reducing and the composition is normally selected according to the material being used.

1. *Argon*—This is the most common gas employed because of its low cost compared with the other inert gases, and it can be used for most materials.
2. *Argon–H_2*—Hydrogen can be added (up to 5%) to produce a slightly reducing atmosphere. The gas is hotter and, being slightly more constricted, produces deeper penetration and higher welding speeds.
3. *Helium and helium/argon*—A typical gas mixture is 75/25, helium/argon mixture which produces a higher heat input and higher welding speeds; the greater heat input is produced by the higher ionisation potential which is 25 eV for helium compared with 16 eV for argon. Besides the higher cost (approximately three times the cost of argon), the arc may be more difficult to ignite, especially in pure helium.
4. *Nitrogen*—As nitrogen is diatomic, on reassociation at the workpiece surface it is capable of transferring more energy than the monatomic argon or helium. The gas is used to weld high-conductivity metals, e.g. copper, but because of nitrogen pick-up in the weld pool, it cannot be used with steels due to the reduction in the toughness.

10.1.6.4 Power source

The power source necessary to maintain the TIG arc has a drooping voltage-current characteristic which provides an essentially constant-current output even when the arc length is varied over several millimetres. Hence, the natural variations in arc length which occur in manual welding have

Table 10.2 Recommended electrode diameter and vertex angle for TIG welding at various current levels

	D.c. electrode-negative		*A.c.*
Welding current (A)	*Electrode diameter** (mm)	*Vertex angle* (°)	*Electrode diameter†* (mm)
<20	1.0	30	1.0–1.6
20–100	1.6	30–60	1.6–2.4
100–200	2.4	60–90	2.4–4.0
200–300‡⇐	3.2	90–120	4.0–4.8
300–400‡⇐	3.2	120	4.8–6.4

* Thoriated tungsten.

† Zirconiated tungsten, balled tip, electrode diameter depends on degree of balance on a.c. waveform.

‡ Use current slope-in to minimise thermal shock which may cause splitting of the electrode.

little effect on the welding current. The capacity to limit the current to the set value is equally crucial when the electrode is inadvertently short circuited onto the workpiece. Otherwise, excessively high currents would be drawn, damaging the electrode and even fusing the electrode to the workpiece.

In practice, the power source is required to reduce the high-voltage mains supply (240 or 440 V and a.c.) to a relatively low voltage (60–80 V, a.c. or d.c.) supply. In its basic form, the power source comprises a transformer to reduce the mains voltage and to increase the current and the rectifier, placed on the secondary side of the transformer, to provide the d.c. supply. Traditional power-source designs employ a variable reactor, moving coil or moving iron transformers, or a magnetic amplifier to control the welding current. Such equipment has the highly desirable features of simple operation and robustness, making it ideally suited to application in aggressive industrial environments. The disadvantages are relatively high material costs, large size, limited accuracy and slow response. Electronic power sources (described earlier) have become available which do not suffer from these disadvantages:

(1) thyristor (SCR) phase control;
(2) transistor, series regulator;
(3) transistor, switched; and
(4) a.c. line rectifier plus inverter.

The major operating features of these systems are described in Section 10.1.1 and the advantages/disadvantages compared with conventional power sources are given in *Table 10.1*. Of the above power sources, the transistor based control systems offer greater accuracy and reproducibility of welding parameters, but tend to be wasteful of electrical energy. The a.c. line rectifier, plus inverter type offers a combination of high electrical efficiency and small size.

Because of the constant-current output characteristics, the arc can be ignited by either touching the electrode to the workpiece or in a contact system by a series of high-frequency high/voltage sparks. The effect of the high-frequency is to ionise the gas between the electrode and the workpiece. As the voltage and frequency are approximately 10–20 kV at 100 MHz, care must be taken to prevent insulation breakdown of the welding control system. Line and air-borne high frequencies can cause problems in instrumentation and electrical equipment in the vicinity of the arc and the power lines of the welding system. High-frequency feedback to the power source can be eliminated by placing an air-cored inductor between the high-frequency generator and the transformer rectifier; the isolator may be built into the high-frequency transformer, as shown in *Figure 10.36*. Care must be taken to ensure that all equipment is properly earthed and that all welding leads are kept as short as possible.

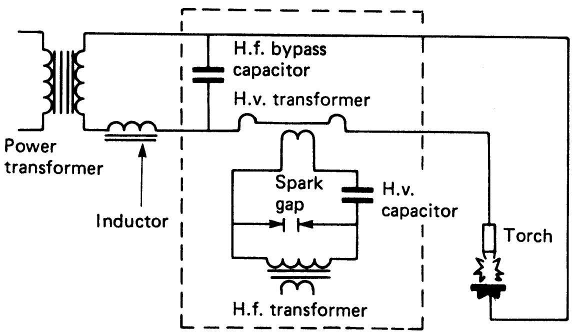
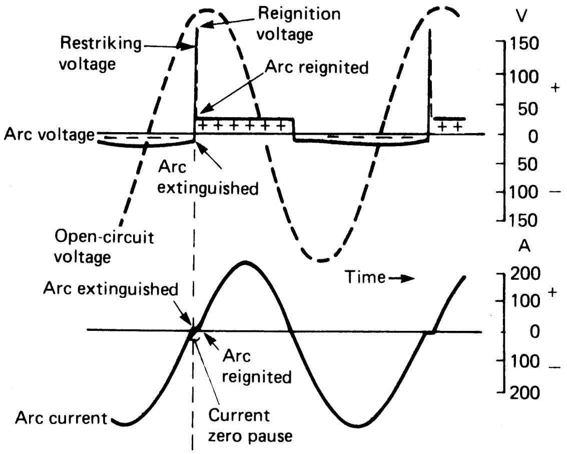


Figure 10.36 High-frequency arc-starting unit for TIG welding. H.f., high frequency; h.v., high voltage

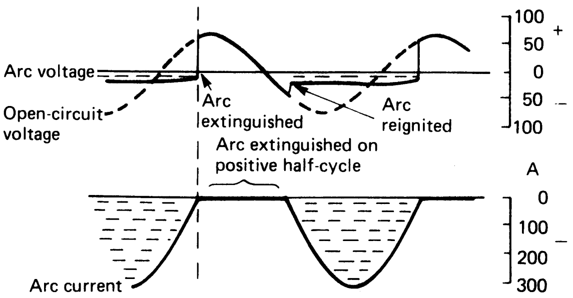
Sine wave a.c. The cyclical nature of the current introduces certain difficulties. When the tungsten electrode changes from positive to negative polarity, a smooth transition takes place because the tungsten electrode (being a thermionic emitter) has an electron cloud available for re-ignition as an arc cathode. When the change in electrode polarity is from negative to positive, a cathode root or a group of multiple cathode roots have to form on the plate. This function requires a high restriking voltage to re-ignite the arc, which is over 150 V when welding aluminium.

For the usual inductive supply, the arc voltage and current waveforms (*Figure 10.37*) both lag considerably on the open-circuit voltage. As a result, a high restriking voltage is available (*Figure 10.37*(*a*)). If the arc fails to re-ignite because of insufficient restrike voltage, a rectifying arc can occur, with the current flowing predominantly in the negative half-periods. Under low-voltage conditions, it is possible to secure positive half-period current by use of auxiliary equipment to give, for example, spark re-ignition. The sparks must be properly timed, as otherwise some degree of rectification will occur.

A more precise method of obtaining the electrode-positive half cycle is the use of the surge injection technique. With a surge injector added to a welding transformer, the open-circuit voltage can be reduced to 50 V. The basic circuit of the



(a) Self reignition



(b) Complete rectification

Figure 10.37 Waveforms of voltage and current for TIG a.c. welding

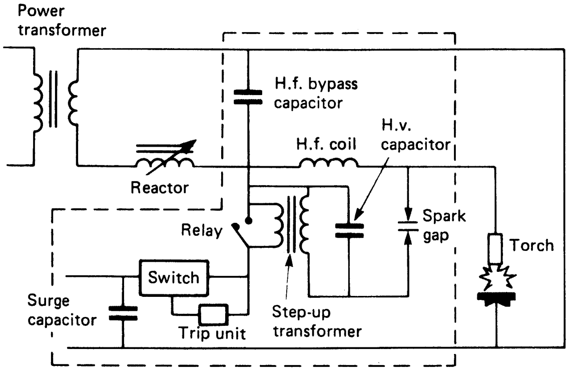


Figure 10.38 Surge injector unit and welding circuit. H.f., high frequency; h.v., high voltage

surge injector together with the high-frequency arc starter is shown in relationship to the welding circuit in *Figure 10.38*.

The operation of the circuit for starting is as follows. When full open-circuit voltage is applied to the system, the relay contact is opened, and the trip unit operates the switch to discharge the surge capacitor into the primary of the step-up transformer. The voltage induced in the secondary builds up until the breakdown voltage of the spark gap into the torch is reached. When the arc has been established, the voltage applied to the relay falls to the arc voltage level and the relay contact closes, the surge capacitor then being discharged directly into the arc. The instant of discharge is governed by the trip unit and is so timed as to occur at arc extinction when the polarity is changing to the electrode-positive half-cycle. The surge capacitor, which is charged to a voltage of sufficient amplitude, is then used to provide an artificial restrike voltage.

Square wave a.c. An alternative design of power source which is becoming more popular is the square wave power source. The principal feature of such designs is that the output current assumes a more square waveform, compared with the conventional sine wave (*Figure 10.10*). Two types of power source are available, differing in the manner in which the square waveform is produced. Whilst a 'squared' sine wave-form is generated by using inverted a.c., a more truly square waveform is produced by a switched d.c. supply (see *Figure 10.11*). In either case the importance for TIG welding is that the current is held at a relatively high level prior to zero and then transfers rapidly to the opposite polarity. In comparison, the current developed by sine wave power sources decreases more slowly to current zero, and likewise the current built up after re-ignition is at a much lower rate.

As shown in *Figure 10.39*(*a*), if a square wave a.c. derived from a switched d.c. supply is used at 75 V open circuit and 160 A r.m.s. welding current, a voltage of 50 V and a circuit current of some 160 A are both obtained within 0.02 ms from zero. With a squared sine wave, a voltage across the gap of above 50 V is achieved in 0.02 ms and a circuit current of 110 A is attained in 0.1 ms from zero (*Figure 10.39*(*b*)). In comparison, the equivalent rise time for a conventional sine wave supply is 0.15 ms to achieve 5 V across the arc gap, and a relatively long time of approximately 3 ms to achieve 110 A from zero.

The benefit of square wave a.c. is that, aided by the inherent high surge voltage associated with the rapid current reversal, a.c. TIG can in some instances be practised at 75 V r.m.s. without the need for high-frequency spark injection to be superimposed for arc re-ignition.

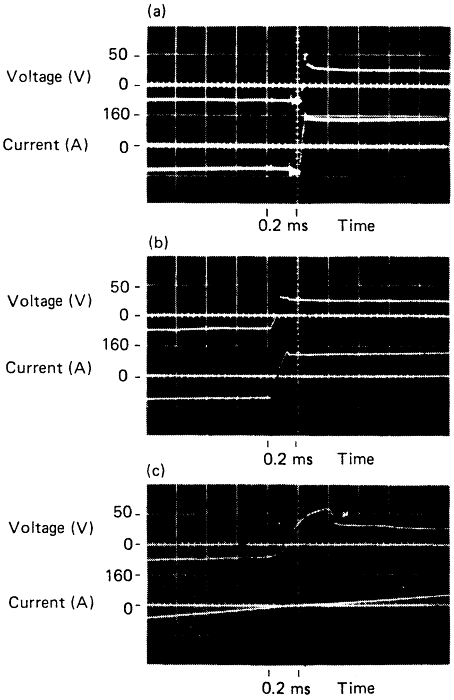


Figure 10.39 Typical positive re-ignition waveforms of voltage and current when welding at 160 A r.m.s. (a) Square wave supply at 75 V open-circuit voltage. (b) Square sine wave at 79 V open-circuit voltage. (c) Sine wave supply at 75 V open-circuit voltage

An additional feature of square wave a.c. power sources is the capacity to imbalance the current waveform, i.e. to vary the proportion of electrode-positive to electrode-negative polarity. In practice, the percentage of electrode-positive polarity can be varied from 30 to 70% at a fixed repeat frequency of 50 Hz. By operating with a greater proportion of electrode-negative, heating of the electrode can be substantially reduced compared to that experienced with a balanced waveform. Although cleaning of the oxide on the surface of the material is normally sufficient with 30% electrode-positive, the degree of arc cleaning may be increased by operating with a higher proportion of electrode-positive polarity (up to a limit of approximately 70%).

10.1.6.5 TIG hot wire

The TIG hot wire variant was developed as a means of achieving very high deposition rates without reducing the high weld quality normally associated with TIG welding. The essential feature is that filler wire is fed directly into the back of the weld pool and resistance heated using a separate power source (see *Figure 10.40*(*a*)). This second power source is usually a.c., to minimise any interference with the welding arc through the magnetic field generated by the current flowing in the wire. In a low current manual system, a single power source can be used to provide both the arc and resistance heating currents, but in this case the hot wire current is d.c. (*Figure 10.40*(*b*)).

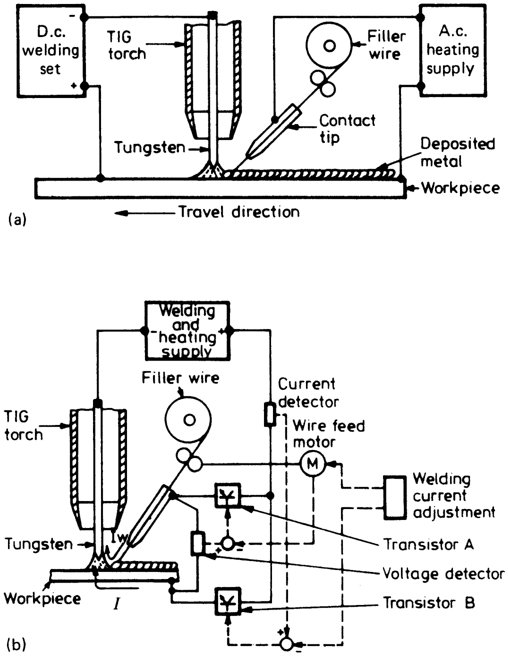


Figure 10.40 Electrical system and torch arrangement for the TIG-hot wire process. (a) Separate hot wire a.c. power source; (b) Single d.c. power source for arc and wire currents: I, arc current; I_w, wire heating current; $I + I_w$, welding current

In operation, the arc melts the base metal to form the weld pool. The filler wire, heated to its melting point by its own power source, enters the weld pool behind the arc to form the weld bead (*Figure 10.41*). Smooth feeding of the wire, control of the angle of entry into the weld pool, and a stable power source are all essential for stable operation, otherwise random arcing from the filler wire will occur with the resulting pool disturbances, causing porosity.

The main advantage of the process is that deposition rates can be achieved which approach those obtainable with MIG welding, i.e. 5–20 kg/h. Hence, TIG hot wire is used for welding, thicker section material, where the significantly higher depletion rates compared with those of the TIG cold wire process, can be fully exploited without any reduction in weld quality. In a specific example—welding steel tube of approximately 20 mm wall thickness-the deposition rate could be increased by a factor of 4. As the number of passes was reduced from approximately nine to five, the overall joint completion rate was reduced by a factor of almost 3.

TIG hot wire is also used for high deposition rate cladding, and here deposition rates of 10 kg/h are readily obtained. Even higher deposition rates of the order of 14 kg/h can be achieved, but only through the use of oscillation of the torch.

10.1.6.6 Pulsed current

Pulsing the welding current at a frequency of 0.1–10 Hz is used to improve control over the arc stability and weld pool behaviour. The essential feature is that a high current pulse is applied, causing rapid penetration of the workpiece.

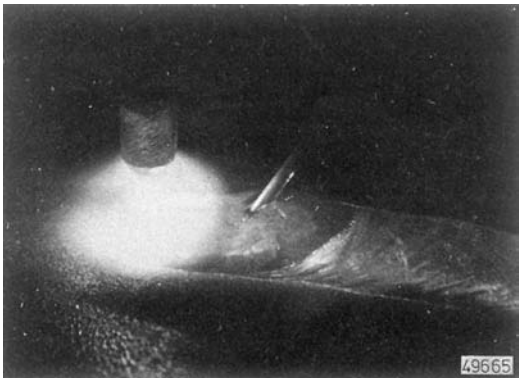


Figure 10.41 Photograph showing the TIG arc and the (hot) wire entering the weld pool behind the arc

If this high current is maintained, excessive penetration and ultimately burn-through occur. Therefore, the pulse is terminated after a pre-set time and the weld pool is allowed to solidify under a low background or pilot arc. Thus, the weld progresses in a series of discrete steps with the pulse frequency balanced to the welding speed to give approximately 60% overlap of the weld spots. The surface appearance of a typical pulsed current TIG weld is shown in *Figure 10.42*.

The pulsed technique has been found to be particularly beneficial in controlling the penetration of the weld bead, even with extreme variation in heat sink. Such variations are experienced either through component design, e.g. thick to thin sections, or from normal production variations in component dimensions, fit-up, clamping and heat build up. In conventional continuous-current welding, where a balance must always be achieved between the heat input from the arc, the melting to form the weld pool, and the heat sink represented by the material of the component being welded, the degree of penetration is greatly influenced by these variations. However, in pulsed operation, the rapidly penetrating weld pool during the high current pulse and the solidification of the weld pool between pulses, markedly reduce the sensitivity to process variation through the effects of heat build up and/or disparity in heat sink.

Despite the obvious advantages of the pulsed process to production, the technique may appear to be a further complication of the process in that a greater number of welding parameters must be considered, i.e.

(1) pulse time;
(2) pulse level;
(3) background time; and
(4) background level.

The technique can be simplified in the first instance from the knowledge that, for a given material, there is a preferred pulse level which is based on its diffusivity and, to a lesser extent, on its thickness. The preferred currents are approximately 400 A for copper, 150 A for stainless steel, and 50 A for lead. Thus, for a given component, the operator need only set the pulse time to achieve penetration which, as shown in the nomograph in (*Figure 10.43*(*a*)), is determined solely by thickness. For example, for welding 2 mm stainless steel at 150 A, a 0.5 s pulse would be demanded, whilst for a 1 mm thick material, the pulse time would be reduced to 0.1 s at the same current level. The background parameters are considerably less critical in the pulsing operation. The background level is normally set at approximately 15 A

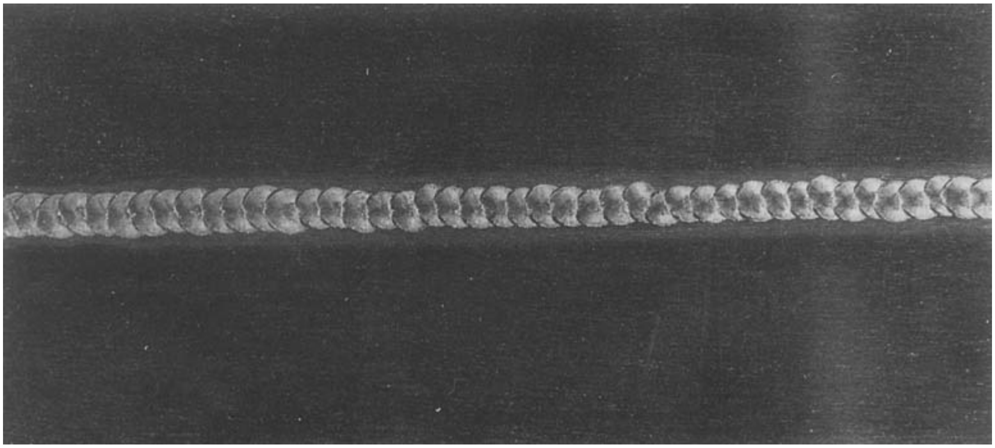


Figure 10.42 Photograph of a pulsed TIG weld

which provides the greatest possible heat dissipation during this period, whilst being high enough to maintain a stable arc. The background period is normally equal to the pulse period, but may be some two to three times greater when welding thicker sections.

The nomograph is presented only as a guideline for the initial selection of welding parameters and must be treated with caution, particularly when welding at the extremes of the thickness range, i.e. sections of > 0.3 mm and < 1.0 mm. In both instances, the preferred pulse current level will be outside the above theoretical operating ranges. For example, in the welding of stainless steel, practical trials have established that for a thickness of 4 mm the preferred pulse parameters are 200 A and 0.75 s, whilst for 0.5 mm thick material the preferred pulse parameters are 50 A and 0.1 s (*Figure 10.43*(*b*)).

Welding thick sections at too low a pulsed current can result in the loss of most of the advantages of pulsing, controlled depth of penetration and tolerance to variation in heat sink, as the weld pool takes a long time to penetrate the material and thermal diffusion occurs ahead of the fusion front. In welding thinner sections with too high a pulsed current, the excessive arc forces may cause cutting and splashing of the weld pool, resulting in a poor bead profile and electrode contamination.

The capacity to use lower pulsed currents and longer pulse time (see *Figure 10.43*(*b*)) is also of particular importance when using power sources which have a limited response, a low rate of rise and fall of the current between the background and the pulsed current levels. For instance, power sources in which the current is controlled by a magnetic amplifier are generally limited to pulses of 0.2 s duration, whilst in thyristor controlled types the response is markedly improved and pulses as short as 0.03 s can be generated. However, for complete flexibility, transistor controlled or inverter power sources are used which can generate pulses within an almost unlimited frequency range up to 10 kHz. An added advantage of these power sources is the capacity to reproduce accurately complex pulse waveshapes which can be of benefit in controlling the weld pool and solidification structure.

10.1.7 Plasma welding

10.1.7.1 Principles of operation

In plasma welding, the arc is formed between a non-consumable tungsten electrode and the workpiece as in TIG welding. However, the electrode is positioned within the body of the torch and the plasma forming gas is separated from the shielding gas (*Figure 10.44*). Thus, the emanating plasma is constricted by a fine bore copper nozzle which produces a columnar, deeply penetrating arc, compared to the more conical TIG arc.

The penetration capacity of the arc is determined by the degree of constriction of the plasma (diameter of the bore of the nozzle) and the plasma gas flow rate. The electrode angle has no effect on penetration and is usually maintained at 30°. However, as in TIG welding, the gas composition has a secondary influence on penetration. In this instance hydrogen, which increases the temperature of the arc (as shown by the increase in arc voltage), is particularly effective. Helium is also used to increase the temperature of the plasma but, because of its lower mass, penetration can decrease in certain operating modes.

The properties of the constricted plasma with variable arc force, which results from varying the plasma gas flow rate, have led to three distinct welding process variants:

(1) microplasma welding, 0.1–1.5 A;
(2) medium current plasma welding, 15–100 A; and
(3) 'keyhole' plasma welding, >100 A.

10.1.7.2 Microplasma

Microplasma welding has been so termed because a very stable arc can be maintained, even at welding current as low as 0.1 A. It is possible to vary the arc length over a comparatively wide range, up to 20 mm, without adversely affecting arc stability and because of the columnar nature of the plasma, without causing excessive spreading of the arc. With TIG welding, the arc is more sensitive to variation in torch distance, both with regard to stability and to spreading of the arc, due to its conical shape.

10.1.7.3 Medium current

At higher currents, that is up to 100 A, the plasma arc is similar to the TIG arc, although it is slightly 'stiffer' and more tolerant to variation in arc length. The plasma gas flow rate can also be increased to give a slightly deeper penetrating weld pool, but with high flow rates there is a risk of shielding gas and air entrapment in the weld pool

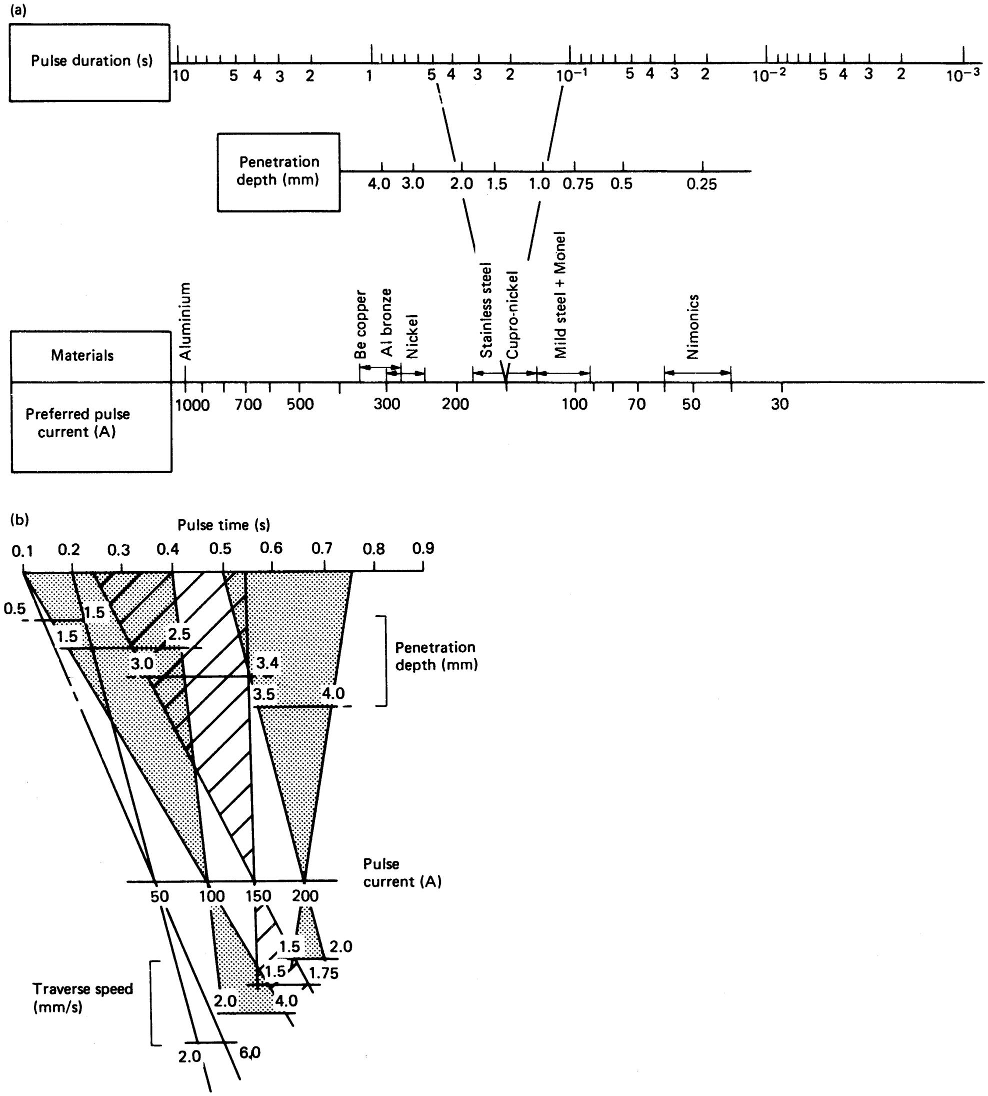


Figure 10.43 Nomographs as an aid to the selection of pulse parameters in pulsed TIG welding. (a) Theoretical pulse parameters according to material type; pulse duration determined by thicknes of material. (b) Selection of pulse parameters in 304 stainless steel based on practical welding trials. Note that the pulse current level is selected according to the material-thickness range

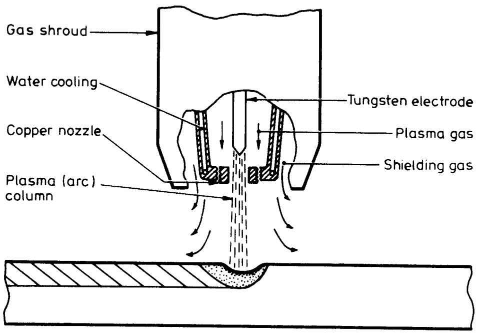


Figure 10.44 The plasma process showing constriction of the arc by a copper nozzle

through excessive turbulence in the gas shield and agitation of the weld pool.

10.1.7.4 *Keyhole*

The most significant difference between the no and plasma welding arcs lies in the keyhole technique. A combination of high welding current and plasma gas flow rates forces the plasma jet to penetrate completely the material, forming a hole as in laser or electron beam welding (*Figure 10.45*). During welding, this hole progressively cuts through the metal with the molten metal flowing behind to form the weld bead under surface-tension forces. The deeply penetrating plasma is capable of welding, in a single pass, relatively thick sections up to approximately 10 mm. However, despite the tolerance of the plasma process to variation in

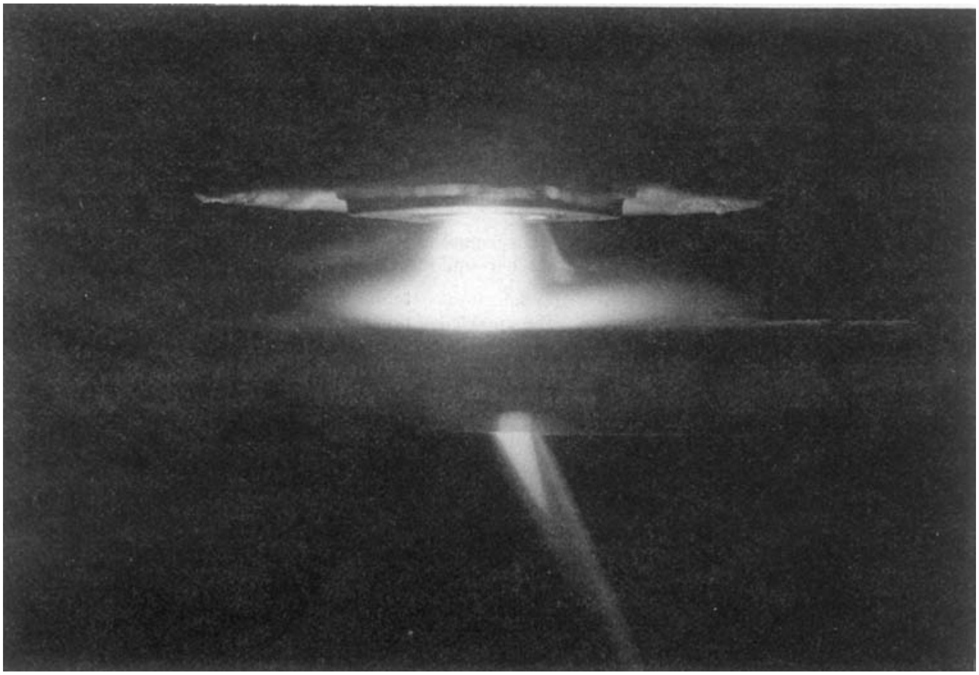


Figure 10.45 Formation of the keyhole and efflux plasma in plasma keyhole welding of 5.0 mm stainless steel. The efflux plasma is inclined backwards as the arc cuts through the plate

torch-to-workpiece distance, this technique is more suitable to mechanised welding, as the welding parameters, i.e. welding current, plasma gas flow rate and welding speed, must be carefully balanced to maintain the stability of the keyhole and the weld pool. Instabilities can easily result in the loss of the keyhole, giving only partial penetration of the weld bead and increasing the risk of porosity.

10.1.7.5 A.c. arc

Sine wave a.c. The plasma arc is not readily stabilised with sine wave a.c. for two reasons: arc re-ignition is difficult when operating with a constricted plasma, and a long arc length and the progressive balling of the electrode tip (during the electrode positive half-cycle) severely disturbs arc root stability. Thus, plasma welding of aluminium is not widely practised, although successful welding has been reported using d.c. (negative polarity) and helium shielding gas.

Square wave a.c. The recent advent of the square wave power supplies described in Section 10.1.6.3 has made it possible to stabilise the a.c. plasma arc without the need for continuously applied high frequency for arc re-ignition. In addition, by operating with only 30% electrode-positive, the electrode is kept so cool that a pointed electrode tip, and hence arc stability, can be sustained. It is particularly important, however, that to limit electrode/nozzle erosion, the maximum current is reduced to less than that which can be operated with a d.c. plasma arc. For example, using 30% electrode-positive, the current rating of a 4.8 mm diameter tungsten electrode with a 40°∢tip angle would be reduced from 175 A (d.c.) to approximately 100 A (a.c.). Furthermore, any increase in the proportion of electrode positive polarity, so as to improve arc cleaning, would significantly reduce the maximum operating current.

Despite the reduction in the maximum current at which the various electrode sizes can operate, stabilisation of the a.c. arc represents a significant advance in plasma welding. Until now, in the welding of aluminium, no advantage could be taken of the deep penetration capability of the plasma arc—because of the need to use a blunt electrode, the alternative a.c. TIG process produces shallow penetration.

10.1.7.6 Electrode and nozzle

The electrode in the plasma system is normally tungsten–2% thoria. Typical electrode diameters, vertex angles and plasma nozzle bore diameters for the various current ranges are given in *Table 10.3*. At low and medium currents, the electrode is sharpened to a point whilst at high currents, it is blunted to approximately 1 mm diameter tip.

The plasma nozzle bore diameter, in particular, must be selected carefully and it is prudent to employ a nozzle the current rating of which is well in excess of the operating current level. The plasma gas flow rate can also have a pronounced effect on the nozzle life with too low a flow rate possibly leading to excessive erosion. Multi-port nozzles, which contain two additional small orifices on each side of the main orifice, can be used at high current to improve control of the arc shape. Use of an oval or elongated plasma arc has been found to be beneficial in high current welding, particularly when operating in the keyhole mode.

10.1.7.7 Plasma and shielding gas

Argon is the preferred plasma gas as it gives the lowest rate of electrode and nozzle erosion. Helium can be used for

Table 10.3 Maximum current for plasma welding for selected electrode diameter, vertex angle and nozzle bore diameter. (Levels are for guidance only, it is important to refer to manufacturer's recommended operating conditions for specific torch and plasma nozzle designs)

Maximum current (A)						*Plasma**		*Shielding*†⇐	
Torch rating (A)									
20	*100*	*200*	*400*	*Electrode diameter* (mm)	*Vertex angle* (°)	*Nozzle bore diameter* (mm)	*Flow rate* (l/min)	*Shroud diameter* (mm)	*Flow rate* (l/min)
Microplasma									
5				1.0	15	0.8	0.2	8	4–7
10						0.8	0.3		
20						1.0	0.5		
Medium current									
	30			2.4	30	0.79	0.47	12	4–7
	50					1.17	0.71		
	75					1.57	0.94		
	100					2.06	1.18		
		50		4.8	30	1.17	0.71	17	4–12
		100				1.57	0.94		
		160				2.36	1.42		
		200				3.20	1.65		
			180	3.2	60‡⇐	2.82	2.4	18	20–35
			200			2.82§	2.5		
High current									
			250	4.8	60‡⇐	3.45§	3.0		20–35
			300			3.45§	3.5		
			350			3.96§	4.1		

* Argon plasma gas.
† Argon and argon–5% H_2 shielding gas.
‡⇐Electrode tip blunted to 1 mm diameter.
§ Multi-port nozzle.

medium and high current operations to increase the temperature of the plasma which, in the melt (non-keyhole) mode, will often promote higher welding speeds. However, the use of helium as the plasma gas can reduce the current-carrying capacity of the nozzle. Furthermore, because of its lower mass, weld-pool penetration will be reduced which, in certain materials, will make the formation of a keyhole difficult. For this reason, helium is seldom used for the plasma gas when operating with the keyhole mode. Hydrogen is often added to the shielding gas, argon plus 2.5%H_2 being the most common mixture, but up to a maximum of 15%H_2 can be used to produce a hotter arc and a more reducing atmosphere. Hydrogen also constricts the arc, which can increase the depth of the weld pool penetration and promote higher welding speeds.

Helium or a helium–argon mixture, typically 75% helium–25% argon, can also be used as the shielding gas. Whilst a hotter arc will be generated, it is less constricted, which can result in a wider weld bead compared with argon or argon–hydrogen shielding.

10.1.7.8 Pulsed current

Similar benefits of improved arc stability and better control over the behaviour of the weld pool can be derived from pulsing the welding current. However, pulsing has special advantages for improving the keyhole mode of operation which, with continuous current, requires careful setting of the welding parameters.

The same principle applies in that a high current pulse causes rapid penetration of the material and establishes a stable keyhole and weld pool. If this high current were maintained, the keyhole would continue to grow, causing excessive penetration and, ultimately, cutting would occur. Therefore, the pulse is terminated after a pre-set time and the weld pool allowed to solidify under a low background or pilot arc. It is equally important that the plasma gas flow be maintained during this period so that the keyhole does not close and, on re-applying the pulse current, the plasma can quickly penetrate the plate, re-establishing a stable keyhole and weld pool. Thus, welding progresses in a series of discrete steps with the pulse frequency balanced to the traverse rate to produce overlapping weld spots (see *Figure 10.46*). In pulsing, the important variables are:

Welding current	*Plasma gas*
Pulse time	Pulse level
Pulse level	Background level
Background time	
Background level	

Selection of welding parameters can be simplified, first with the knowledge that the pulse time is determined more by the physical requirements of forming the keyhole and weld pool at a given traverse rate, than by the plate thickness or material composition. For most materials, within a plate thickness of 3–5 mm, a minimum pulse time of 0.1 s is required to re-establish the keyhole and weld pool. At longer pulse times, the excess energy is largely dissipated in the efflux plasma. The background time is usually set equal

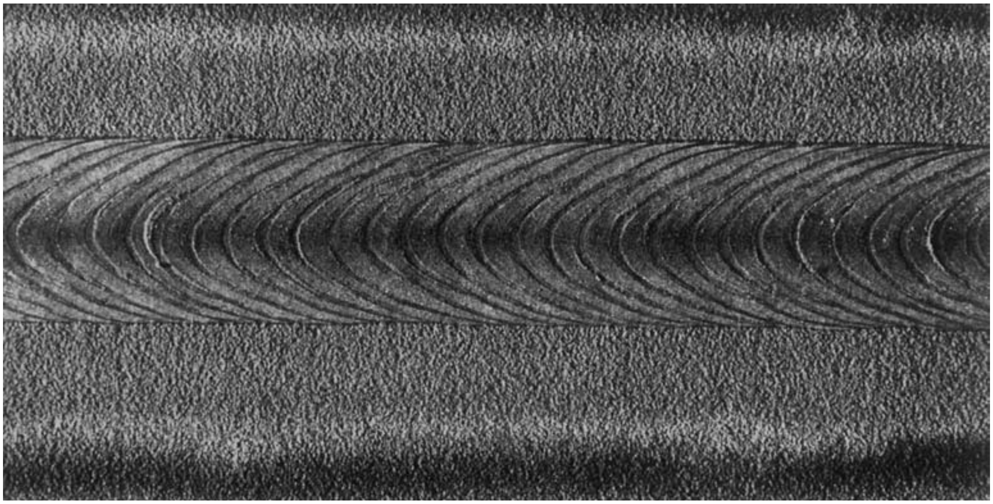


Figure 10.46 Appearance of the pulsed (keyhole) plasma arc weld

to the pulse time, which is sufficient for solidification between pulses. Thus, the pulse frequency is determined by the welding speed and the need for at least 60% overlap of the pulses to provide a continuous seam. For instance, when welding 4.4 mm stainless steel at a speed of 0.15 m/min, a suitable frequency is 2 Hz.

It follows that the pulsed current level and plasma gas flow rate are the major welding parameters which must be set to give an over penetrating plasma for a particular material composition and plate thickness combination. The background current is held low to give rapid cooling between pulses, while the plasma flow rate is held constant to maintain the keyhole. For instance, when welding 4.4 mm austenitic stainless steel, the pulsed current and plasma gas flow rate are typically 140 A and 2 l/min, respectively. However, when welding the same steel in 5.0 mm thickness, the pulsed current and plasma gas flow rate are increased to 190 A and 2.3 l/min and all other parameters are held constant.

10.1.8 Electroslag and electrogas welding

10.1.8.1 Principles of operation

Electroslag welding, originally developed for the welding of thick mild and low-alloy steels, is restricted to welding in the vertical or near-vertical position. *Figure 10.47* shows the basic arrangement.

Once established, the electroslag process is basically arcless, the heat required to melt the wire and fuse the parent material being supplied by resistive heating of the molten slag bath. To ensure uniform fusion of the joint faces, it is necessary with thick material either to use more than one electrode or to oscillate the electrode(s) across the joint. As the weld progresses, the level of the weld pool rises and the welding head and water-cooled shoes are moved slowly up the joint. A typical speed for electroslag welding of 76 mm thick mild steel would be 1 m/h at 550 A and 44 V. Since the process is only stable once a slag of appropriate depth has been established, it is essential that a run-on plate is provided so that the defective start position of the weld can be removed on completion of the joint.

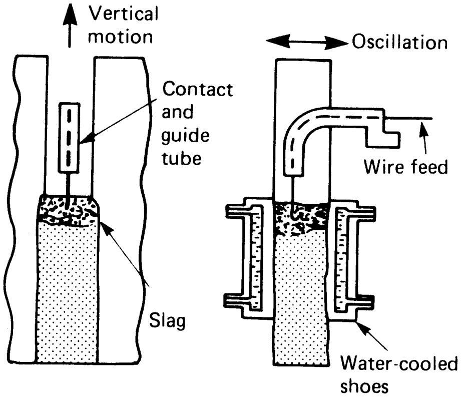


Figure 10.47 Electroslag process

As the electroslag process is virtually a continuous casting process, the resulting welds have large grain size and a tendency towards columnar growth. Because of this, the weld metal and heat-affected zone generally have extremely poor (mechanical) properties. While this can be mitigated by a post-weld, normalising heat treatment, such treatment may cancel out the economic advantages that the process otherwise offers.

Although electroslag welding is mainly used for the joining of mild and low-alloy steels, satisfactory welds have been made in high-alloy steels, titanium and aluminium.

10.1.8.2 Consumable-guide welding

Consumable-guide welding is a much simplified version of electroslag welding. This simplification stems mainly from the use of a wire guide which is progressively melted into the weld. The consumable guide (or guides) eliminates the need for a moving welding head and also results in faster welding speeds. The two moving water-cooled copper shoes used for electroslag welding are generally replaced by two pairs of shoes which are leap-frogged up each side of the joint as welding progresses.

10.1.8.3 Electrogas welding

Electrogas welding is very similar in principle to electroslag welding in that it is used for welding thick section material in a single pass and in the vertical position. The main difference being that the heat for welding is generated by an arc which is formed between a flux-cored electrode and the molten weld pool. The flux from the electrode forms a protective layer over the weld pool, but additional protection in the form of a gaseous shield (usually CO_2) is used. The electrogas process is generally faster than electroslag welding when used on relatively thin sections. The resulting weld metallurgy is similar to a high-current submerged-arc weld.

10.1.9 Metal cutting and gouging

Metal can be removed from workpieces by use of coated electrode, plasma-arc and carbon arc processes.

10.1.9.1 Coated-electrode process

Flux-coated electrodes are available in two types. The first has a specially formulated coating that gives a deeply penetrating and forceful arc, tending to create a weld pool and to blow out the molten metal: a hole or recess is left, which can be carried along the material. When thick material is being cut, a sawing action is employed. The second type of electrode is a hollow electrode through which a high-pressure stream of oxygen is passed. The cutting action is akin to that with oxyacetylene.

Electrode techniques are, in the main, limited to ferrous materials.

10.1.9.2 Plasma-arc process

The plasma-arc process can give neat and accurate cutting. The principles are basically the same as in plasma-arc welding, but the force of the plasma column is increased by higher gas flow through the nozzle and a higher power output. Virtually any conducting material can be cut, at speeds faster than those of other methods, e.g. 8 m/min for 6 mm thick aluminium.

The most common gases are argon–H_2, nitrogen, oxygen and air. The use of air has increased in recent years especially for cutting thin sheet steel, e.g. automotive and ‘do-it-yourself’ users. The torch which is suitable for air and oxygen employs a water-cooled, copper–hafnium tipped electrode.

The power source is similar to the ones used for plasma welding, but the output (cutting) voltage can be as high as 200 V. Safety precautions in the use of these power sources are essential (see appropriate standards e.g. IEC 60974-8)

10.1.9.3 Carbon arc process

In this process an arc between a carbon electrode and the workpiece is used to create a molten pool of metal which is blown away by a jet of compressed air. The method is useful for all grades of ferritic steels.

10.2 Resistance welding

Resistance welding processes involve the welding of two or more metal parts together in a localised area by the application of heat and pressure. The heat is generated by the resistance to the passage of a high current through the metal parts held under a preset pressure. Copper or copper alloy electrodes are normally used to apply pressure and convey the electrical current through the workpieces. In the case of spot and seam welding, the electrodes or wheels are shaped to concentrate the force and current. In projection welding and butt welding the shape of the components dictates the weld area. No consumable materials such as welding wire, fluxes or gases are required.

The heat generated depends on the current (I), the time of application (t) and the resistance (R) and is proportional to I^2Rt. The heat generated is governed largely by the bulk resistance of the materials being joined and the interface resistance at the contact between the materials. In some cases, a fused zone is produced, as in the spot or seam welding of sheet materials. However, in many cases of projection welding, and particularly resistance butt and flash welding, a forge weld is produced without melting. The plastic deformation of the heated parts in contact produces a bond in the solid state analogous to the blacksmith’s weld.

Resistance welding can be used to join a wide range of materials and the ease of welding depends on the metallurgical compatibility of the materials and their electrical resistivity. Mild steel is readily weldable, having a resistivity some six times that of the copper alloy electrode. Surfaces should be free from rust, scale, dirt and other materials likely to hinder current flow but light oils or lubricants do not normally interfere with the welding. Aluminium alloys, with their higher conductivity, require up to three times the current required for steel to develop the heat for welding. In addition, the surface resistance of aluminium alloys is strongly influenced by the oxide film and this needs to be closely controlled for high quality welds. Anodised or heavily passivated surfaces can be insulating and therefore impossible to weld.

10.2.1 Welding equipment

Figure 10.48 shows the essential features of a spot welding machine. The three basic units are a structural frame, a force application system and the electrical system. The frame provides strength and rigidity to react the electrode force without undue flexure of the machine. The electrical circuit on the low voltage secondary side of the transformer may constitute part of the frame or, in the case of the welding gun, provide most of the structural strength. The force application system is normally pneumatic. Regulated air pressure is fed to a cylinder to provide the electrode force. Servomotors are increasingly being used to provide programmable control of electrode position, approach speed and force. Hydraulic systems are used occasionally for compact, high force applications, and springs are used for very small hand or foot operated machines. The electrical system comprises a welding transformer and a timer/controller unit.

10.2.1.1 Machine types

Pedestal or bench mounted machines are fixed types and the workpieces are fed into the machines manually or mechanically using a carousel or pick and place device. Gun welders are used for welding fixed, often larger structures and the gun is manipulated either manually, being suspended from a counterbalance system, or using a robot. The transformer may either be remote from the gun, connected by a heavy water cooled kickless cable, or integral with the gun itself. In the latter case, the transformer can be of a substantially smaller size because of the reduced secondary impedance, and current is supplied at mains voltage through smaller

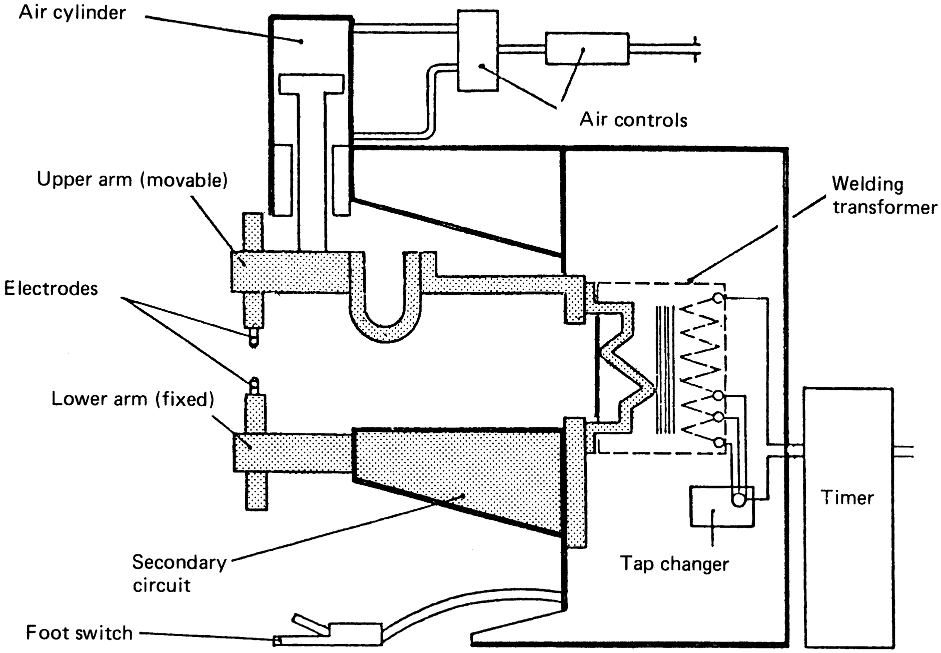


Figure 10.48 A pedestal spot welding machine

cables. This allows the gun to be more manoeuvrable despite its greater weight, often in the region of 60 to 100 kg. Modern residual current circuit breakers provide the necessary safety protection for manual operation of integral transformer guns.

Multiwelders can be used to weld components in a dedicated machine where each spot is welded with a separate pair of electrodes. One electrode is operated with its own cylinder and the opposing electrode is fixed. A frame provides the structural support and reacts the electrode force. A number of transformers may be used, each usually serving two or more electrode pairs from a two turn secondary with separate outputs. If more than one electrode pair is supplied from one transformer connection, the machine is sequenced so that only one weld is made at a time, the other electrodes remaining open. In order also to prevent too high a mains current demand from simultaneous firing of the transformers, the weld times are sequenced or cascaded. Where two electrode pairs are connected to one transformer, it is important that the circuits are similar so that the impedance, and thus the current drawn, is the same in each circuit.

10.2.1.2 Resistance welding power supplies

The most common power supply comprises a single-phase a.c. transformer, see *Figure 10.49*. This converts the mains supply primary voltage to a low (2–20 V) secondary welding voltage. The turns ratio of the transformer is the number of turns of the primary conductor, divided by the number of turns of the heavy secondary conductor (usually 1 or 2). This is the ratio by which the voltage is reduced and the mains current magnified. The open circuit secondary voltage may be considered nominally constant, so the current drawn on the welding circuit depends on the circuit impedance according to Ohm's law. This circuit is virtually a dead short with a resistance in the region of $10^{-3}\,\Omega$. Thus, several thousand amperes of welding current can be developed at the low secondary voltage.

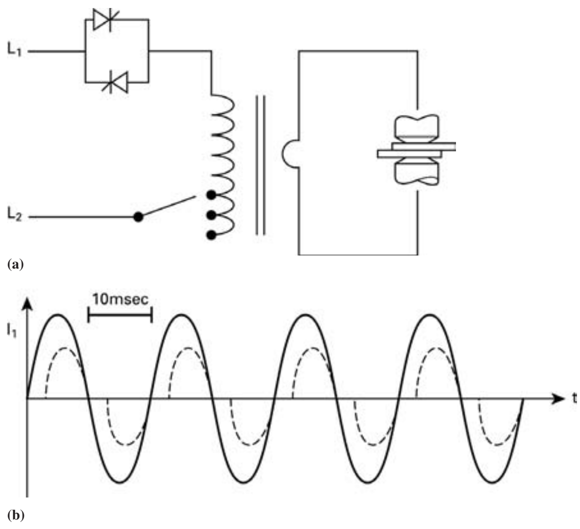


Figure 10.49 Single-phase a.c. power supply. (a) Welding circuit; (b) Current waveform, broken line illustrates phase shift control of current level

The power capacity of a machine is normally quoted in kVA at 50% duty cycle. This refers to the power that may be drawn by the transformer over a long period of time, with current flowing for 50% of that time, without causing the transformer to overheat.

The maximum allowed current on the primary side of the transformer at 50% duty cycle is then the kVA rating divided by the mains voltage. In practice, much higher primary current levels can be drawn by resistance welding machines for short times, and this is acceptable, provided account is taken of the actual duty cycle (current on time/total time) expressed as a percentage.

The allowable power at duty cycle $x\%$ may be calculated as follows:

$$\mathrm{kVA}_{x\%} = \mathrm{kVA}_{50\%}\sqrt{\frac{50}{x}}$$

The single-phase a.c. transformers are of the shell/core type of a compact design and the insulation is cured to provide a robust unit capable of the mechanical and electrical loading typical of resistance welding equipment. Water cooling may be provided on the low voltage secondary conductors, thus improving its thermal rating. A number of standards cover the specific requirements of transformers such as ISO5826 (BS7125) and BS EN ISO 7284. Many transformers have tap switches so that the secondary voltage can be changed to provide step changes in the welding current range. Fine control of current is by phase shift control using the timer/controller.

It is the available welding current rather than kVA rating that determines the suitability of a machine to weld a particular component. While the maximum short circuit current may be known for a particular machine, the welding current available will be much lower. This is because of the added resistance of the component being welded and the associated electrode or tooling. The secondary circuit impedance is also influenced by an inductive component, which is related to the area within the machine throat. Current is reduced if this area is increased, and even more so if there is steel within the throat. Thus, the arm spacing and routing of jumpers or flexible connections is important in order to minimise the losses.

A number of alternative power supply types are available, as follows:

Secondary rectified d.c. This type is used for high power applications as it uses a balanced three-phase supply and the rectified d.c. welding current is not subject to the inductive power losses suffered by a.c. current, see *Figure 10.50*. Single-phase d.c. machines are available but are much less common.

Frequency converter d.c. Primary rectification of a three-phase supply avoids the need to rectify high welding currents. A short duration d.c. pulse (less than 0.2 sec) is then delivered to the welding transformer. The d.c. polarity is changed for each pulse to avoid transformer saturation. This type of equipment has been widely used for high power applications, such as high quality welding of aluminium alloys.

Inverter welders Medium frequency inverter welders are used to enable more efficient, lighter weight welding transformers to be used, typically for robot guns. A three-phase supply is first rectified then chopped with a transistor unit to give medium frequency, typically up to 1000 Hz ac. This a.c. at about 600 V is then converted in the welding transformer to low voltage/high current at the same frequency. This is immediately rectified to give a d.c. welding current. These supplies are finding increasing use in robotic spot welding in the automotive industry and also for miniature applications, where fine control of the current pulse shape can be achieved.

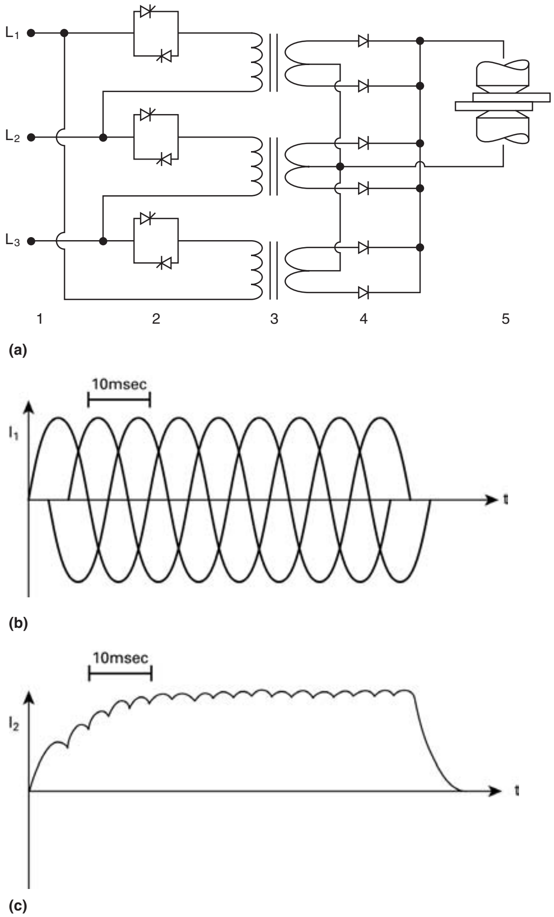


Figure 10.50 Three-phase secondary rectified d.c. power supply
(a) Welding circuit. 1—mains supply, 2—thyristor current control, 3—welding transformer, 4—secondary rectification, 5—welding electrodes (b) Primary current waveform. (c) Welding current waveform.

Capacitor discharge A short duration d.c. pulse, typically less than 10 msec, is achieved by discharging a bank of capacitors through a welding transformer. A low power demand is required to charge the transformers. Such equipment is common for miniature applications, but large machines are available for large projection welding operations. These are often portal machines and have high force capacity with a fast follow-up, low inertia head. The heat input is relatively low compared to a.c. or d.c. power supplies.

Transistorised power supplies These are generally small to miniature supplies, up to about 5 kA. The current is pure d.c. and the pulse can be shaped accurately for fine control of the resistance heating of difficult material combinations or configurations, particularly in electrical and electronic applications.

In some cases, especially miniature applications in dissimilar materials, d.c. current is used and its polarity may be used to advantage in modifying the heat balance within a weld. The heating may be biased towards one electrode due to the Peltier effect.

10.2.1.3 Mains supply

When connecting resistance welding equipment to the mains supply, a number of factors need to be considered. As the primary current demand is high, the high voltage transformer (e.g. 11 kV) supplying the system should be of sufficient capacity that there is not an undue voltage drop on the mains voltage side. Cables connecting the machines do not necessarily need to be rated at the maximum continuous thermal rating for the size of machine being connected, as the duty cycle of the machine is usually low (except for seam welding). Account should be taken of the maximum anticipated duty cycle and any relevant local regulations. It is also important to consider the voltage drop associated with the supply cables, according to their size, length and the current demand. A reduction in supply voltage at the welding machine will reduce its capacity proportionally. The cable layout should also be so organised as to minimise reactive losses. Resistance welding equipment, particularly a.c. types have a reactive load and consequently a relatively low power factor. Occasionally, capacitive power factor correction equipment is used in the plant to reduce the voltage/current phase difference.

10.2.1.4 Resistance welding electrodes

Resistance welding electrodes play an important part in the achievement of good and consistent weld quality. The materials used are normally copper alloys, internally water-cooled, which provide a combination of high electrical and thermal conductivity, together with mechanical strength. This is required to withstand the electrode forces and a certain amount of hammering of the electrodes on initial contact. The dimensions of spot welding electrodes are covered by various ISO standards.

Electrode materials for resistance welding are covered by ISO 5182. The highest conductivity class 1 alloys, such as Cu/0.1%Zr, are used for spot welding coated steels and aluminium alloys. Harder, class 2 alloys with a conductivity of 75 to 80% IACS (International Annealed Copper Standard), such as Cu/1%Cr, are used for steels in general. Class 3 alloys, with 30 to 45% IACS, provide additional hardness for spot welding stainless steels and for projection welding dies, although tungsten/30%Cu inserts are also often used in projection welding. Refractory metals such as molybdenum or tungsten are also used for certain applications.

The diameter of the electrode contact face for spot welding should approximate to $5\sqrt{t}$, where t is the single sheet thickness in millimetres (up to 3 mm). This allows a weld of a similar diameter to be produced. In order to maintain weld quality, the electrode contact area and profile should be maintained by periodic electrode dressing or replacement by re-machined electrodes. An electrode dressing tool or form tool should be used where possible for greatest reliability. Automatic dressing equipment can be used in robotic applications. Typically, a few hundred spots for aluminium alloys, a few thousand for coated steels and several thousand for uncoated steels, are possible before electrode maintenance is required.

10.2.2 Welding process

10.2.2.1 Welding sequence control

In order to achieve the highest quality and repeatability in resistance welding, the control of welding current and its duration must be accurate. Modern electronic timers provide two functions: (a) the synchronous timing of the weld sequence and current pulse, measured in cycles of mains frequency, and (b) control of the current level by phase shift control. Furthermore, on air-or hydraulic-operated equipment, the timer controls a solenoid valve and the welding sequence is interlocked once initiated. This takes the control of the machine out of the hands of the operator and prevents the electrode force being released until the end of the hold time. The basic welding sequence shown in *Figure 10.51* comprises the 'squeeze time', during which the electrode force rises to its preset level, the 'weld time' during which current flows and the weld is formed and the 'hold time during which the weld is allowed to cool under pressure. The electrodes are then released. When spot welding 1 mm low carbon steel, the weld time would be about 0.2 sec, the welding current about 7 kA and the whole sequence less than 1 sec. More complex sequences can be used for special applications, and features include multiple current pulses, slope up of current and variation of force during the weld.

Solid state switches (thyristors) are used to control the flow of primary current to the transformer, the timer can be set to deliver the required duration of current, i.e. the number of cycles at 50 Hz mains frequency. The magnitude of the current can be adjusted finely by altering a heat or phase shift control. This changes the point at which the thyristors switch on during each half-cycle. The solid-state timers allow precise setting of time and current and often have constant-current control. This is based on feedback of the actual current measurement using a toroidal coil to adjust the phase shift (heat) control during the weld pulse. This compensates for mains voltage fluctuations and variation in secondary circuit impedance. Many also have programmable control to enable automatic equipment to select preset welding programmes, where one machine is required to weld a variety of thickness combinations, for example. Additional functions include diagnostics, fault detection and stepper functions, which allow welding current to be increased progressively to compensate for gradual electrode wear.

Prior to the introduction of modern controllers, weld timing developed from relatively crude control. Early machines that worked on a mechanical lever system incorporated the contactor for primary current within the lever system. Thus, the time of the weld was left to the judgement of the operator. This, however, led to unsuitable and unreliable welding conditions being used. By introducing simple timers and an electrically operated contactor, the duration of current flow was more reliable. Mechanical contactors were superseded by mercury vapour switches (ignitrons), which have now been replaced by thyristors.

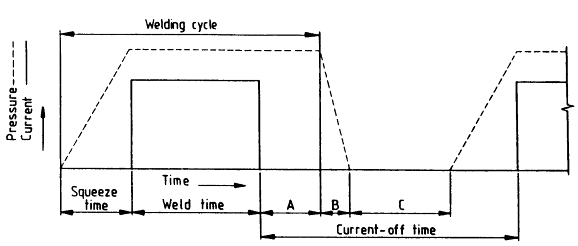


Figure 10.51 Timing sequence for spot, stitch and projection welding. A—hold time (forging time); B—pressure decay time (not critical); C—pressure off time

10.2.2.2 Welding parameters

The main welding parameters are electrode force, weld time and current. Recommended settings are available in the literature for many standard materials and applications, and guidelines are presented for spot welding uncoated and coated low carbon steels in BS 1140.

The electrode force required for spot welding low carbon steel is typically 1.5 to 2 kN per mm of its sheet thickness and is slightly higher for coated steels. The force required for stronger and high hot-strength materials such as stainless steel or nickel alloys may be 2 to 4 times this level.

Weld times for low carbon steels are normally about 7 to 10 cycles per mm of the single sheet thickness in the range 0.4 to 3 mm. The shorter times should be used in conjunction with the higher forces. Higher welding current is required at shorter weld times. However, it is not possible to compensate for insufficient current by excessively long weld times.

Welding current is adjusted to achieve a weld of the required size once other parameters are fixed. The current needed depends on the material type, the size of the electrode tip used and the other parameters set. Typically 7 to 10 kA is required for 1 mm and 15 to 20 kA for 3 mm uncoated steel. Higher current is required for coated steels because of the lower interface resistance.

It is important that welding current flows only through the electrode contacts and is not shorted through other parts of the tooling or components. Current can be lost through adjacent welds (current shunting). The degree of shunting depends on the material thickness and weld spacing. Welding current must be increased to compensate for close pitch welds (e.g. less than 3 times the tip diameter for thin sheet). Steel in the throat of the machine also causes a reduction in current on a.c. machines because of the increased inductive effect. Automatic compensation is possible in this case by using a constant current controller.

10.2.2.3 Weld quality and testing

The principal quality requirement of spot welds in sheet materials is weld size, and this is indicated by a destructive peel or chisel test to tear a plug or button from one of the sheets. The minimum acceptable weld diameter depends on the application standard but is typically $3.5\sqrt{t}$ or $4\sqrt{t}$ for sheet thicknesses up to 3 mm. A plug failure should normally occur and an interface failure may indicate the weld is weak due to lack of fusion or embrittlement.

Surface defects such as splash metal, cracks and excessive deformation are normally unacceptable, but metallographic examination of internal defects is usually only necessary for aircraft quality work.

For projection or butt welded joint configurations, testing in tension, shear, bend or torsion can be conducted to suit the component and service requirements.

Quality control in resistance welding is generally by process control and periodic destructive testing. Monitoring or routine checking of current, weld time and force can be done to ensure consistency of the main process parameters. A range of commercial in-process weld quality monitors and feedback controllers are also available that examine additional parameters such as weld resistance, weld energy or weld expansion. In order to exploit such monitors, the process must first be under control. A positive correlation between monitor output and quality must be demonstrated, and the monitor should be properly set up and maintained. Confidence in such monitors is easily lost if these criteria are not met.

Post-weld non-destructive testing of resistance welds is limited. Ultrasonic testing of spot welds requires a special-purpose high-frequency probe, incorporating a water column retained by a plastic membrane bubble, which is applied to the spot weld indentation. Skill and extensive training are required to interpret the multiple reflections on the flaw detector, but substantial reduction of destructive tests has been achieved by automotive manufacturers in particular.

10.2.2.4 Twin-spot welding

There are a number of cases where it is convenient to use twin-spot welding. These include welding components where access is limited to one side or where components are to be welded to large sheets. It is often possible to use a twin-spot method of operation where both electrodes are on one side of the material, the current being taken through a copper backing bar or plate below the bottom sheet (*Figure 10.52*(*a*)). If the lower sheet is significantly thicker than the upper sheet, the copper backing plate may be replaced by an insulated support to react to the load, or eliminated completely where a closed section is being welded.

An alternative technique, referred to as 'push–pull', involves initiating separate transformers simultaneously each side of the component, feeding two pairs of electrodes (*Figure 10.52*(*b*)). Series and push–pull welding allow small,

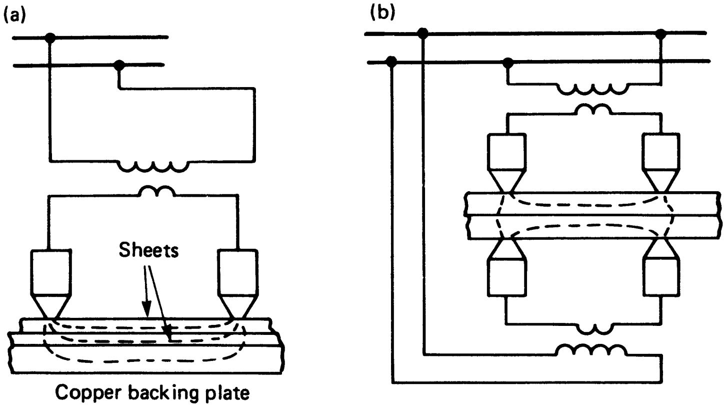


Figure 10.52 Twin-spot welding: (a) series weld; (b) push–pull weld

low inductive loss secondary circuits to be used, and minimise the amount of steel within the throat of the machine. This is an advantage on large components as smaller capacity equipment can be used.

10.2.3 Projection welding

In projection welding, the size and position of the weld or welds are determined by the design of the component to be welded. The force and current are concentrated in a small contact area between the components to be welded. This occurs naturally in cross wire welding, or is deliberately introduced by machining or forming. An embossed dimple is used for sheet joining and a V projection or angle can be machined on a solid component to achieve an initial line contact with the component to which it is to be welded. Nuts and studs are also widely projection welded. *Figure 10.53* shows some typical configurations.

Some of the advantages of projection welding are:

- Output is increased because a number of welds can be made simultaneously.
- Electrode life is extended because electrodes with large contacting surfaces may be used and these may have harder, sintered copper/tungsten inserts.
- The process is versatile, allowing a wide range of component designs to be resistance welded.
- Welds may be more closely spaced than practicable with spot welding without compensation for current shunting, and
- Minimum surface marking can be achieved on one side of the joints.

Consistency of projection height and shape, together with good electrode alignment and flatness are essential to ensure uniform share of the available force and current. Welding times are short, to ensure that sufficient heat is generated in the small contact area before the projection collapses completely. Rapid follow-up of the welding head ensures that the weld is consolidated as the projection collapses.

Projection welding machines operate on principles similar to those of spot welding plant, but are designed for much higher electrode force and current because it is usually required that several projections be welded simultaneously. Very large projection welding machines are available which can apply a total electrode force of several tonnes and deliver a welding current of over 500 kA. The larger machines are generally three-phase d.c., which gives a balanced mains demand and permits more even current distribution owing to a minimal inductive effect. Large capacitor discharge machines with low inertia heads have also been used for projection welding applications. The weld time of only several msec duration results in a weld with low residual heat input. Although the welding current is much higher than for normal a.c. supplies, the mains current demand is low during the few seconds charging stage.

10.2.4 Seam welding

Seam welding (*Figure 10.54*) consists of making a series of overlapping spot welds by means of copper alloy wheel electrodes to produce a leak-tight joint. The electrode wheel applies a constant force to the workpieces and rotates continuously at a controlled speed. The welding current is either pulsed to give a series of discrete spots, or continuous for certain high-speed applications, where up to about 6 m/min can be achieved for 1 mm low carbon steel.

In the conventional technique, the track width of the welding wheel is approximately $5\sqrt{t}$ mm and the minimum acceptable weld width is typically $4\sqrt{t}$ mm. However, a leak-tight weld as strong as the parent material can be achieved with a weld width only a little wider than the sheet thickness. In narrow wheel seam welding, a radiused electrode face is used, giving a smaller contact area (*Figure 10.55*(*a*)). Weld widths of 2–3 mm are made in sheet thicknesses up to about 2 mm. Narrow wheel seam welding is particularly suitable for coated steels. The wheels are driven by a roller, which bears on the edge of the wheel and planishes the contact face, to maintain a uniform condition.

When welding speed is increased above about 20 m/min for very thin material, the spacing of welds made at each half cycle of the mains frequency becomes too great to produce a leak-tight weld. Higher frequency supplies, typically 600 Hz are used to weld tin cans in 0.2 mm thick steel at speeds to over 70 m/min. In this case, a mash weld is produced (*Figure 10.55*(*b*)) where the initial overlap is only 1.5 times the sheet thickness, and the seam is crushed during welding. The effect of electrode contamination is avoided by using a copper wire (formed to a flat oval) which is passed over each electrode in turn. Thus, the electrode face is continuously renewed. D.c. current is not used for such applications as it gives less effective heat generation in this case.

10.2.5 Resistance butt and flash welding

Resistance butt welding is the simplest form of resistance welding and is used predominantly to butt join wires and rods, including small diameter chain. The components to be joined are clamped in opposing dies, with a small stick-out, and butted under pressure. Current from a resistance

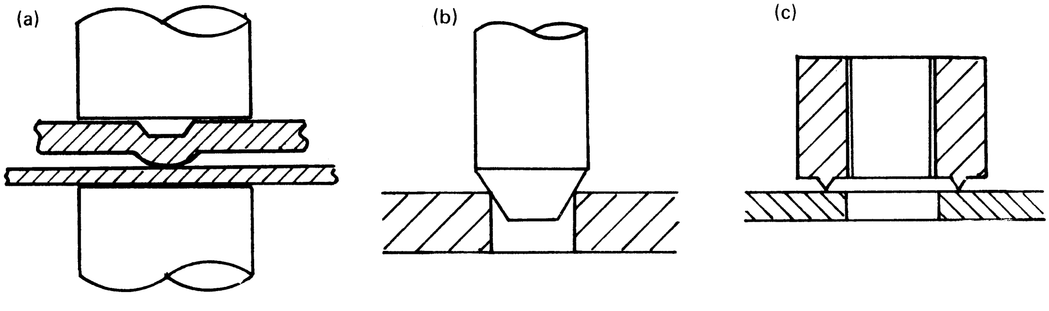


Figure 10.53 Examples of projection welding configurations: (a) embossed projection; (b) stud to plate; (c) annular projection

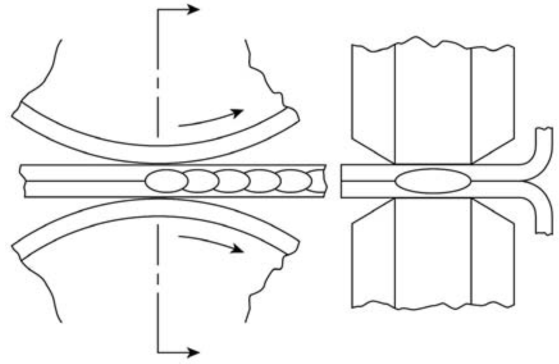


Figure 10.54 Conventional wide wheel seam welding

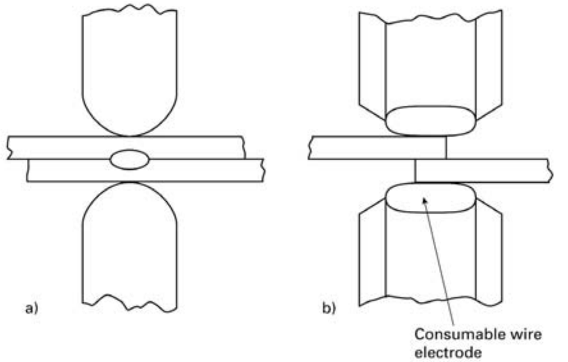


Figure 10.55 Alternative seam welding techniques: (a) narrow wheel seam; (b) mash welding with consumable wire electrode (shown prior to welding)

welding transformer is passed between the dies, causing heating of the weld area. The material deforms under the applied load giving a forge weld. The welding current is normally terminated once a preset reduction in length has occurred, although pressure must be maintained until after current has stopped flowing. Resistance butt welding is not normally suitable for larger components such as thin strip because of unevenness of heat generation. However, the use of d.c. power supplies together with programmed force and current cycles, has enabled components such as automotive road wheels to be welded on automatic equipment at up to 700 rims/hour.

Flash welding is used for a wide range of component shapes and sizes from bicycle wheel rims to rails. More efficient energy input, and a more localised and evenly heated zone, can be achieved compared with resistance butt welding. In this case, the components are clamped between dies and moved slowly towards each other with the current switched on. Current flows through successive point contacts, which heat rapidly, melt and blow out of the joint, giving the characteristic flashing action as the forward motion continues. After a preset material loss has occurred, sufficient to heat the material behind the interface to its plastic state, the components are forged together to expel melted material and contaminants, and complete a solid phase forge weld. The flash or upset metal may be removed while still hot using a shearing tool. When welding larger sections, the parts may be preheated to promote easier flashing. This is done by advancing and retracting the components to make repeated short-circuit contacts under pressure. The welding sequence and the welding machine are shown schematically in *Figures 10.56* and *10.57*.

Machines are predominantly single-phase a.c. in the power range 20–500 kVA. Such machines are designed with as small as possible secondary inductive loop as the flashing action is characterised by repeated breaks in the secondary circuit which cause transient arcs. The lower the inductance, the lower is the energy dissipated in the arc and the smoother are the melted surfaces prior to final forging. This helps to reduce the risk of weld interface flaws. Large d.c. flash welders have been produced to reduce the mains power demand, but there is a risk of sustained arcs produced during flashing, leading to undesirable deep melted areas at the interface. Modulated current may be used to counteract this tendency.

The flashing sequence must be accurately controlled and the forge applied with sufficient force and speed for best results. The forward movement of the components is normally accelerated as flashing progresses. The flashing speed

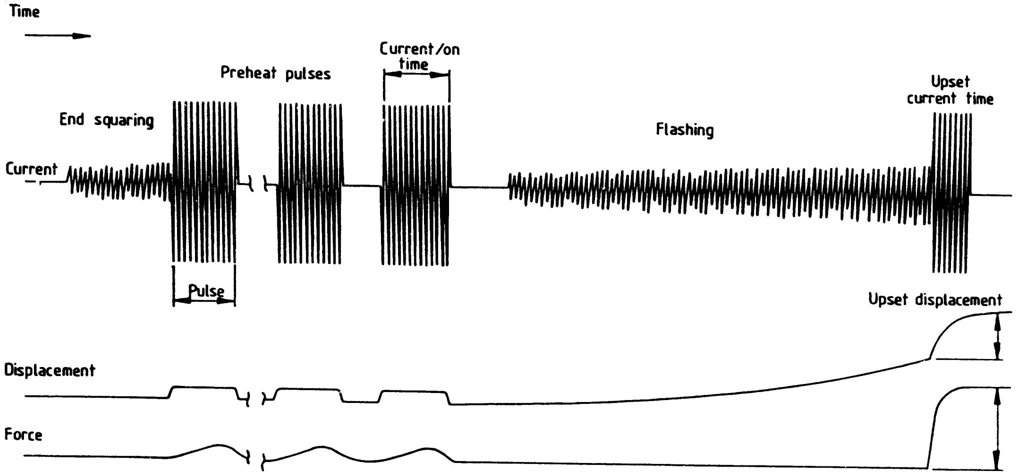


Figure 10.56 Generalised diagram of the flash welding sequence

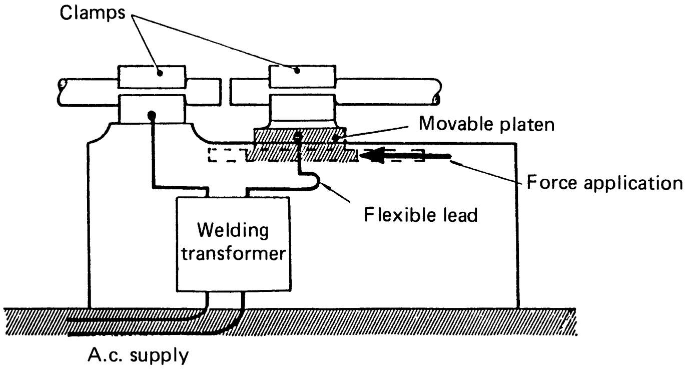


Figure 10.57 A flash welding machine

should be as fast as is practicable for a given applied voltage, whilst avoiding premature butting or 'freezing' of the parts. This ensures that the flashing action is continuous and not so coarse that deep craters are blown in the contacting surfaces with a risk of interfacial flaws. These take the form of a film or planar distribution of oxide inclusions known as 'flat spots'. Feedback control of the travel speed maintains an optimum flashing action.

Welding conditions vary considerably for different materials and component size, but may be generalised as follows for a material of thickness *t*, flashing length 0.7–1.5*t* and upset length 0.2–0.5*t*. Forge forces are between 60 N/mm^2 for low carbon steels and 150 N/mm^2 for stainless steels, with even higher forces required for heat-resisting materials. Flashing speeds are typically 0.5 to 3 mm/s, and upset speeds 25 to 350 mm/s. Guidance on process control and weld testing is provided in BS 6944 for butt welds in steel and BS 4204 for tubular joints in pressure applications.

10.2.6 Safety aspects of resistance welding

There are a number of potential hazards in resistance welding. Although the machines are intrinsically safe, it is important to observe good welding practice, provide adequate training and adopt the appropriate safety measures. Machines should be manufactured and installed to the appropriate standards such as EN 50063.

Mechanical hazards involve the risk of trapping fingers or other parts of the body between electrodes or other moving parts. Safety devices include various types of guard, interlocked two-hand button operation and low force electrode approach. Where practicable, spot welding electrodes should have a working gap of no more than about 6 mm.

Splash metal may be expelled under pressure from the weld. Eye protection and suitable protective clothing should therefore be worn. Burns and lacerations may result from careless handling of hot assemblies or materials with burrs or sharp edges. Suitable gloves and protective clothing should be worn.

Electrical hazards result from inadvertent contact with live terminals. Exposed conductors do not normally exceed 20 V, but mains voltage is connected to the control cabinet and to the transformer taps and primary windings. The machine should be installed and enclosed to the appropriate standards, using the correctly rated cables and protection devices. Equipment should be switched off at the mains before removing covers or opening doors, such as for the purpose of changing taps where a tap switch is not fitted externally. Ideally, such doors should be provided with safety interlocks. An additional hazard may be that the strong magnetic fields produced close to resistance welding equipment could affect the operation of heart pacemakers.

Fumes result from the vaporisation or burning of metal or organic coatings on material being welded, or from interweld adhesives, sealants, etc. This is not normally a major problem and adequate ventilation is usually sufficient. Local extraction may be required in some cases, depending on the type and concentration of the fumes.

10.3 Fuses

Fuses, which are now produced in vast numbers each year to protect electric circuits, were used as long ago as 1864 to protect submarine cables. In theory any conducting metal can be used as a fusible element. In practice, however, a variety of metals are used, ranging from cheaper materials such as copper to rarer and, therefore, more expensive materials such as silver.

The term 'fuse' is used in national and international standards to describe a complete assembly. In its simplest form, this consists of a piece of metal wire connected between two terminals on a suitable support; and at its most complex as a cartridge fuse-link mounted in a carrier and fuse base.

Modern cartridge fuse-links contain fusible elements mounted in rigid housings of insulating material. The housings are filled with suitable exothermal and arc-quenching powders, such as silica, and they are sealed by metal end-caps which carry the conducting tags or end connections. A typical fuse-link is shown in *Figure 10.58*. The metal parts, other than the fusible elements, are invariably of copper, brass, steel or composites and they must be capable of operating under the exacting thermal, mechanical and electrical conditions which may arise in service. The materials used for the fusible elements must enable predictable performances to be obtained under a wide range of conditions, from normal thermal cycling to the violent changes of state that occur when elements are subjected to arcing during the interruption of faults.

10.3.1 Fuse technology

Fuses operate for long periods during which they carry currents at levels up to those associated with healthy conditions on the circuits protected by them and they must be capable of interrupting overcurrents up to the maximum levels possible when faults occur.

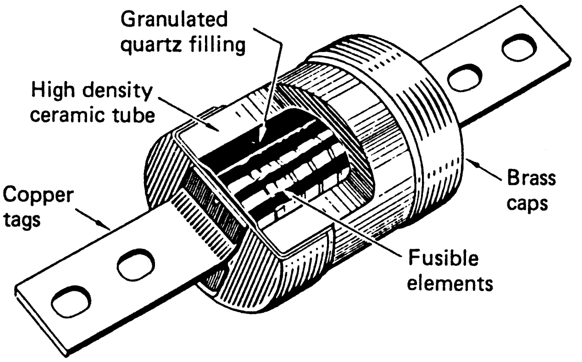


Figure 10.58 HRC fuse-link

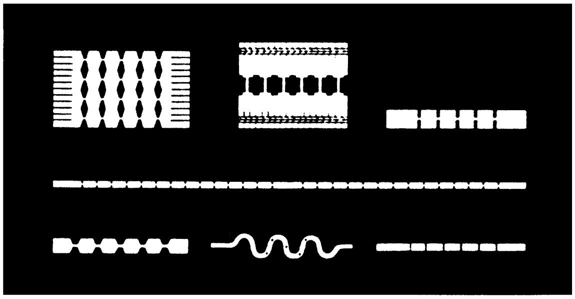


Figure 10.59 Shaped fuse elements

To satisfy these requirements, a fuse must be able to carry normal load currents and even transient overloads (and the thermal cycling which accompanies them) for a service life of at least 20 years, without any change of state that might affect its electrical performance. This property of 'non-deterioration' requires that the fusible element be both thermally and chemically compatible with the ambient media. It must also respond thermally to overcurrents by melting and subsequently interrupting its circuit.

The melting of an element is followed by a period of arcing during which the electrical energy input can be very high, its magnitude and the duration of arcing being dependent on the protected circuit. Successful fault interruption implies that the arcing is wholly contained within the fuse-link and the level at which this can be achieved is termed the breaking or rupturing capacity of the fuse-link. It must be recognised that unsuccessful interruption can result in disastrous changes to a circuit and its surroundings.

The operating time of a fuse-link varies inversely with the level of an overcurrent and discrimination is obtained in networks by choosing fuses with the necessary time/current characteristics and current ratings.

An important property of a high-breaking-capacity fuse-link is its ability to limit the energy fed to a fault, by melting and achieving arc extinction long before the fault current can rise to the levels which the circuit could otherwise produce, i.e. the 'prospective' values. This property is achieved by selecting the necessary element material and geometry.

10.3.2 Element materials

Silver, copper, tin, lead, zinc or alloys of these materials are used to produce fuse elements. In the past, silver was used in the majority of fuse-links but copper has been employed increasingly in recent years because of its much lower cost. Copper is now used in most industrial type fuse-links but silver is still required to obtain the performance required of fuse-links needed to protect semiconductor devices. The other materials listed above are used in fuse-links intended for low-power circuits.

Apart from its cost, silver is ideal for elements because of its physical properties. It is reasonably immune to corrosion in normal atmospheric conditions and is chemically compatible with silica and other media by which it is surrounded in fuse-links. Even when oxidation of an element does occur at elevated temperatures, the conductance is hardly affected because the conductivity of silver oxide is close to that of the parent metal. Silver is ductile and easily fashioned into the shapes needed, some of which are shown in *Figure 10.59*. It is easily joined or connected to other materials and is hard enough to be mechanically self-supporting. It can also be combined with other dissimilar materials (e.g. to produce the M effect referred to later) to produce eutectic alloys without affecting its stability during thermal cycling. The physical break-up which follows the melting of an element is regular and predictable for material of the prescribed purity. The vaporised metal can be made to disperse within the arc-quenching media to combine and condense so that the resulting 'fulgurite' becomes an insulator.

A silver element may be heated almost to melting and then allowed to cool without its state changing significantly. Such situations arise in service when large overcurrents are interrupted elsewhere in a circuit and it is important that the fuse-links involved should not, as a result, be weakened or change their designed time/current characteristics.

Whilst the properties of copper are not quite so good as those of silver they are, nevertheless, of standards which allow satisfactory elements to be produced for industrial fuse-links.

10.3.3 Filling materials

Cartridge fuse-links are invariably filled with granular quartz of high chemical purity the grain sizes being in the region of 300 μm. The filling material conducts heat energy away from fuse elements and, therefore, to obtain consistent performance it is necessary that a high and constant packing density be achieved. To further improve the consistency and raise the heat conduction it is now usual to employ inorganic binders in the filling material. This enables shorter operating times to be achieved and better performance is also obtained in d.c. applications.

10.3.4 Fuse-links with short operating times

All modern high-breaking-capacity fuse-links contain elements, usually with restrictions, of small cross-sectional areas connected between relatively massive end connections which act as heat sinks. To obtain rapid operation this principle is employed to a high degree. The extent to which the mass of the heat sink can be increased while reducing the length of the relatively thin element is determined by the requirement that the fuse should withstand the system voltage after the current has been interrupted (i.e. must not restrike). Considerable ingenuity has reconciled these two mutually incompatible requirements. More than one restriction may be used in series along the length of an element to cater for increased voltage, but this aggravates the problem of dissipating heat from the elements. The solution lies in an increase in the transfer of radial heat through the surrounding media. Thus the fuse element must not be looked at in isolation, but as a composite whole with the rest of the assembly.

The fashioning of fuse elements to produce elaborate shapes is economically limited by the means available to achieve the shape required. A variety of means is employed and these often influence the choice of material as regards its physical constants, e.g. the purity of silver as a factor in hardness, etc.

10.3.5 The M effect

The M effect, deriving from an exposition by Metcalf, refers to exploiting the thermal reactions of dissimilar metals in the control of time/current characteristics. The thermally most stable fuse element is a simple homogeneous metal. Such an element provides the highest degree of non-deterioration and reliability with adequate breaking capacity at higher overcurrents, but it may be insensitive at lower overcurrents. A lower melting temperature metal with higher resistivity and, therefore, greater thermal mass, can be made to respond more sensitively to lower overcurrents but may be unreliable at higher currents.

The M effect is a means by which these extremes can be combined to produce a desired characteristic, but it needs to be used with care in design to avoid compromising non-deterioration properties. An element incorporating the M effect is shown in *Figure 10.60*(*a*).

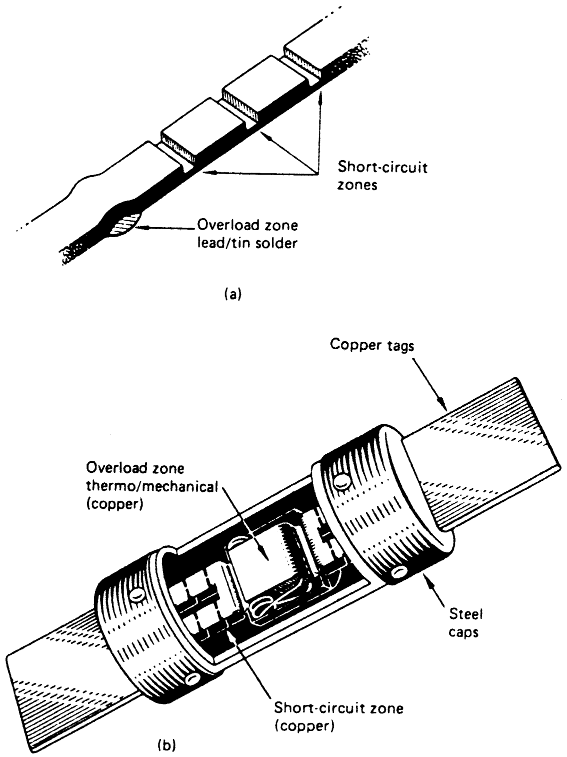

Short-circuit zones
Overload zone lead/tin solder
(a)
Copper tags
Overload zone thermo/mechanical (copper)
Steel caps
Short-circuit zone (copper)
(b)

Figure 10.60 (a) Fuse element (English Electric); (b) dual-element fuse.

10.3.6 Composite or dual-element fuses

Satisfactory operation throughout the overcurrent and short-circuit ranges is sometimes obtained in the same package by combining what are, in effect, two fuses connected in series in the same cartridge (*Figure 10.60*(*b*)). Typical of these is the so-called dual-element design common in the USA. The short-circuit zone is similar to the homogeneous element used in single-purpose HRC fuses. The overload zone may take the form of a massive slug of low-melting-point alloy, or some electromechanical device, e.g. two copper plates soldered together and stressed by a spring so that when the solder melts the plates spring apart to interrupt the current. The variables in such designs are considerable and many ingenious ideas have been exploited with some success.

10.4 Contacts

Contacts may be classified according to the load they control and are here discussed under four basic headings:

(1) low voltage, light current;
(2) low voltage, high current;
(3) medium voltage (<660 V) and power levels; and
(4) high voltage, high power.

An indication of the physical properties of contact materials and the performance and application of contact alloys is given in *Tables 10.4* and *10.5*.

10.4.1 Low-voltage, low-current contacts

These contacts are required to make and break a very low electrical duty so contact erosion is not a problem. Ideally the contacts should have a low contact resistance which does not introduce electrical noise by electrothermal or electromechanical means. Importance is therefore placed on surface contamination and deterioration in use and in storage. Silver–nickel alloys are commonly used in low-current control circuits, but for low noise a plated gold surface may be applied which survives a signal level service but rapidly exposes the silver–nickel for higher current applications. Rhodium or palladium may also be used in combination with gold for higher mechanical duty applications.

10.4.2 Low-voltage, high-current contacts

Separation of contacts carrying a current will create an arc. Since an arc requires a minimum voltage to maintain itself, the arc rapidly extinguishes for low voltage. However, inductance in the load will cause an arc of longer duration and results in contact burning. The high current requires a low contact resistance and large contacts with high contact force are used. Contact materials may be silver–nickel, though some automotive applications will use copper alloys for economy.

Contact resistance is a function of the contact materials, the force applied and, to some extent, the shape of the contacts. The resistance of a pair of contacts may be expressed as $r = k/f^n$, where f is the force applied, and k and n are constants depending on the contact materials and shape.

Typical values for f and n for copper are given in *Table 10.6*. At the instant of contact closure a single contact point

Table 10.4 Physical properties of contact materials

Material		*Density* $\left(\begin{array}{c}\mathrm{kg/m^3}\\ \times 10^{-3}\end{array}\right)$]	*Melting point** (°C)	*Boiling point*†⇐ (°C)	*Hardness* (HV)		*Tensile strength* (N/mm^2)		*Elongation* (%)		*Thermal conductivity at 20°*C (W/K-m)	*Electrical conductivity* (m/Ω-mm^2)
					Soft	*Hard*	*Soft*	*Hard*	*Soft*	*Hard*		
Pure metals												
Silver	Ag	10.5	961	2200	30	80	200	360	30	2	419	62
Gold	Au	19.3	1063	2370	25	60	140	240	30	1	297	44
Platinum	Pt	21.5	1769	4400	40	95	140	400	50		72	9.5
Palladium	Pd	12.0	1552	4000	40	100	200	480	44	2	72	9
Rhodium	Rh	12.4	1966	4500	130	280	420		9		88	22
Indium	Ir	22.5	2454	5300	220	350					59	19
Copper	Cu	8.9	1083	2300	50	100	200	450	33	2	394	58
Tungsten	W	19.1	3400	6000	250	450	1000	5000			167	18
Molybdenum	Mo	10.2	2610	5560			600	2500			142	19
Iron	Fe	7.9	1539	2740	90	150					75	10
Nickel	Ni	8.9	1453	2730	80	200	450	900	50	2	92	14
Power engineering materials‡⇐												
Fine grain silver	0.15% Ni	10.5	960	2200	55	100	220	360	25	1	415	58
Silver–copper	3% Cu	10.4	900	2200	65	120	250	470	25	1	372	52
(hard silver)	5% Cu	10.4	850	2200	70	125	270	550	20	1	335	51
	10% Cu	10.3	780	2200	75	130	280	550	15	1	335	50
	20% Cu	10.2	780	2200	85	150	320	650	15	1	335	49
Silver–nickel	10% Ni	10.3	961	2200	50	90	220	400	20	1		54
	15% Ni	10.2	961	2200	55	92	240	420	17	1		
	20% Ni	10.1	961	2200	60	95	280	450	15	1		47
	30% Ni	10.0	961	2200	65	105	330	530	8	1		42
	40% Ni	9.8	961	2200	70	115	370	580	6	1		37
Silver–cadmium oxide	10% CdO	10.2	961	2200	70	100	300	460				48
	15% CdO	10.1	961	2200	80	125						42
Silver–cadmium oxide	10% CdO	10.2	961	2200	50	80	230	450				48
	12% CdO	10.2	961	2200	60	95						47
	15% CdO	10.1	961	2200	65	115	280	420				45.5
Silver–zinc oxide	8% ZnO	10.2	961	2200	60							49
Silver–graphite	3% C	9.1	961	2200	40							47
	5% C	8.6	961	2200	40							43.5
Tungsten–silver	20% Ag	15.4	961	2200	180	240					245	26–28
	35% Ag	14.8	961	2200	100	130					280	34–36
	50% Ag	13.5	961	2200	900	100						40
	65% Ag		961	2200	80	90						
	80% Ag		961	2200	70	80						42
Tungsten–carbide-	20% Ag	13.3	961	2200	400	470						25–35
silver	60% Ag	11.2	961	2200	100	130						
Tungsten–copper	15% Cu	16.0	1083	2300	190	260						
	20% Cu	14.7	1083	2300	180	240						24–26
	25% Cu	14.3	1083	2300	170	220						
	30% Cu	13.1	1083	2300	160	200						28–30
	40% Cu	12.7	1083	2300	140	170						
	50% Cu	11.8	1083	2300	130	150						
Molybdenum–silver	25% Ag		961	2200	190	210						
	35% Ag	10.3	961	2200	160	180						22
	50% Ag	10.3	961	2200	130	150						24
Light current materials‡⇐												
Fine grain silver	0.15% Ni	10.5	960	2200	55	100	220	360	25	1	415	58
Silver–palladium	60% Pd	11.4	1330	2200	100	170	380	720			29.3	2.4
	50% Pd	11.2	1290	2200	90	160	360	700	15	1	34	3.3
	40% Pd	11.1	1225	2200	70	140	350	630	20	1	46	4.9
	30% Pd	10.9	1155	2200	65	120	300	600	20	1	59	6.7
Palladium–copper	15% Cu	11.4	1370	2300	100	260	400	800	42	1		2.5
	40% Cu	10.5	1200	2300	120	280	450	850	39	1	37.7	2.7
Gold–silver–copper	20% Ag	15.1	865	2200	125	230	480	820	20	1	68	7.1
	10% Cu											
	25% Ag	15.2	940	2200	90	185	400	700	25	2	67	7.1
	5% Cu											
Gold–silver–nickel	26% Ag	15.4	990	2200	80	120	350	570	20	1		8.3
Gold–nickel	5% Ni	18.2	995	2370	105	160	380	640	25	1	52	7.3
Gold–platinum	10% Pt	19.5	1100	2970	45	160	260	410	20	1	54.5	8
	25% Pt	19.9	1220	2970	80	155						3.6
Gold–silver	8% Ag	18.1	1035	2200	30	100	150	320	25	1	147	15.8
	20% Ag	16.5	1035	2200	35	90	190	390	25	1	75	10.5
	30% Ag	15.4	1025	2200	40	95	220	380	25	1		9.8
Gold–silver–platinum	25% Ag	16.0	1050	2200	30	150	300	500	25	1		6.1
	6% Pt											
Platinum–tungsten	5% W	21.4	1830	4400	160	250						2.3
Platinum–nickel	8.5% Ni	19.2	1670	2370	180	260						3.7

* For alloys the solidus point is given, for sintered materials the melting point of the lowest-melting component.

† The boiling point of the lowest-boiling component is given.

‡⇐The composition is given in weight %.

Table 10.5

Material	*Properties*	*Area of application*
Base: tungsten or molybdenum W/Ag 80/20 … 20/80	Very low wear, decreasing with increasing W content. High contact resistance, increasing with increasing W content. Resistance increases during life. High contact forces necessary. Bad arc mobility properties. Not workable.	Low voltage and high voltage circuit-breakers, miniature circuit-breakers (in particular American systems). Railway switches
WO/Ag 80/20 … 40/60	Slightly better than W/Ag for erosion. Suppression of forming of tungstate	As W/Ag, in particular for simple pairs of contacts (no special arcing contacts)
Mo/Ag 80/20 … 50/50	Similar to W/Ag	Similar to W/Ag
W/Cu 85/15 … 50/50	Similar properties to A/Ag, but more prone to forming of oxide	High voltage-load breaking switches and circuit-breakers (contacts in air, oil, SF_6); transformer tap-changers (contacts under oil). Electrodes for spark erosion, electrolytic removal and welding.
Base: silver		
Ag	Highest electrical and thermal conductivity. Oxidation resistant but formation of sulphide. Material transfers. Easily worked	Control switches, microswitches, regulators and selector switches: voltages $U >$ 60 V; currents $l < 10$ A
Ag/Ni 99.85/0.15 (fine-grain silver) Ag/Cu 97/3 … 90/10 (hard silver)	Similar to Ag, but lower erosion. Contact resistance increases with increasing base metal content. Welding tendency low for peak currents below 100 A	Control-, micro-, selector-, relax switches, regulators, miniature circuit-breakers with ratings up to 5 A
Ag/Ni 90/10 … 80/20	Contact resistance similar to hard silver, but less increase in resistance during life. Lower erosion. No welding for current peaks up to 100 A. Low and flat material transfer when switching d.c. Erosion debris on insulating materials non-conducting. Good arc extinguishing properties	Control switches, regulators, selector switches for d.c. and a.c. up to 100 A. Switches for domestic appliances. Miniature circuit-breakers up to 25 A rating. Motor control switches, contactors up to 25 A rating. Automative switches. M.c.b.s for d.c. and a.c. (unequal pairs with Ag/C). Controllers
Ag/Ni/ 70/30 … 60/40	Contact properties similar to 10–20% Ni, but higher contact resistance and lower wear (increasing with increasing Ni content)	Circuit-breakers for d.c. and a.c. Automative horn switches. Controllers
Ag/CdO 90/10 … 85/15	Contact resistance somewhat higher than for Ag/Ni 90/10. No welding up to peak currents of 3000 A. Low arc erosion in the range 100 to 3000 A. Very good arc extinction properties. Unfavourable arc movement properties. Limited workability	Low-voltage contactors, motor and motor-protection switches with ratings from 10 A. Low-voltage circuit-breakers with ratings up to about 100 A. Miniature circuit-breakers and earth leakage circuit-breakers with peak currents up to 3000 A. Lighting switches
Ag/C 97/3 … 95/5	Low contact resistance. Very high reliability against welding (increasing with increasing C content). Good friction properties. High wear. Bad arc-mobility properties. Bad workability	Miniature circuit-breakers and earth leakage circuit-breakers. Low voltage circuit-breakers (unequal pairs with Ag/Ni). Capacitor protective relays. Sliding contacts with self-lubrication
Ag/ZnO 92/8	Similar properties to Ag/CdO, but arc erosion in the current range 100–3000 A somewhat larger and in the range 3000–5000 A smaller	Low-voltage circuit-breakers with ratings up to 200 A. Earth leakage circuit-breakers
Base: palladium		
Pd	Highly resistant to corrosion, but prone to catalytic reaction with organic materials (brown powder). Highly resistant to arc erosion. Low electrical conductivity	Switching contacts at voltages $U =$ 20–60 V

cont'd

Table 10.5 *(continued)*

Material	*Properties*	*Area of application*
Ag/Pd 70/30 … 50/50	Generally resistant to corrosion, but worse than Au alloys. For Ag compared to Ag/Pd 70/30 about 7 times faster and compared with Ag/Pd 50/50 about 100 times faster formation of surface films. Highly wear resistant	Switching contacts at voltages $U = 20$–60 V, e.g. for telephone relays and selectors. Usual material in telecommunications. Sliding contact in precision potentiometers
Pd/Cu 85/15 and 60/40	Corrosion behaviour similar to Pd, but at 40% Cu thin oxide layers form high resistance to arc erosion. Low tendency to transfer	Switching contacts at voltages $U = 6$–60 V. High switching currents
Base: platinum		
Pt/W 95/5 Pt/Ni 91.5/8.5	Resistance to corrosion better than for Pd alloys, but also formation of 'brown powder'. Low, even transfer. Very highly wear resistant	Switching contacts at high load currents and very long life
Base: gold		
Au, Au/Pt 90/10	Highest resistance to corrosion, contact resistance constant over long periods. Prone to cold welding. Material transfer	Opening contacts for very small currents and voltages (dry circuits), e.g. in measuring devices, need switches
Au with hardening additives, electrolytically produced (hard gold)	Similar to Au, but slightly higher contact resistance and less prone to cold welding	Plugs, slide rails, rotary and sliding switches. PCB edge connections
Au/Ag 92/8 … 70/30 Au/Ag/Pt 69/25/6	Good resistance to corrosion. Higher hardness and resistance to wear and less prone to transfer than Au	Switching contacts with voltage <24 V and for small currents, e.g. circuits. Plugs for frequent operation
Au/Co 95/5	Resistance to corrosion similar to, hardness and wear resistance higher than Au/Ag alloys. Slight tendency to transfer. Less malleable	Switching contacts for long life, e.g. for flashers, measuring devices, clocks. Plugs with long life
Au/Ag/Cu 70/25/5, 70/20/10 Au/Ag/Ni 71/26/3	Good resistance to corrosion but slightly less than Au/Ag, decreasing with higher base metal content. Transfer worse than for Au/Ni and Au/Co	Switching contacts, e.g. telegraph relays at voltage < 24 V. Plugs for normal life at contact forces at about 0.5 N

Table 10.6 Constants for copper contract: $r = k/f^n$ *with* r in ohms and f in newtons

Form	*Surface condition*	*n*	*k*
Point	Normal	0.5	0.0007
Line	Normal	0.7	0.0015
Plane	Normal	1.0	0.004
(160 mm^2)	Lubricated	1.0	0.003
	Tinned	1.0	0.012
	Fine-ground, new	2.0	5

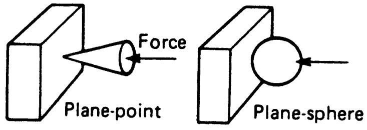
only may be considered; increasing force distorts the surface and more contact points are established in parallel with the first. With a 'soft' material and with a high force, contact is eventually established over an area which may represent a large proportion of the available contact area, resulting in a low contact resistance. This system would soon cause mechanical erosion of the contact surfaces and so such techniques are only applied to contacts which are only occasionally separated such as plugs and sockets or isolators. Sliding of the contacts may also be permitted to ensure clean surfaces are presented at each contact 'make'.

The contact shape has importance if the contacts are not expected to erode during their lifetime and thus change their original shape.

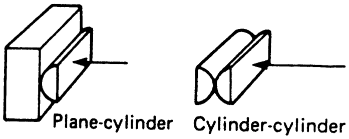
10.4.3 Contact design

The passage of current at the contacting face will cause heating and this may cause a local softening of the material with a resulting increase in contacting area and reduced contact resistance. This would appear to be an advantage but represents a dangerous condition since welding may occur. In the minimal case the contacts may be separated mechanically and the weld broken but in the worst case the contacts become permanently joined! Welding may also occur due to contacts arcing and particularly so when contacts 'bounce' while carrying current.

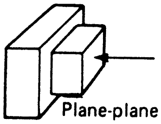
To discourage welding, contact 'alloys' are available which contain low resistance silver and a hard material such as nickel or tungsten or an oxide of cadmium, tin or zinc. Graphite may also be included to reduce welding but at the expense of increased contact resistance. Some forms of contact are shown in *Figure 10.61*.



(a) Point contacts



(b) Line contacts



(c) Plane contacts

Figure 10.61 Basic forms of contact

10.4.3.1 Medium voltage (up to 660 V)

Make and break contacts for this industrial range are required to provide a useful life of many thousands of operations. The duty may be a motor load where the starting current is typically six times the running current. For this duty (AC3), a long life can be expected since little erosion occurs at contact close. However, bounce can cause welding so contacts need to be rated according to the duty. Contacts which make and break equal currents (AC4) would erode more rapidly. At high currents shaped contacts may be used so that a defined area of the contact breaks last and carries the eroding arc. The remaining area, by staying clean, provides a low resistance for the continuous load current. The best combination of contact force, size, shape and material has to be made for a contactor to control a wide range of loads; capacitive loads (fluorescent lights) are particularly prone to weld problems. Silver–cadmium oxide is a common choice of contact material because of its good erosion and weld resistance properties. Silver–tin oxide is a recent alternative. The granular structure of these sintered materials has a great influence on their performance. The ratio of silver to cadmium/tin/nickel is high and thus acceptably low contact-resistance is obtained. However, the thermal rating of the contacts normally decides their dimensions.

10.4.3.2 High voltage–high power (≥ 3 kV)

This application is generally required for a low duty since the operating rate is slower. The higher voltages may require a multiple contact break and external influences are used to extend the arc into a shute where it is cooled and extinguished. Sulphur hexafluoride gas and oils may be used to assist in deionising the arc. (The special behaviour of an arc in a vacuum is used in the vacuum contactor which is capable of interrupting current at a non-current zero (current chopping).) Contact materials for which high-power duty have to withstand higher temperatures and tungsten alloys are common. In oil, copper is acceptable for the load-carrying part of the contact since oxidation does not occur and the arc tips only are fitted with tungsten alloy.

The effect of a current passing through closed contacts is represented by I^2rt. The contact resistance, r, is considered low and stable, but t is large so the resulting heat needs to be dispersed. This is performed by conduction into the mass of the supporting contact backing and, in many constructions, into the cable connections. During the contact make and break, r is larger but t is small. The heat is dispersed only into the contact area, so the thermal capacity and thermal conduction of the immediate contact assembly is most significant. Good ventilation also assists in reducing contact temperature.

When separating contacts which are passing a current, the final contact is small so the resultant resistance produces considerable heat and high temperatures are reached. $\theta = kV^2/\gamma\rho$, where V is the contact voltage, γ is the thermal conductivity and ρ is the electrical resistivity of the contact materials. V has a major influence and temperatures in excess of 10 000 K are readily attained. Such temperatures cause vaporisation, thermionic emission and, together with electromagnetic forces, these cause destruction of the contact surfaces evidenced as erosion.

When the contacts are subjected to an arc during circuit interruption, it is advantageous to minimise the duration of the arc. An a.c. circuit carries current which passes through zero twice per cycle and at this instant there is no energy to support the arc which then extinguishes. The rise in voltage across the contacts may restrike the arc if any ionisation remains. The contacts must therefore be separated to a sufficient distance and the gap cooled and ventilated. For a d.c. circuit, no such current zero occurs and the contact separation is required to break the arc. This may be assisted by multiple break contacts, extending the arc by magnetic fields, forcing the arc against cooling plates and generally by a very fast separation of the contacts. The use of blow-out coils to create a magnetic field is commonly used in large units but these are not effective at low currents and a permanent magnet may be added to assist the low current performance.

The life of contacts is a function of the current and the number of operations, there being both mechanical wear and electrical surface disruption. With d.c. there is a tendency for material to transfer from one contact to another in a unidirectional manner and advantage can be taken of this by the use of dissimilar contact materials or even dissimilar contact sizes. With a.c., the reversal of the arcing current will average the erosion so that the contacts erode equally. The unidirectional transfer by d.c. generates a 'pip and crater' condition which may reach the state where the 'pip' wedges into the crater effectively locking the contacts together. This effect is noticeable on small contacts which have too small a separation force to break the pip clear of the crater. Very low force contacts as in reed switches are particularly susceptible and gold contacts may even cold weld when left closed for a long period. Precautions are necessary to reduce these effects and for d.c. low current applications RC suppression and resistance to limit current pulses through the contacts is generally used.

10.5 Special alloys

Many alloys have been developed for special applications either at elevated temperatures for heating elements or as heat-resisting materials or at room temperature where a minimum change of resistance or dimensions is required.

10.5.1 Heating alloys

There is a considerable range of alloys used for heating elements for a wide range of applications including electric fires, storage heaters and industrial and laboratory furnaces. These alloys usually contain nickel together with chromium, copper and iron in varying proportions and often with small amounts of other elements. Similar alloys are also used for the construction of fixed and variable resistors. For heating elements, a considerable resistivity is required to limit the bulk of wire required. In addition, the temperature coefficient of resistivity should be small so that the current remains reasonably constant at constant applied voltage. *Table 10.7* gives the properties and trade names of a range of resistance heating alloys.

The operating temperature of these alloys is dependent on the cross-section of the wire or strip and on the atmosphere in which the material is to be used. The manufacturer's literature should be consulted before any application is finalised.

For higher temperatures, ceramic rods are used. Silicon carbide may be used in applications ranging from below 600°C up to 1600°C in either air or controlled atmospheres, although the type of atmosphere will determine the recommended element temperature. For even higher temperatures there are various cermets. Molybdenum disilicide ($MoSi_2$) with additions of a ceramic glass phase may be used up to 1900°C and a zirconia material (Zircothal) is used up to 2200°C. These maximum temperatures depend on the type of atmosphere in which they are to be used.

10.5.2 Resistance alloys

Alloys for standard and fixed resistors are required to have a low-temperature coefficient of resistivity in the region of room temperature.

Manganin (84% Cu, 4% Ni, 12% Mn) This has been the traditional material for high-grade standard resistors. Its resistivity is about 0.40 μΩ-m and its temperature coefficient is about 1×10^{-5}/°C.

Karma and Evanohm Trade names for quaternary alloys (73% Ni, 21% Cr, 2% Al, 2% Fe or Cu) which are being used increasingly for standard resistors, especially those of high value. The resistivity is about 1.30 μΩ-m and the temperature coefficient is $\pm 0.5 \times 10^{-5}$/°C. Each of the above alloys has a low thermo-e.m.f. against copper. Normally joining the above alloys to copper should be by argon arc welding or if this is not possible hard soldering may be used.

Constantan, Eureka Advance and Ferry Proprietary names for copper-nickel alloys (55% Cu, 45% Ni) which are used for heavy-duty and fixed resistors, potentiometers and strain gauges. They have a resistivity of about 0.50 μΩ-m and the temperature coefficient varies between $\pm 4 \times 10^{-5}$/°C. The high thermo-e.m.f. against copper (−40 μV/°C) is a disadvantage for d.c. resistors but the effect is usually negligible in a.c. resistors. These alloys may be soft soldered satisfactorily.

10.5.3 Controlled-expansion alloys

These give a range of thermal expansion required in precision parts, control devices and glass-to-metal seals. The lowest expansion alloy is a 36% nickel–iron alloy and is variously called Invar, Nilex or Nilo. The expansion coefficient of these alloys can be less than 1 ppm/°C although this can only be attained over a limited temperature range. Other alloys in the nickel–iron series with additions of cobalt or chromium can be tailored to give the same expansion coefficient as various types of glass for use as metal-to-glass or ceramic seals for television tubes, integrated circuits and fluorescent lights. The expansion coefficient of these alloys will be in the range 4–10 ppm/°C.

10.5.4 Heat-resisting alloys

A range of nickel–chromium based alloys has been specifically developed to meet strict limitations on the permissible creep of vital components in gas turbines in severe conditions of time, mechanical stress and working temperature. A 43/37/18/2% iron–nickel–chromium–silicon alloy is heat resisting in oxidising conditions up to 950°C or higher if the atmosphere is reducing. Developed originally for wire-woven conveyor belts for electric furnaces, it is now used also for a wide range of high-temperature applications.

Table 10.7 Resistance heating alloys*

	Nichrome 80 *Brightray S* *Nikrothal 80*	*Nichrome 60* *Brightray B* *Nikrothal 60*	*Nichrome 40* *Brightray F* *Nikrothal 40*	*Resistalloy 134*	*Alferon Y* *Fecralloy* *Kanthal AF*	*Kanthal APM* *Fecralloy 145* *Alferon 25*
Nominal composition (%)						
Ni	Balance	59	38	—	—	—
Cr	20	16	18	13/18	16	22
Fe	1	Balance	Balance	Balance	Balance	Balance
Y	—	—	—	—	0.4	—
Al	—	—	—	3.5/4.2	6	4.5 to 5.8
Si	1.5	0.35	2.2	—	—	—
Mn	0.4	1.0	0.5/1.2	—	—	—
Maximum cycling temperature (°C)	1150	1100	1050	1050	≤1375	≤1400
Resistivity (μΩ-m)						
At 20°C	1.08	1.12	1.06	1.25	1.37	1.30
At 1000°C	1.15	1.26	1.30	1.38	1.45	1.50

---

* These alloys are supplied by, Resistalloy, Sheffield (Fecralloy); Kanthal, Stoke On Trent (Nikrothal); Inca, Hereford (Brightray); and British Driver Harris, Stockport (Nichrome and Alferon).

10.6 Solders

Soldering is a process whereby metal components are joined together using a low-temperature filler metal, which is usually a tin-containing alloy. To assist in the wetting of the basis metal by molten solder, a flux, which is a weak acid, must be present to dissolve the thin oxide films already present on the surface of the components and to prevent further oxidation during heating of the joint.

10.6.1 Fluxes

Soldering fluxes are liquid or solid materials which, when heated, are capable of promoting or accelerating the wetting of metals by molten solder. Fluxes are usually divided into three groups by a classification based on the nature of their residues, namely corrosive, intermediate and non-corrosive fluxes. The National Standard for soft-soldering fluxes, BS 5625 (1980), incorporates a larger number of categories

Table 10.8 Rare and precious metals used for contacts

Metal or alloy	*Melting point* (°C)	*Vickers hardness (annealed)*	*Density* (kg/m^3)	*Resistivity at 20*°C (Ω-m $\times 10^8$)
Light-duty contacts				
Gold	1064	20	19 200	2.2
Platinum	1770	65	21 450	10.6
10% Iridium–platinum	1780	120	21 600	24.5
20% Iridium–platinum	1815	200	21 700	30.0
25% Iridium–platinum	1845	240	21 700	32.0
30% Iridium–platinum	1885	285	21 800	32.3
25% Iridium–ruthenium–platinum	1890	310	20 800	39.0
7% Platinum–silver–gold	1100	60	17 100	16.8
30% Silver–gold	1025	32	16 600	10.4
30% Silver–copper–gold	1014	95	14 400	14.0
10% Silver–copper–gold	861	160	13 700	12.5
Rhodium	1960	40	12 400	4.9
Iridium	2447	220	22 400	5.1
Palladium	1554	40	12 000	10.8
40% Silver–palladium	1290	95	11 900	35.8
40% Copper–palladium	1200	145	10 400	35.0
Medium–duty contacts				
10% Gold–silver	965	30	11 400	3.6
20% Palladium–silver	1070	55	10 700	10.1
10% Palladium–silver	1000	40	10 600	5.8
5% Palladium–silver	965	33	10 500	3.8
Fine silver	961	26	10 500	1.6
0.2% Magnesium–0.2% nickel–silver	961	140	10 400	2.8
1% Graphite–silver	961	40	9 900	1.8
2% Graphite–silver	961	40	9 700	2.0
Standard silver	778	56	10 300	1.9
10% Copper–silver	778	60	10 300	2.0
10% Cadmium oxide–silver	850	50	9 800	2.1
10% Nickel–silver	961	40	10 300	2.0
15% Cadmium oxide–silver	850	60	10 000	2.3
20% Copper–silver	778	85	10 200	2.1
20% Nickel–silver	961	48	10 100	2.1
Cadmium–copper–silver	800	65	10 100	4.2
50% Copper–silver	778	95	9 700	2.1
Heavy–duty contacts				
10% Cadmium oxide–silver	850	55	10 000	2.1
15% Cadmium oxide–silver	850	65	9 800	2.3
40% Tungsten carbide–silver	960	90	11 900	2.5
45% Tungsten carbide–silver	960	95	12 200	2.8
50% Tungsten–silver	960	125	13 600	2.8
50% Tungsten carbide–silver	960	160	12 500	3.0
55% Tungsten–silver	960	140	13 400	3.0
60% Tungsten carbide–silver	960	200	13 200	4.8
65% Tungsten–silver	960	185	14 800	3.3
73% Tungsten–silver	960	220	15 600	4.0
78% Tungsten–copper	1080	240	15 200	6.1
68% Tungsten–copper	1080	160	13 600	5.3
60% Tungsten–copper	1080	140	12 800	4.3

which gives an indication of the chemical nature of each flux type and their application.

For electrical components and other applications where corrosive residues could be difficult to remove, a non-corrosive rosin flux is used. Solder wire with a continuous core, or cores, of rosin flux can be used for manual soldering operations. The National Standard outlining the requirements of such material is BS 441 (1980).

10.6.2 Solder types

A selection of solder alloys are available which melt at temperatures ranging between 60 and 310 °C. British Standard grades of solder, their maximum levels of impurities that are permissible and typical applications are listed in BS 219 (1977). For the soldering of electrical connections and high-quality sheet metal work, an alloy containing 60% Sn, 40% Pb (grade K) is often used. For the machine soldering of electronic assemblies a solder of equivalent alloy composition but with a lower level of impurities is recommended (grade KP). Tin solder alloys with lower tin contents are used for general engineering and the joining of copper conductors and lead sheathing, etc.

The shear strengths of soldered joints are generally within the range 20–60 N/mm^2 at room temperature. As the temperature is increased the strength of joints made with tin–lead solders can decrease significantly. For this reason several solder alloy compositions, such as 95% Sn, 5% Sb (grade 95A) and 96.5% Sn, 3.5% Ag (grade 96S) are recommended for use at service temperatures in excess of 100°C. There are various methods of mechanically attaching two components prior to soldering in order to give added joint strength.

Depending on the soldering method employed solder can be used in the form of a bath of molten metal, sticks, solid wire, flux-cored wire, powder, solder creams, solder paint, or as preforms stamped out of thin foil. Solder creams, which are a mixture of oxide free solder powder, flux and rheology modifiers, can be pre-placed onto the area to be soldered by screen printing or syringe dispensing prior to heating. This technique is called 'reflow soldering' and is now widely used for the joining of surface mount electronic components to printed circuit boards.

10.7 Rare and precious metals

One of the most important and widespread electrical uses of the rare and precious metals is for contacts in applications, ranging from everyday electrical appliances to heavy-duty switchgear and contact-breakers, as well as in scientific and precision instruments, and communication equipment. Contacts can be broadly divided into light, medium and heavy duty and in *Table 10.8* the materials within these groups are approximately arranged in order of descending cost.

Light-duty contacts require that the surfaces do not corrode appreciably, so the more noble metals and alloys are often used, while currents are low, so that resistivity is less important.

Medium-duty contacts handle heavier currents, so that low resistivity is important and, since contact forces are normally high, slight corrosion/tarnishing is less important, but higher hardness becomes desirable.

For heavy-duty applications severe arcing and heavy mechanical wear must be expected, so that higher resistivity can be accepted, in the interests of high hardness and arc resistance.

The choice of contact material is very much a compromise between the intrinsic initial cost, the ease and cost of replacement and the electrical and mechanical properties of the alloy.

Platinum and rhodium–platinum alloys are extensively used for high-temperature thermocouples, which are accurate and particularly stable, as well as for the elements of high-temperature furnaces. Iridium–platinum, and rhodium–platinum are also used as electrodes in cathode-ray tubes. Caesium salts are used in the manufacture of photo-electric cells.

Pure silver is commonly used for electrical fuses and also in certain types of batteries and in capacitors, while a wide range of precious metals and alloys are used for thermal fuses acting as overtemperature cut-outs in electric furnaces.

A number of precious metal alloys are used for precision variable resistances, where the contact resistance at the wiper brush must be minimised. Precious metals are extensively used in integrated circuit technology where resistance to tarnishing and oxidation are important as well as the high ductility essential to the drawing of fine wire and thermocompression bonding. The primary material used in these applications is gold, although platinum, palladium, osmium and iridium are also used. Gold is employed in fine-wire interconnections in packaging of devices and as a eutectic alloy with silicon, used to bond the chip to the header. In thin- and thick-film circuits gold alloys with chromium, copper, nickel, platinum and silver are extensively used in terminations.

New developments in circuit technology requiring ever higher density interconnections utilise tape automated bonding (TAB) and flip-chip techniques. In these techniques wire interconnections have been replaced by direct soldered connections between solder wettable pads on the chip and the substrate. The development of 'solder bumps' used in the bonding involves a gold plating stage to protect the metallised aluminium pads prior to deposition of the high-temperature solders necessary in these techniques.

10.8 Temperature-sensitive bimetals

Temperature-sensitive bimetals, commonly known as thermostatic bimetals, are produced by bonding together two metals having different coefficients of expansion and cold rolling the composite into strip. When subjected to a temperature change, the strip alters curvature in a precise and calculable manner. The bimetals can be used in forms such as the deflection of a straight strip, the rotation of spirals or helices and the snap action of dished discs. Applications include temperature indicators, thermostatic controls, energy regulators, temperature compensation and automotive fuel control devices.

The alloys used for the low-expansion components are normally Invars, 36% or 42% nickel–iron. The high-expansion components are mainly alloys based on manganese, iron or nickel. Alloys have been developed for special applications, such as shower temperature control units and steam traps, where the corrosion resistance is specially important. A range of bimetals with closely controlled resistivities are available for devices, such as overload circuit-breakers, where the bimetal is heated by the direct passage of current. These include a number of trimetals in which the centre component is a low resistivity metal such as copper or nickel.

Table 10.9 Thermostatic bimetals

*Bimetal type**	*Deflection Constant, K* (°C^{-1})	*Range of maximum sensitivity* (°C)	*Modulus of elasticity* (Kg/mm^2)	*Electrical resistivity* (μΩ-m)
200	19.3×10^{-6}	−25–200	13 500	1.11
140	14×10^{-6}	0–175	16 000	0.76
400	11.8×10^{-6}	0–310	16 000	0.70
188†	8.8×10^{-6}	0–130	17 500	0.87
200R17‡⇐	18.9×10^{-6}	−25–200	13 500	0.16
R 5 M‡⇐	13.4×10^{-6}	−20–200	16 000	0.06

* Telcon Limited.
† Corrosion-resistant type.
‡ Trimetals.

A straight bimetal strip of length L, thickness t, and width w, fixed at one end and free to move at the other, will produce a free end deflection of $d = 1.1\,K\,(\Delta T)L^2/t$ for a temperature change ΔT.

The force developed, if the free end is restrained from moving, is 1.1 $KE(\Delta T)wt^2/4L$, where E is its modulus of elasticity. Similarly, for a bimetal spiral or helical coil of radius r, the angular deflection is 130 $KL/(\Delta t)/t$ and the restrained force is 0.19 $KE(\Delta T)wt^2/r$.

The properties of a representative range of bimetals are given in *Table 10.9*.

10.9 Nuclear-reactor materials

10.9.1 Introduction

A detailed description of the engineering design and mode of operation of the principal reactor types is given in Chapter 19 of this book. Reference to that chapter shows how varied is the detailed design of the various commercial types and how this variation requires large differences in the materials used, particularly within the reactor core. In this section, these diverse materials are identified and the reasons for their selection highlighted.

Before proceeding, however, it is worth noting that all the important commercial reactor types are ‘thermal’, i.e. neutrons have average energies of <0.1 MeV. Such reactors use materials in six common applications:

(1) as fuel;
(2) as fuel cladding;
(3) as pressurised coolant;
(4) as a moderator which reduces neutron energies to thermal values;
(5) as a pressure vessel which contains the high-pressure coolant; and
(6) as radiation shields.

There is continuing interest in the *fast-reactor system* so-called because it produces nuclear fission using energetic, i.e. ‘fast’, neutrons with energies of >0.8 MeV. Such reactors have some of the components of the present ‘thermal’ designs (e.g. fuel, cladding and coolant), but they do not require a moderator or a pressure vessel since the coolant does not operate at high pressure. Again, Chapter 19 provides further details. Although prototypical reactors have been in operation for many years (e.g. in the UK and France), a larger European fast reactor power station is currently being designed. Final choices of alloys to be used within this system have yet to be made in important areas, e.g. the fuel cladding, and so the information given below will refer to the UK prototype fast reactor (PFR) at Dounreay in Scotland.

Table 10.10 summarises the various materials used within the cores of the major commercial reactor types but also including the PFR.

10.9.2 Fuels

The only naturally occurring fuel is the isotope ^{235}U, present in natural uranium to about 0.7% . Uranium occurs as a complex silicate ore, chiefly pitchblende, which also contains lead, thorium, iron, calcium, radium, bismuth, antimony and zinc.

However, kasolite (essentially, a lead uranyl silicate) and carnotite ($K_2O{\cdot}2UO_3{\cdot}V_2O_5{\cdot}3H_2O$) are also sources of uranium.

There are various methods of uranium extraction which depend in detail on the type of ore and the impurities it contains. In general, the ore is crushed (e.g. by ball milling), in order to increase chemical reaction rates, and then dissolved in either sulphuric or nitric acid. Thus, as an example, boiling of the ore in concentrated nitric acid would produce a nitrate solution of uranium as well as the principal metal impurities. Of the latter, radium may be precipitated by the addition of barium sulphate whereas iron, lead and manganese are removed by adding sodium carbonate. Uranyl nitrate may be decomposed thermally to produce UO_3. This higher oxide may then be reduced partially to UO_2 by high-temperature exposure to hydrogen-bearing gases or may be reduced completely to uranium metal by fluorination to UF_4 and subsequent reduction by magnesium.

Most modern designs of commercial nuclear reactors use uranium fuel (as UO_2) enriched in the ^{235}U isotope (see, for example, *Table 10.10*). Such enrichment is achieved by further fluorination of UF_4 to UF_6 and then making use of the different gaseous diffusion rates of $^{235}UF_6$ and $^{238}UF_6$ to achieve partial separation of the isotopes. An alternative technique of separation relies on the different masses of the isotopes and their response to centrifugal action. Such physical (rather than chemical) processes of separation are needed because the chemical properties of the isotopes are identical.

Early commercial reactors in Europe, USA, USSR and Japan relied on natural uranium metal as the fuel, but this technology has now been superseded by the world-wide use of UO_2 pellets. This change allows higher fuel temperatures to be used and higher conversion rates of the ^{235}U fuel. Both fuel types are enclosed in metal cans in order to

Table 10.10 A summary of the principal materials within the core of the main commercial reactor types

Reactor type	*Fuel*	*Fuel cladding*	*Coolant*	*Moderator*
MAGNOX	U metal (natural)	Magnox Al-80 (Mg/0.8% Al)	CO_2	Graphite
AGR	UO_2 (2% enriched)	20 Cr/25 Ni steel	CO_2	Graphite
PWR	UO_2 (3% enriched)	Zircaloy-4 (Zr/1.5 Sn/0.10 Cr/0.20 Fe)	H_2O	H_2O
BWR	UO_2 (3% enriched)	Zircaloy-2 (Zr/1.5 Sn/0.10 Cr/0.20 Fe/0.05 Ni)	H_2O	H_2O
CANDU	UO_2 (natural)	Zircaloy-4	D_2O	D_2O
RBMK-1000	UO_2 (2% enriched)	Zr–Nb alloy	H_2O	Graphite
PFR	PuO_2/UO_2 (24–30% enriched)	Type 316 stainless steel	Liquid sodium	None

MAGNOX = ~~m~~agnox alloy clad/U-metal fuel system
AGR = ~~a~~dvanced gas-cooled reactor
PWR = ~~p~~ressurised water reactor
BWR = ~~b~~oiling water reactor
CANDU = ~~C~~anada uranium deuterium
RBMK-1000 = ~~R~~ussian hybrid design
PFR = ~~p~~rototype fast reactor

contain the products of the nuclear reaction and to provide mechanical support and stability.

^{235}U fissions on interaction with a neutron having energy in the thermal range, typically $\leqslant 0.1$ MeV—hence, its use in so-called 'thermal' reactors. By contrast, 'fast' reactors derive their name from the use of energetic neutrons (e.g. >1 MeV) to provide fission. ^{235}U cannot be used as a fuel in such cases and present fast-reactor designs employ ^{239}Pu as a fuel mixed as an oxide with UO_2.

10.9.3 Fuel cladding

The fuel cladding must provide structural support to the fuel stack and contain the fission products. Cladding materials must have a relatively low neutron absorption, have suitable low- and high-temperature strength and ductility, have good chemical compatibility with both the fuel and reactor coolant and have a resistance to property degradation as a result of neutron irradiation. Even though these requirements are particularly onerous, a number of cladding alloys have found service in commercial reactors.

The first generation of gas-cooled reactors used a magnesium–aluminium (Mg, 0.8% Al) cladding alloy, known as Magnox. This name has been adopted as the generic title of this reactor system. The alloy has a low melting temperature (ca. 923 K) and this, coupled with the use of uranium metal as fuel, means that only low T_2 temperatures can be used (ca. 685 K).

Accordingly, within the UK, the second generation of gas-cooled reactors (known as advanced gas-cooled reactors (AGRs)) made use of higher melting temperatures of both fuel and cladding: UO_2 for the former and 20% Cr, 25% Ni stainless steels for the latter. Although the latter material begins to melt at about 1630 K, normal operating temperatures will rarely exceed 1100 K, corresponding to a T_2 temperature of about 930 K. This relatively modest operating temperature, in relation to the solidus temperature, is dictated by the need to limit oxidation of the cladding in the CO_2 based coolant during normal operation and also to prevent melting under various postulated fault conditions. In addition, the conventional niobium stabilised steel cladding which is generally used, has relatively low creep strength at normal operating temperatures and this can result in mechanical interaction between the cladding and the UO_2 pellets during temperature and power changes. In the event that AGRs would ever be required to load follow, an alternative stainless-steel cladding, dispersion strengthened with titanium nitride particles, has been developed. The creep rate of this alloy at operating conditions is about a factor 10^3 less than that of the conventional alloy and, as a consequence, mechanical interaction with the pellet stack is much reduced.

However, the main development in thermal reactor systems world-wide has been to use pressurised water as coolant and a zirconium alloy as fuel cladding (*Table 10.10*). This alloy is nearly ideal under such conditions, having good neutron economy but also with high creep strength at operating temperatures (ca. 600 K) so that mechanical interaction with the fuel pellets is limited. Nevertheless, zirconium is chemically reactive both with the coolant and also with fission products contained within the fuel rod, particularly iodine. Operating constraints may be applied to limit such attack and the associated loss of load-bearing sections of the clad.

An important topical example is the potential problem of excessive waterside clad corrosion of zircaloy-4 in the pressurised water reactor (PWR) system at high power outputs or after extended dwell periods. Although zircaloy reacts readily with water at reactor temperatures to form ZrO_2, the rate of reaction decreases with time as a protective film of the oxide forms over the clad surface. For film thicknesses greater than about 2.5 μm, however, mechanical failure of the layer occurs and the corrosion rate increases. This sequence of events can repeat itself over the exposure period in reactor. This, in itself, is not a particular problem but the adherent surface oxide on the clad surface acts as a thermal insulation barrier which progressively raises the temperature of the oxide–metal interface as the oxide layer thickens. Obviously, the increase in temperature, for a given oxide thickness, is greater the greater is the heat flux, i.e. the greater is the power rating of the fuel. Since the corrosion rate appears to be determined by the temperature of the oxide–metal interface, a positive feedback is created which produces a progressive increase in corrosion rate with time.

Whereas all PWR systems can readily achieve their original design fuel burn-ups of 30 GW-d/te-U, waterside corrosion of zircaloy is now widely perceived to be the most economic threat to achieving significantly higher burn-ups, e.g. to 50 GW-d/te-U at useful power ratings. An associated issue is that a fraction, perhaps 10%, of the hydrogen released from water during the oxidation of the zirconium alloy enters the cladding and can produce mechanical embrittlement by precipitation of zirconium hydride platelets. This effect also is exacerbated at high burn-ups and, together with the issue of water-side corrosion, presents corrosion scientists with the challenge of improving the system's economics whilst still retaining high reliability and safety.

10.9.4 Coolant

Both the Magnox reactor system and the AGR use a pressurised CO_2 based coolant. Additions of CO, CH_4 and H_2O are made to this in a closely controlled manner to optimise gaseous reaction within the primary circuit. It is important to operate in a coolant which does not produce excessive oxidation of the graphite moderator since, in these reactor systems, the moderator also has a structural role and defines the location of fuel within the reactor core. This would be quite easily achieved by increasing the methane and/or carbon monoxide concentrations of the coolant but this also increases its carbon potential which could result in deposition of carbon on the fuel cladding and other core components. This is particularly important in the AGR since the steel cladding in this case contains both iron and nickel: elements which catalyse the deposition of carbon. Such deposits reduce the heat-transfer efficiency and will produce an increase in temperature of the cladding. Whereas, in practice, it is difficult to avoid some deposition, particularly in the AGRs, unacceptably bad behaviour can be prevented by good coolant control. Satisfactory current operation is achieved using a coolant with 1% CO.

Table 10.10 demonstrates that in the other principal thermal reactor designs the coolant is water, either light (H_2O) or heavy water (D_2O). Strict compositional control is also required in these cases. An excellent example of the balance that again needs to be struck between obtaining satisfactory fuel cladding behaviour without prejudicing other circuit or core components is offered by the PWR. In this case, the coolant is light water which contains up to 1200 volume parts per million (vpm) of boric acid which aids in reactivity control of the reactor. This addition lowers the pH of the water and produces an acidic solution which could dissolve iron rich components of the primary circuit. To minimise this, the coolant pH is raised (an ideal value would be to between 7.1 and 7.4) by the addition of an alkali. In nearly all Western plants, the alkali used is lithium hydroxide (LiOH), but its concentration tends to be limited to a maximum value of 2.2 ppm of lithium equivalent. This produces a coolant pH of 6.9, i.e. a little more acidic than the ideal. The reason for this is the wide-spread concern that higher lithium levels would increase the corrosion rate of the zircaloy cladding. This particular aspect is the subject of much world-wide investigation at the moment.

The generation of large quantities of heat in the relatively small core of a fast reactor necessitated the use of a liquid sodium coolant to achieve adequate heat transfer in the early conceptual designs of this system. This choice has been maintained in present prototypes. An advantage of such a coolant is that pressurisation is unnecessary but there are obvious disadvantages in the necessity to avoid water ingress from the secondary side of the heat-exchanger system and also in the need for separate heating circuits to avoid freezing during shutdown conditions. With recent advances in gas-circulator technology, it is now feasible to design a gas-cooled fast reactor using a pressurised CO_2 coolant.

10.9.5 Moderator

Thermal reactors require moderators to slow down fast neutrons to a sufficiently low energy to permit a fission reaction to occur with the ^{235}U isotope. In addition, the loss of neutrons within the moderator must be kept low in order that sufficient remain within the core for a chain reaction to proceed. Of the common moderator materials, *viz.* light water, heavy water and graphite, the last two provide good neutron economy and permit the use of unenriched fuel, e.g. in the Magnox and Canada Uranium Deuterium (CANDU) systems (*Table 10.10*); reactors moderated by light water, e.g. PWR and boiling water reactor (BWR), require some enrichment in the ^{235}U content of the fuel.

In the commercial gas-cooled reactors, graphite is used as a moderator. In practice, each pile consists of a vertical arrangement of fuel elements penetrating a large block of graphite fabricated from individual bricks. As has been pointed out earlier, the moderator in these cases has a structural role also and its oxidation in the CO_2 coolant is controlled by additions of CO, CH_4 and H_2O. Although the Magnox reactors operate with unenriched fuel, the AGR fuel pellets need to contain about 2% ^{235}U to compensate for neutron losses within the steel components of the core (e.g. the fuel cladding is an advanced stainless steel).

The use of heavy water as a moderator has been pioneered by the Canadians in their CANDU system. In this design (see Chapter 19), the moderator 'block' is termed a *calandria* and is a large stainless-steel vessel holding the heavy water at atmospheric pressure. Some 400 horizontal pipes or 'channels' exist within this block and these contain individual fuel elements enclosed in a pressure tube. The pressurised coolant which passes through each pressure tube is also heavy water and contributes to the overall moderation. Again it is possible to use unenriched fuel (as UO_2 in this case).

Light-water reactors, as their name implies, use light water as both coolant and moderator. However, there is no physical segregation of function, the coolant/moderator being allowed to flow freely through the reactor core. This leads to a particularly simple design and relatively small structures. Nevertheless, enriched fuel needs to be used (typically, up to 4% ^{235}U) to compensate for poor moderator economy.

A hybrid design has been developed on a large scale in the USSR. This, identified as the RBMK-1000 type in *Table 10.10*, uses light water as coolant but with graphite moderation. This combination of coolant and moderator can result in large positive void coefficients of reactivity; that is, an increase in core reactivity will occur following a reduction in coolant flow or increase in coolant temperature. For this particular design, such an increase in reactivity also produces an increase in reactor power, particularly at low power levels. The combination makes control difficult and was a contributing factor to the accident which befell this reactor type at Chernobyl.

10.9.6 Pressure vessel

All thermal reactors use a pressurised coolant so that the integrity of the primary system is a principal factor in reactor operation and safety. Most designs rely on a single, large pressure vessel but pressure-tube reactors enclose each fuel channel in a miniature cylindrical vessel or pressure tube.

The most widespread material for the construction of large pressure vessels is mild or low-alloy steel, manufactured to a wall thickness which varies with reactor type, e.g. 120 mm in the early Magnox stations to greater than 200 mm for PWR. A problem with such materials is the need always to demonstrate that brittle fracture of the vessel will not occur, particularly since neutron irradiation may itself tend to cause embrittlement. Only relatively simple cylindrical shapes are used for these vessels and contain a minimum of through-wall penetrations. As a consequence, the pressurised primary circuit extends beyond this pressure boundary and so similarly high standards of integrity need to be applied to the other components of this circuit, e.g. pipework, the recirculation pumps, steam drums, steam generators.

These difficulties were overcome in the later Magnox designs and in the AGR stations by the use of prestressed concrete vessels. These tend to be very large, e.g. 30 m high and 26 m in diameter, and contain the primary circuit components. The obvious disadvantages are the need for large on-site construction facilities, subsequent limited access and the requirement to use internal thermal insulation (based on stainless-steel-foil packages) to limit the temperature rise of the vessel.

A quite different approach is offered by the pressure tube reactors, e.g. CANDU and RBMK-1000. Here, the pressurised circuit is split into numerous small vessels or pressure tubes, each of which contains a single fuel channel. The tube wall is of a zirconium alloy (e.g. Zr, 2.5% Nb) which is separated by a gaseous (typically nitrogen) thermal insulation gap from the walls of the calandria in the CANDU system or is located within a graphite ring in the RBMK-1000 system. The obvious advantage is that the failure of a single pressure tube will not prejudice the integrity of the reactor, provided that such failure cannot propagate from one channel to another, e.g. by the release of high-energy debris. It is for this reason that much attention is paid to the corrosion rate of the inner surface of the pressure tube and the associated pick up of deuterium or hydrogen, since subsequent solid-state reaction with zirconium can lead to the precipitation of brittle intermetallic compounds.

10.9.7 Shield

The reactor shield must attenuate neutrons and γ-radiation. In most cases, the cheapest material is heavy concrete such as barytes concrete in which crude barium sulphate replaces the normal aggregate.

10.10 Amorphous materials

For many years soft magnetic alloys have been made by a casting, hot rolling and cold rolling regime. Development of such materials has concentrated on the reduction of impurities and the promotion of appropriate grain growth and metallurgical structure. More recently, a class of materials has emerged which uses quite different principles. Amorphous metals have no definite crystal structure and are cast from a melt in such a way that cooling to final thickness is very rapid (of the order of 10^6 °C/s). By an appropriate choice of composition which will include a ferromagnetic metal and 'glass forming' elements, plus the rapid cool, a structure is produced which is 'amorphous' in that no definite crystals are formed (as confirmed by X-ray crystal analysis). In such a structure, domain walls experience no lower energy in one position than another and so can move very freely to effect changes in magnetisation.

To achieve such very rapid cooling the material has necessarily to be very thin (of the order of 20–50 μm) and of quite high electrical resistance (of the order of 1.00–1.5 μΩ-m) due to the presence of glass formers such as boron. The combination of freely movable domain walls, high resistivity and low thickness combine to give a material of very low power loss.

Potential application areas for this new class of material can be broadly divided into two categories: power transformers, and electronic/high-frequency devices.

Table 10.11 Comparison of commercial grain-oriented silicon–iron and amorphous material

	3% Silicon–iron 0.3 mm thick	*Iron based amorphous alloy Metglas 2605S-2 0.33 mm thick*	*Consolidated amorphous strip POWERCORE 0.13 mm thick*
Curie temperature (°C)	745	415	415
Crystallisation temperature (°C)	—	550	540
Maximum working temperature (°C)	≈650	150	125
Tensile strength (MPa)	320–360	≥1500–2000 (as cast)	—
Yield strength (MPa)	300–320	1500–2000 (as cast)	—
Resistivity (μΩ-m)	0.45–0.48	1.37	1.37
Laminations factor (%)	95–98	80	90
Loss at 1.2 T, 50 Hz (W/kg)	0.64	0.11	0.12
Loss at 1.5 T, 50 Hz (W/kg)	0.83	0.27	0.28
Specific apparent power at 1.3 T (V-A/kg)	0.69	0.54	0.25
Specific apparent power at 1.5 T (V-A/kg)	0.94	2.33	1.3
Coercive force (A/m)	6.4	1.6–2.5	2.0
Saturation induction (T)	2.03	1.56	1.56

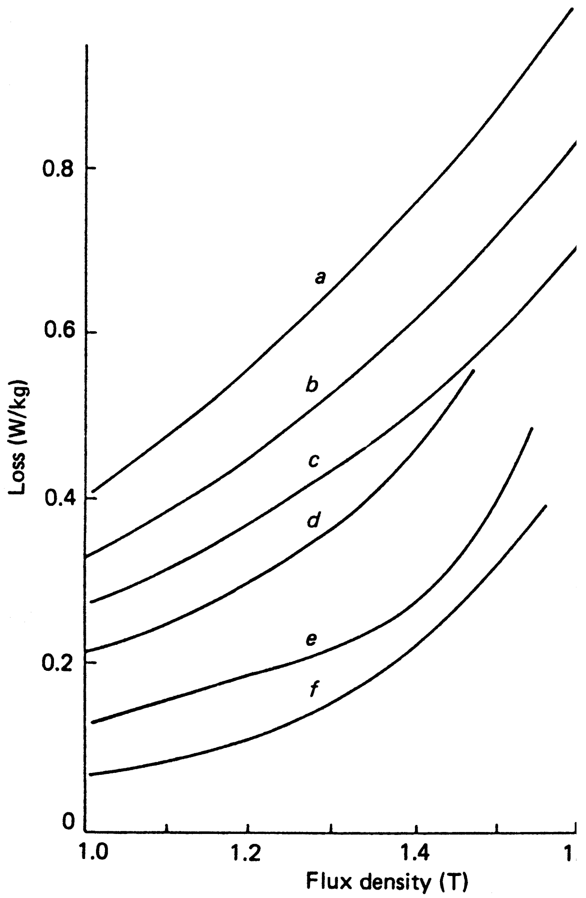


Figure 10.62 Variation of loss with flux density in various materials. (a) 0.3 mm conventional grain oriented; (b) 0.3 mm high permeability; (c) 0.23 mm domain refined; (d) 6.5% semicrystalline silicon–iron; (e) POWERCORE; (f) Metglas 2605S-2

10.10.1 Power-transformer materials

The presence of a high alloy content means that the saturation induction of amorphous materials is a lot lower than for silicon steels and operating inductions are usually confined to 1.5 T and below. At such inductions, power losses are of the order of one-third of those for grain-oriented electrical steels.

The mechanical properties of amorphous materials are less convenient. Amorphous materials are hard, brittle and difficult to cut accurately. Also, amorphous metals are very stress sensitive, so that cores must be built very carefully to avoid the degradation of magnetic properties which can follow from stress.

The production of power-transformer cores from amorphous metals is still at an early stage of development; however, transformers with wound cores designed to use thin amorphous ribbon directly are under widespread evaluation in the USA. In Europe where three-phase cut cores are more common, the employment of amorphous cores is more difficult.

Allied Signal Inc. in the USA has produced a multi-layered version of amorphous ribbon in which six or more layers of basic ribbon are bonded to give a composite strip which can be cut and stacked into three-phase cores. This material is sold by Allied Signal Inc. as 'Powercore'.

Magnetic and physical properties of a grade of amorphous foil material (Allied Signal Metglas 2605S-2) and of Powercore are given in *Table 10.11*.[1] *Figure 10.62* compares the loss versus induction performance of amorphous and cold-rolled materials.[1]

The future of amorphous materials in the electrotechnical world will depend on the equilibrium which may be reached between attractively low power losses, physical difficulty of employment, stress sensitivity, reduced saturation induction and cost.

Any trend making for increased energy costs and a drive towards the conservation of energy is likely to favour amorphous materials for at least some applications.

10.10.2 High-frequency-device materials

The amorphous alloys developed to compete with silicon iron for the large power transformer market have compositions which are based on relatively cheap elements, i.e. iron, boron and silicon. These alloys have a high magnetostriction which inhibits their ability to develop high permeabilities and low coercive force. It was therefore necessary to develop a different type of alloy based on cobalt or nickel, with a low magnetostriction, to enable the amorphous materials to extend their application potential. The typical properties of these alloys are detailed in *Table 10.12*.

Table 10.12 Properties of low magnetostriction amorphous alloys*

Material	$Fe_{40}Ni_{40}(MoBSi)_{20}$	$(CoFe)_{70}(MoBSi)_{30}$
Magnetostriction ($\times 10^{-6}$ m/m)	8	0.3
Coercive force (A/m)	3	0.5
Saturation flux density (T)	0.8	0.55
Permeability at 0.4 A/m	25 000	250 000
Maximum permeability	200 000	500 000
Loss at 0.2 T, 100 kHz (W/kg)	65	35
Resistivity (μΩ-m)	1.3	1.3
Curie temperature (°C)	250	250
Crystallisation temperature (°C)	450	450
Hardness (VPN)	800	1000
Maximum operating temperature (°C)	120	80

*Alloys of this type are supplied by Vacuumschmelze, Germany and Allied Signal Inc., USA.
VPN = Vickers pyramid hardness.

Important characteristics to be noted are their low losses at high frequency, low coercive force, high permeability, and very high hardness. Hitherto, due to their expense and limited forms of supply, the commercial application of high-frequency-devices has been restricted to areas where these properties can be used to best advantage. These include inductive components for switched-mode power supplies, pulse transformers, transductors, electromechanical sensors and tape-recorder heads where their high mechanical hardness provides excellent wear resistance.

Great care has to be taken with amorphous alloys not to exceed the maximum operating temperature recommended by the manufacturers since the amorphous structure is unstable at elevated temperatures and can cause the alloy to crystallise. The magnetic performance of the crystalline alloy is substantially inferior to that of the amorphous form. It has recently been discovered, however, that if the crystal size is restricted to nanometres, in an alloy of composition $Fe_{73.5}Cu_1Nb_3\ Si_{16.5}B_6$, low field permeabilities of 10^4 can be obtained. These nanocrystalline alloys could find wider application than the zero magnetostriction amorphous alloys due to their low basic raw-material cost.

Reference

1 MOSES, A. J. *IEE Proceedings*, **137**, 233 (1990)

Section C
Control

11

Electrical Measurement

M J Cunningham MSc, PhD, MIEE, CEng
University of Manchester

G L Bibby BSc, CEng, MIEE
Formerly University of Leeds

Contents

11.1 Introduction 11/3

11.2 Terminology 11/3

11.3 The role of measurement traceability in product quality 11/3

11.4 National and international measurement standards 11/3
  11.4.1 Establishment of the major standards 11/3

11.5 Direct-acting analogue measuring instruments 11/4
  11.5.1 Direct-acting indicators 11/4
  11.5.2 Direct voltage and current 11/5
  11.5.3 Alternating voltage and current 11/6
  11.5.4 Medium and high direct and alternating voltage 11/7
  11.5.5 Power 11/7
  11.5.6 Maximum alternating current 11/8
  11.5.7 Power factor 11/8
  11.5.8 Phase sequence and synchronism 11/8
  11.5.9 Frequency 11/8

11.6 Integrating (energy) metering 11/9
  11.6.1 Single-phase meter 11/9

11.7 Electronic instrumentation 11/10
  11.7.1 Digital voltmeters 11/10
  11.7.2 Digital wattmeters 11/16
  11.7.3 Energy meters 11/16
  11.7.4 Signal generators 11/16
  11.7.5 Electronic analysers 11/16
  11.7.6 Data loggers 11/17

11.8 Oscilloscopes 11/17
  11.8.1 Cathode-ray tube 11/17
  11.8.2 Deflection amplifiers (analogue c.r.o.) 11/19
  11.8.3 Instrument selection 11/20
  11.8.4 Operational use 11/21
  11.8.5 Calibration 11/21
  11.8.6 Applications 11/21
  11.8.7 Digital storage oscilloscopes 11/22
  11.8.8 Digital oscilloscope characteristics 11/22

11.9 Potentiometers and bridges 11/23
  11.9.1 D.c. potentiometers 11/23
  11.9.2 A.c. potentiometers 11/26
  11.9.3 D.c. bridge networks 11/27
  11.9.4 A.c. bridge networks 11/28

11.10 Measuring and protection transformers 11/32
  11.10.1 Current transformers 11/32
  11.10.2 Voltage transformers 11/33

11.11 Magnetic measurements 11/34
  11.11.1 Instruments 11/35
  11.11.2 Magnetic parameters 11/35
  11.11.3 Bridge methods 11/36

11.12 Transducers 11/36
  11.12.1 Resistive transducers for temperature measurement 11/37
  11.12.2 Thermistors 11/38
  11.12.3 p–n Junctions 11/38
  11.12.4 Pyrometers 11/38
  11.12.5 Pressure 11/38
  11.12.6 Acceleration 11/38
  11.12.7 Strain gauges 11/39
  11.12.8 Magnetostrictive transducers 11/40
  11.12.9 Reactance sensors 11/40
  11.12.10 Stroboscope 11/41
  11.12.11 Photosensors 11/41
  11.12.12 Nuclear radiation sensors 11/42

11.13 Data recording 11/42

11.1 Introduction

With increased interest in quality in manufacturing industry, the judicious selection of measuring instruments and the ability to demonstrate the acceptability of the measurements made are more important than ever. With this in mind, a range of instruments is described and instrument specification and accuracy are considered.

11.2 Terminology

The vocabulary of this subject is often a source of confusion and so the internationally agreed definitions[1] of some important terms are given.

Measurement: The set of operations having the object of determining the value of a quantity.

Measurand: A quantity subject to measurement.

Metrology: The field of knowledge concerned with measurement. This term covers all aspects both theoretical and practical with reference to measurement, whatever their level of accuracy, and in whatever field of science or technology they occur.

Accuracy: The closeness of agreement between the result of a measurement and the true value of the measurand.

Systematic error: A component of the error of measurement which, in the course of a number of measurements of the same measurand, remains constant or varies in a predictable way.

Correction: The value which, added algebraically to the uncorrected result of a measurement, compensates for an assumed systematic error.

Random error: A component of the error of measurement which, in the course of a number of measurements of the same measurand, varies in an unpredictable way.

Uncertainty of measurement: An estimate characterising the range of values within which the true value of a measurand lies.

Discrimination: The ability of a measuring instrument to respond to small changes in the value of the stimulus.

Traceability: The property of the result of a measurement whereby it can be related to appropriate standards, generally international or national standards, through an unbroken chain of comparisons.

Calibration: The set of operations which establish, under specified conditions, the relationship between values indicated by a measuring instrument or measuring system and the corresponding known value of a measurand.

Adjustment: The operation intended to bring a measuring instrument into a state of performance and freedom from bias suitable for its use.

Influence quantity: A quantity which is not the subject of the measurement but which influences the value of a measurand or the indication of the measuring instrument.

11.3 The role of measurement traceability in product quality

Many organisations are adopting quality systems such as ISO 9000 (BS 5750). Such a quality system aims to cover all aspects of an organisation's activities from design through production to sales. A key aspect of such a quality system is a formal method for assuring that the measurements used in production and testing are acceptably accurate. This is achieved by requiring the organisation to demonstrate the traceability of measurements made. This involves relating the readings of instruments used to national standards, in the UK usually at the National Physical Laboratory, by means of a small number of steps. These steps would usually be the organisation's Calibration or Quality Department and a specialist calibration laboratory. In the UK, the suitability of such laboratories is carefully monitored by NAMAS. A calibrating laboratory can be given NAMAS accreditation in some fields of measurement and at some levels of accuracy if it can demonstrate acceptable performance. The traceability chain from production to National Measurement Laboratory can, therefore, be established and can be documented. This allows products to be purchased and used or built into further systems from organisations accredited to ISO 9000 with the confidence that the various products are compatible since all measurements used in manufacture and testing are traceable to the same national and international measurement standards.

11.4 National and international measurement standards

At the top of the traceability chain are the SI base units. Definitions of these base units are given in Chapter 1. The SI base units are metre, kilogram, second, ampere, kelvin, candela and mole. Traceability for these quantities involves comparison with the unit through the chain of comparisons. An example for voltage is shown in *Figure 11.1*.

The measurement standards for all other quantities are, in principle, derived from the base units. Traceability is, therefore, to the maintained measurement standard in a country. Since establishing the measurement standards of the base and derived quantities is a major undertaking at the level of uncertainty demanded by end users, this is achieved by international cooperation and the International Conference of Weights and Measures (CIPM) and the International Bureau of Weights and Measures (BIPM) have an important role to play. European cooperation between national measurement institutes is facilitated by Euromet.

11.4.1 Establishment of the major standards

11.4.1.1 Kilogram

The kilogram is the mass of a particular object, the International Prototype Kilogram kept in France. All other masses are related to the kilogram by comparison.

11.4.1.2 Metre

The definition of the metre incorporates the agreed value for the speed of light ($c = 299\,792\,458$ m/s). This speed does not have to be measured. The metre is not in practice established by the time-of-flight experiment implied in the definition, but by using given wavelengths of suitable laser radiation stated in the small print of the definition.

11.4.1.3 Second

The second is established by setting up a caesium clock apparatus. This apparatus enables a man-made oscillator to be locked onto a frequency inherent in the nature of the caesium atom, thus eliminating the inperfections of the man-made oscillator, such as temperature sensitivity and ageing.

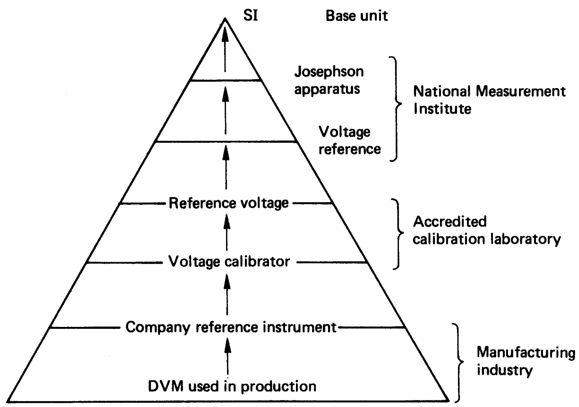


Figure 11.1 The traceability chain. DVM, digital voltmeter

Using this apparatus, the second can be established extremely close in size to that implied in the definition. Time and frequency can be propagated by radio signals and so calibration of, for example, a frequency counter consists in displaying the frequency of a stable frequency transmission such as BBC Radio 4 on 198 kHz. Commercial off-air frequency standards are available to facilitate such calibration.

11.4.1.4 Voltage

Work by Kibble[2] and others has led to the value of 483 597.9 GHz/V being ascribed to the height of the voltage step which is seen when a Josephson junction is irradiated with electromagnetic energy. A Josephson junction is a device formed by the separation of two superconductors by a very small gap. The voltage produced by such a junction is independent of temperature, age, materials used and apparently any other influencing quantity. It is therefore an excellent, if somewhat inconvenient, reference voltage. It is possible to cascade thousands of devices on one substrate to obtain volt level outputs at convenient frequencies. Industrial voltage measurements can be related to the Josephson voltage by the traceability hierachy.

11.4.1.5 Impedance

Since 1990 measurements of electrical resistance have been traceable to the quantised Hall effect at NPL.[3] Capacitance and inductance standards are realised from DC standards using a chain of AC and DC bridges. Impedance measurements in industry can be related through the traceability hierarchy to these maintained measurement standards. Research is in progress at NPL and elsewhere aimed at using the AC quantum hall effect for the direct realisation of impedance standards.

11.4.1.6 Kelvin

The fundamental fixed point is the triple point of water, to which is assigned the thermodynamic temperature of 273.16 K exactly; on this scale the temperature of the ice point is 273.15 K (or 0°C). Platinum resistance and thermocouple thermometers are the primary instruments up to 1336 K (gold point), above which the standard optical pyrometer is used to about 4000 K.

11.5 Direct-acting analogue measuring instruments

The principal specification for the characteristics and testing of low-frequency instruments is IEC 51 (BS 89). The approach of this standard is to specify intrinsic errors under closely defined reference conditions. In addition, variations are stated which are errors occurring when an influencing quantity such as temperature is changed from the reference condition. The operating errors can, therefore, be much greater than the intrinsic error which is likely to be a best-case value.

Instruments are described by accuracy class. The class indices are 0.05, 0.1, 0.2, 0.3, 0.5, 1, 1.5, 2, 2.5, 3 and 5 for ammeters and voltmeters. Accuracy class is the limit of the percentage intrinsic error relative to a stated value, usually the full-scale deflection. For example, for a class index of 0.05, the limits of the intrinsic error are ±~~0~~.05% of the full-scale value. It must be stressed that such performance is only achieved if the instrument is at the reference conditions and almost always extra errors in the form of variations must be taken into account.

11.5.1 Direct-acting indicators

These can be described in terms of the dominating torque-production effects:

(1) *Electromagnetic torque*: moving-coil, moving-iron, induction and electrodynamic (dynamometer).
(2) *Electric torque*: electrostatic voltmeter.
(3) *Electrothermal torque*: maximum demand.

A comparison is made in *Table 11.1*, which lists types and applications.

11.5.1.1 Torque effects

Instrument dynamics is discussed later. The relevant instantaneous torques are: *driving* torque, produced by means of energy drawn from the network being monitored; *acceleration* torque; *damping and friction* torque, by air dashpot or eddy current reaction; *restoring* torque, due usually to spring action, but occasionally to gravity or opposing magnetic field.

The higher the driving torque the better, in general, are the design and sensitivity; but high driving torque is usually associated with movements having large mass and inertia. The torque/mass ratio is one indication of relative performance, if associated with low power demand. For small instruments the torque is 0.001–0.005 mN-m for full-scale deflection; for 10–20 cm scales it is 0.05–0.1 mN-m, the higher figures being for induction and the lower for electrostatic instruments.

Friction torque, always a source of error, is due to imperfections in pivots and jewel bearings. Increasing use is now made of taut-band suspensions (*Figure 11.2*); this eliminates pivot friction and also replaces control springs. High-sensitivity moving-coil movements require only 0.005 mN-m for a 15 cm scale length: for moving-iron movements the torques are similar to those for the conventional pivoted instruments.

11.5.1.2 Scale shapes

The moving-coil instrument has a linear scale owing to the constant energy of the permanent magnet providing one of the two 'force' elements. All other classes of indicators are

Table 11.1 Comparison of instruments

Measurement parameter	*Type*	*Advantages*	*Disadvantages*
Direct voltage and current	Moving coil	Accurate: wide range	Power loss 50–100 mW
	Induced moving-magnet	Cheap	Inaccurate; high power loss
Alternating voltage and current	Moving iron	Cheap; reasonable r.m.s. accuracy	No low range; high power loss
	Electrodynamic	Adequate a.c./d.c. comparator; more accurate r.m.s.	Expensive; high power loss; square-law scale
	Induction	Long scale	Single frequency; high power loss; inaccurate
	Moving-coil rectifier	Low power loss (a few mW); wide range; usable up to audio frequency	Non-sinusoidal waveform error
	Moving-coil thermocouple	True r.m.s.; good a.c./d.c. comparator; usable up to radiofrequency	No overload safety factor; slow response; power loss (e.g. 1 W)
Direct and alternating voltage	Electrostatic	Negligible power; good a.c./d.c. comparator; accurate r.m.s.; wide frequency range	High capacitance; only for medium and high voltage; poor damping
A.c. power (active)	Electrodynamic	Accurate; linear scale	Expensive; high power loss
	Thermocouple electronic	High frequency; pulse power; true r.m.s.	Expensive
	Induction	Cheap; long scale	Single frequency; high power loss
A.c. power (reactive)	Electrodynamic	Accurate; linear scale	Power loss; needs phase-rotation network
A.c. maximum current	Thermal maximum demand	Long adjustable time constant; thermal integration; cheap	Relatively inaccurate

inherently of the double-energy type, giving a square-law scale of the linear property (voltage or current) being measured; for a wattmeter, the scale is linear for the average scalar product of voltage and current. The rectifier instrument, although an a.c. instrument, has a scale which is usually linear, as it depends on the moving-coil characteristic and the rectifier effect is usually negligible. (In low-range voltmeters, the rectifier has some effect and the scale shape is between linear and square law.) Thermal ammeters and electrodynamic voltmeters and ammeters usually have a true square-law scale, as any scale shaping requires extra torque and it is already low in these types. Moving-iron instruments are easily designed with high torque, and scale shaping is almost always carried out in order to approach a linear scale. In some cases the scale is actually contracted at the top in order to give an indication of overloads that would otherwise be off-scale. The best moving-iron scale shape is contracted only for about the initial 10% and is then nearly linear. Logarithmic scales have the advantage of equal percentage accuracy over all the scale, but they are difficult to read, owing to sudden changes in values of adjacent scale divisions. Logarithmic scales are unusual in switchboard instruments, but they are sometimes found in portable instruments, such as self-contained ohmmeters.

11.5.2 Direct voltage and current

11.5.2.1 Moving-coil indicator

This instrument comprises a coil, usually wound on a conducting former to provide eddy-current damping, with taut-band or pivot/control-spring suspension. In each case the coil rotates in the short air gap of a permanent magnet. The direction of the deflection depends on the polarity, so that unmodified indicators are usable only on d.c., and may have a centre zero if required. The error may be as low as ±~~0.~~1% of full-scale deflection; the range, from a few microamperes or millivolts up to 600 A and 750 V, which makes possible multirange d.c. test sets. The scale is generally linear (equispaced) and easily read. Non-linear scales can be obtained by shaping the magnet poles or the core to give a non-uniform air gap.

11.5.2.2 Corrections

The total power taken by a normal-range voltmeter can be 50 μW/V or more. For an ammeter the *total* series loss is 1–50 mW/A. Such powers may be a significant fraction of the total delivered to some electronic networks: in such

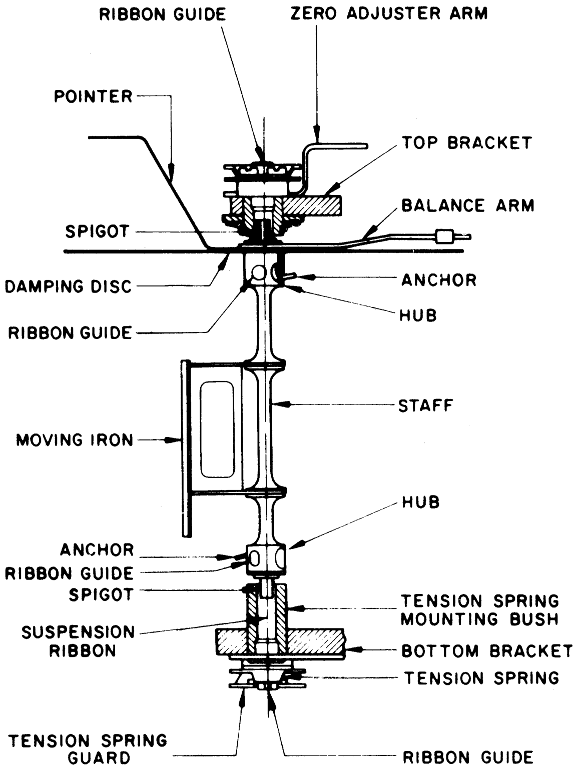


Figure 11.2 Taut-ribbon suspension assembly (Crompton Parkinson Ltd)

cases electronic and digital voltmeters with trivial power loss should be used instead.

11.5.2.3 Induced moving-magnet indicator

The instrument is polarised by a fixed permanent magnet and so is only suitable for d.c. A small pivoted iron vane is arranged to lie across the magnet poles, the field of which applies the restoring torque. One or more turns of wire carry the current to be measured around the vane, producing a magnetic field at right angles to that of the magnet, and this drive torque deflects the movement to the equilibrium position. It is a cheap, low-accuracy instrument which can have a nearly linear centre-zero scale, well suited to monitoring battery charge conditions.

11.5.3 Alternating voltage and current

11.5.3.1 Moving-iron indicator

This is the most common instrument for a.c. at industrial frequency. Some (but not all) instruments may also be used on d.c. The current to be measured passes through a magnetising coil enclosing the movement. The latter is formed either from a nickel–iron alloy vane that is *attracted* into the field, or from two magnetic alloy members which, becoming similarly polarised, mutually *repel*. The latter is more usual but its deflection is limited to about 90°; however, when combined with attraction vanes, a 240° deflection is obtainable but can be used only on a.c. on account of the use of material of high hysteresis.

The inherent torque deflection is square law and so, in principle, the instrument measures true r.m.s. values. By careful shaping of vanes when made from modern magnetic alloys, a good linear scale has been achieved except for the bottom 10%, which readily distinguishes the instrument from the moving-coil linear scale. Damping torque is produced by eddy-current reaction in an aluminium disc moving through a permanent-magnet field.

11.5.3.2 Precautions and corrections

These instruments work at low flux densities (e.g. 7 mT), and care is necessary when they are used near strong ambient magnetic fields despite the magnetic shield often provided. As a result of the shape of the B/H curve, a.c./d.c. indicators used on a.c. read high at the lower end and low at the higher end of the scale. The effect of hysteresis with d.c. measurements is to make the 'descending' reading higher than the 'ascending'.

Moving-iron instruments are available with class index values of 0.3–5.0, and many have similar accuracies for a.c. or d.c. applications. They tend to be single-frequency power instruments with a higher frequency behaviour that enables good r.m.s. readings to be obtained from supplies with distorted current waveforms, although excessive distortion will cause errors. Nevertheless, instruments can be designed for use in the lower audio range, provided that the power loss (a few watts) can be accepted. Ranges are from a few milliamperes and several volts upward. The power loss is usually significant and may have to be allowed for.

11.5.3.3 Electrodynamic indicator

In this the permanent magnet of the moving-coil instrument is replaced by an electromagnet (field coil). For ammeters the moving coil is connected in series with two fixed field coils; for voltmeters the moving and the fixed coils each have series-connected resistors to give the same time-constant, and the combination is connected in parallel across the voltage to be measured. With ammeters the current is limited by the suspension to a fraction of 1 A, and for higher currents it is necessary to employ non-reactive shunts or current transformers.

11.5.3.4 Precautions and corrections

The flux density is of the same order as that in moving-iron instruments, and similar precautions are necessary. The torque has a square-law characteristic and the scale is cramped at the lower end. By using the mean of reversed readings these instruments can be d.c. calibrated and used as adequate d.c./a.c. transfer devices for calibrating other a.c. instruments such as moving-iron indicators. The power loss (a few watts) may have to be allowed for in calculations derived from the readings.

11.5.3.5 Induction indicator

This single-frequency instrument is robust, but limited to class index (CI) number from 1.0 to 5.0. The current or voltage produces a proportional alternating magnetic field normal to an aluminium disc arranged to rotate. A second, phase-displaced field is necessary to develop torque. Originally it was obtained by pole 'shading', but this method is obsolete; modern methods include separate magnets one of which is shunted (Ockenden), a cylindrical aluminium movement (Lipman) or two electromagnets coupled by loops (Banner). The induction principle is most generally employed for the measurement of energy.

11.5.3.6 Moving-coil rectifier indicator

The place of the former copper oxide and selenium rectifiers has been taken by silicon diodes. Used with taut-band suspension, full-wave instrument rectifier units provide improved versions of the useful rectifier indicator.

11.5.3.7 Precautions and corrections

Owing to its inertia, the polarised moving-coil instrument gives a mean deflection proportional to the mean rectified current. The scale is marked in r.m.s. values on the assumption that the waveform of the current to be measured is sinusoidal with a form factor of 1.11. On non-sinusoidal waveforms the true r.m.s. value cannot be inferred from the reading: only the true *average* can be known (from the r.m.s. scale reading divided by 1.11). When true r.m.s. values are required, it is necessary to use a thermal, electrodynamic or square-law electronic instrument.

11.5.3.8 Multirange indicator

The rectifier diodes have capacitance, limiting the upper frequency of multirange instruments to about 20 kHz. The lower limit is 20–30 Hz, depending on the inertia of the movement. Indicators are available with CI numbers from 1.0 to 5.0, the multirange versions catering for a wide range of direct and alternating voltages and currents: a.c. 2.5 V–2.5 kV and 200 mA–10 A; d.c. 100 μV–2.5 kV and 50 μA–10 A. Non-linear resistance scales are normally included; values are derived from *internal* battery-driven out-of-balance bridge networks.

11.5.3.9 Moving-coil thermocouple indicator

The current to be measured (or a known proportion of it) is used as a heater current for the thermocouple, and the equivalent r.m.s. output voltage is steady because of thermal inertia—except for very low frequency (5 Hz and below). With heater current of about 5 mA in a 100 Ω resistor as typical, the possible voltage and current ranges are dictated by the usual series and shunt resistance, under the restriction that only the upper two-thirds of the square-law scale is within the effective range. The instrument has a frequency range from zero to 100 MHz. Normal radiofrequency (r.f.) low-range self-contained ammeters can be used up to 5 MHz with CI from 1.0 to 5.0. Instruments of this kind could be used in the secondary circuit of a r.f. current transformer for measurement of aerial current. The minimum measurable voltage is about 1 V.

The thermocouple, giving a true r.m.s. indication, is a primary a.c./d.c. transfer device. Recent evidence[4] indicates that the transfer uncertainty is only a few parts in 10^6. Such devices provide an essential link in the traceability chain between microwave power, r.f. voltage and the primary standards of direct voltage and resistance.

11.5.3.10 Precautions and corrections

The response is slow, e.g. 5–10 s from zero to full-scale deflection, and the overload capacity is negligible: the heater may be destroyed by a switching surge. Thermocouple voltmeters are low-impedance devices (taking, e.g., 500 mW at 100 V) and may be unsuitable for electronic circuit measurements.

11.5.4 Medium and high direct and alternating voltage

11.5.4.1 Electrostatic voltmeter

This is in effect a variable capacitor with fixed and moving vanes. The power taken (theoretically zero) is in fact sufficient to provide the small dielectric loss. The basically square-law characteristic can be modified by shaping the vanes to give a reasonably linear upper scale. The minimum useful range is 50–150 V in a small instrument, up to some hundreds of kilovolts for large fixed instruments employing capacitor multipliers. Medium-voltage direct-connected voltmeters for ranges up to 15 kV have CI ratings from 1.0 to 5.0. The electrostatic indicator is a true primary alternating/direct voltage (a.v./d.v.) transfer device, but has now been superseded by the thermocouple instrument. The effective isolation property of the electrostatic voltmeter is attractive on grounds of safety for high-voltage measurements.

11.5.5 Power

11.5.5.1 Electrodynamic indicator

The instrument is essentially similar to the electrodynamic voltmeter and ammeter. It is usually 'air-cored', but 'iron-cored' wattmeters with high-permeability material to give a better torque/mass ratio are made with little sacrifice in accuracy. The current (fixed) coils are connected in series with the load, and the voltage (moving) coil with its series range resistor is connected either (a) across the load side of the current coil, or (b) across the supply side. In (a) the instrument reading must be *reduced* by the power loss in the volt coil circuit (typically 5 W), and in (b) by the current coil loss (typically 1 W).

The volt coil circuit *power* corrections are avoided by use of a *compensated wattmeter*, in which an additional winding in series with the volt circuit is wound, turn-for-turn, with the current coils, and the connection is (a) above. The m.m.f. due to the volt-coil circuit current in the compensating coil will cancel the m.m.f. due to the volt circuit current in the current coils.

The volt circuit terminal marked ±⇐is immediately adjacent to the voltage coil; it should be connected to the current terminal similarly marked, to ensure that the wattmeter reads positive power and that negligible p.d. exists between fixed and moving coils, so safeguarding insulation and eliminating error due to electric torque.

When a wattmeter is used on d.c., the power should be taken from the mean of reversed readings; wattmeters read also the active (average) a.c. power. For the measurement of power at very low power factor, special wattmeters are made with weak restoring torque to give f.s.d. at, e.g., a power factor of 0.4. Range extension for all wattmeters on a.c. can be obtained by internal or external current transformers, internal resistive volt-range selectors or external voltage transformers. Typical self-contained ranges are between 0.5 and 20 A, 75 and 300 V. The usable range of frequency is about 30–150 Hz, with a best CI of 0.05.

Three-phase power can be measured by single dual-element instruments.

11.5.5.2 Thermocouple indicator

The modern versions of this instrument are more correctly termed electronic wattmeters. The outputs from current- and voltage-sensing thermocouples, multiplied together and amplified, can be displayed as average power. Typical d.c. and a.c. ranges are 300 mV–300 V and 10 mA–10 A, from pulsed and other non-sinusoidal sources at frequencies up to a few hundred kilohertz. The interaction with the network is low; e.g. the voltage network has typically an impedance of 10 kΩ/V.

11.5.5.3 Induction indicator

The power-reading instrument includes both current and voltage coils. The energy meter is referred to later. The instruments are frequency sensitive and are used only for fixed-frequency switchboard applications. The CI number is 1.0–5.0, usually the latter.

11.5.5.4 Electrodynamic reactive-power indicator

The active-power instrument can read reactive power if the phase of the volt circuit supply is changed by 90°, to give $Q = VI \sin \phi$ instead of $P = VI \cos \phi$ (for sinusoidal conditions).

11.5.6 Maximum alternating current

11.5.6.1 Maximum-demand instrument

The use of an auxiliary pointer carried forward by the main pointer and remaining in position makes possible the indication of maximum current values, but the method is not satisfactory, because it demands large torques, a condition that reduces effective damping and gives rise to overswing. A truer maximum demand indication, and one that is insensitive to momentary peaks, is obtainable with the aid of a thermal bimetal. Passing the current to be indicated through the bimetal gives a thermal lag of a few minutes. For long-period indication (1 h or more) a separate heater is required.

Such instruments have been adapted for three-phase operation using separate heaters. A recent development permits a similar device to be used for two phases of a three-phase supply. The currents are summed *linearly*, and the maximum of the combined unbalanced currents maintained for 45 min is recorded on a kV-A scale marked in terms of nominal voltage.

11.5.7 Power factor

Power factor indicators have both voltage and current circuits, and can be interconnected to form basic electrodynamic or moving-iron direct-acting indicators. The current coils are fixed; the movement is free, the combined voltage- and current-excited field producing both deflecting and controlling torque. The electrodynamic form has the greater accuracy, but is restricted to a 90° deflection, as compared with 360° (90° lag and lead, motoring and generating) of the moving-iron type.

One electrodynamic form comprises two fixed coils carrying the line current, and a pair of moving volt coils set with their axes mutually at almost 90° (*Figure 11.3*). For one-phase operation the volt-coil currents are nearly in quadrature, being, respectively, connected in series with an inductor L and a resistor R. For three-phase working, L is replaced by a resistor r and the ends of the volt-coil circuits taken, respectively, to the second and third phases. A three-phase moving-iron instrument for balanced loads has one current coil and three volt coils (or three current and one voltage); if the load is unbalanced, three current and three volt coils are required. All these power factor indicators are industrial frequency devices.

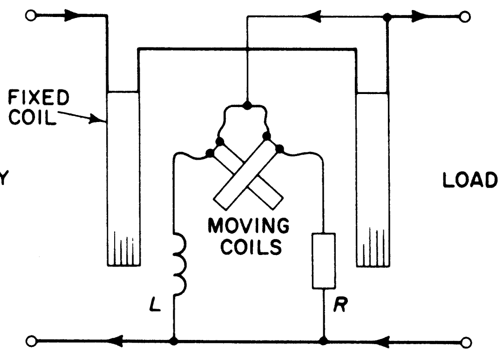


Figure 11.3 Single-phase electrodynamic power factor indicator

The one-phase electrodynamic wattmeter can be used as a phase-sensing instrument. When constant r.m.s. voltages and currents from two, identical-frequency, *sinusoidal* power supplies are applied to the wattmeter, then the phase change in one system produces power scale changes which are proportional to cos ϕ.

Many electronic phase-meters of the analogue and digital type are available which, although not direct-acting, have high impedance inputs, and they measure the phase displacement between any two voltages (of the same frequency) over a wide range of frequencies. Either the conventional solid state digital instruments generate a pulse when the two voltages pass zero, and measure and display the time between these pulses as phase difference, or the pulses trigger a multivibrator for the same period, to generate an output current proportional to phase (time) difference. These instruments provide good discrimination and, with modern high-speed logic switching, the instant for switch-on (i.e. pulse generation) is less uncertain: conversely, the presence of small harmonic voltages around zero could lead to some ambiguity.

11.5.8 Phase sequence and synchronism

A small portable form of *phase sequence* indicator is essentially a primitive three-phase induction motor with a disc rotor and stator phases connected by clip-on leads to the three-phase supply. The disc rotates in the marked forward direction if the sequence is correct. These instruments have an intermittent rating and must not be left in circuit.

The *synchroscope* is a power factor instrument with rotor slip-rings to allow continuous rotation. The moving-iron type is robust and cheap. The direction of rotation of the rotor indicates whether the incoming three-phase system is ‘fast’ or ‘slow’, and the speed of rotation measures the difference in frequency.

11.5.9 Frequency

Power frequency monitoring frequency indicators are based on the response of reactive networks. Both electrodynamic and induction instruments are available, the former having the better accuracy, the latter having a long scale of 300° or more.

One type of indicator consists of two parallel fixed coils tuned to slightly different frequencies, and their combined currents return to the supply through a moving coil which lies within the resultant field of the fixed coils. The position of the moving coil will be unique for each frequency, as the currents (i.e. fields) of the fixed coils are different unique functions of frequency. The indications are within a restricted range of a few hertz around nominal frequency. A ratiometer instrument, which can be used up to a few kilohertz, has a permanent-magnet field in which lie two moving coils set with their planes at right angles. Each coil is driven by rectified a.c., through resistive and capacitive impedances, respectively, and the deflection is proportional to frequency.

Conventional solid state counters are versatile time and frequency instruments. The counter is based on a stable crystal reference oscillator (e.g. at 1 MHz) with an error between 1 and 10 parts in 10^8 per day. The displays have up to eight digits, with a top frequency of 100 MHz (or 600 MHz with heterodyning). The resolution can be 10 ns, permitting pulse widths of 1000 ns to be assessed to 1%: but a 50 Hz reading would be near to the bottom end of the display, giving poor discrimination. All counters provide for measurements giving the period of a waveform to a good discrimination.

11.6 Integrating (energy) metering

Integrating meters record the time integral of active, reactive and apparent power as a continuous summation. The integration may be limited by a specified total energy (e.g. prepayment meters) or by time (e.g. maximum demand). Meters for a.c. supplies are all of the induction type, with measuring elements in accordance with the connection of the load (one-phase or three- or four-wire three-phase). Manufacturing and testing specifications are given in BS 37. Instrument transformers for use with meters are listed in BS 3938 and BS 3941.

11.6.1 Single-phase meter

The rotor is a light aluminium disc on a vertical spindle, supported in low-friction bearings. The lower bearing is a sapphire cup, carrying the highly polished hemispherical hardened steel end of the spindle. The rotor is actuated by an induction driving element and its speed is controlled by an eddy current brake. The case is usually a high-quality black phenolic moulding with integral terminal block. The frame is a rigid high-stability iron casting which serves as the mounting, as part of a magnetic flux path, and as a shield against ambient magnetic fields.

The driving element has the basic form shown in *Figure 11.4*. It has two electromagnets, one voltage-excited and the other current-excited. The volt magnet, roughly of E-shape, has a nearly complete magnetic circuit with a volt coil on the central limb. Most of the flux divides between the outer limbs, but the *working* flux from the central limb penetrates the disc and enters the core of the current magnet. The latter, of approximate U-shape, is energised by a coil carrying the load current. With a condition of *zero* load current, the working flux from the volt magnet divides equally between the two limbs of the current magnet and returns to the volt magnet core through the frame, or through an iron path provided for this purpose. If the volt magnet flux is symmetrically disposed, the eddy current induced in the disc does not exert any net driving torque and the disc remains stationary. The volt magnet flux is approximately in phase quadrature with the applied voltage.

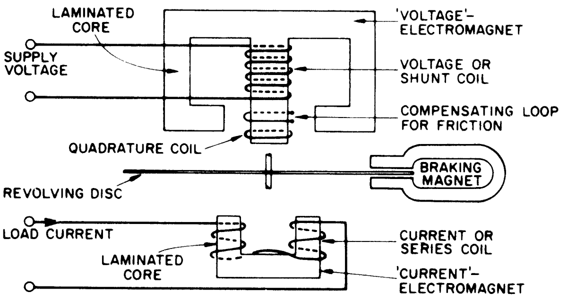


Figure 11.4 Essential parts of an induction meter

When a load current flows in the coil of the current magnet, it develops a co-phasal flux, interacting with the volt magnet flux to produce a resultant that 'shifts' from one current magnet pole to the other. Force is developed by interaction of the eddy currents in the disc and the flux in which it lies, and the net torque is proportional to the voltage and to that component of the load current in phase with the voltage—i.e. to the active power. The disc therefore rotates in the direction of the 'shift'.

The disc rotates through the field of a suitably located permanent isotropic brake magnet, and induced currents provide a reaction proportional to speed. Full-load adjustment is effected by means of a micrometer screw which can set the radial position of the brake magnet.

The volt magnet working flux lags the voltage by a phase angle rather less than 90° (e.g. by 85°). The phase angle is made 90° by providing a closed *quadrature coil* on the central limb, with its position (or resistance) capable of adjustment.

The accuracy of the meter is affected on low loads by pivot friction. A *low-load adjustment*, consisting of some device that introduces a slight asymmetry in the volt magnet working flux, is fitted to mitigate frictional error. There are several methods of producing the required asymmetry: one is the insertion of a small magnetic vane into the path of the working flux. The asymmetry results in the development at zero load of a small forward torque, just sufficient to balance friction torque without causing the disc to rotate.

11.6.1.1 Performance of single-phase meters

The limits of permissible error are defined in BS 37. In general, the error should not exceed ±2% over most of the working range for credit meters. For prepayment meters, +2 and −3% for loads above 1/30 full load at unity power factor, at any price setting to be used, are the specified limits. Tests for compliance with this requirement at and below 1/10 marked current must be made with the coin condition that not less than one nor more than three coins of the highest denomination acceptable by the meter are prepaid.

It is usual for manufacturers to adjust their meters to less than one-half of the permissible error over much of the working range. The mean error of an individual meter is, of course, less than the maximum observable error, and the mean collective error of a large number of meters is likely to be much less than ±1% at rated voltage and frequency. If no attempt is made during production testing to bias the error in one direction, the weighted mean error of many meters taken collectively is probably within 0.5%, and is likely to be positive (see *Figure 11.5*(*a*)).

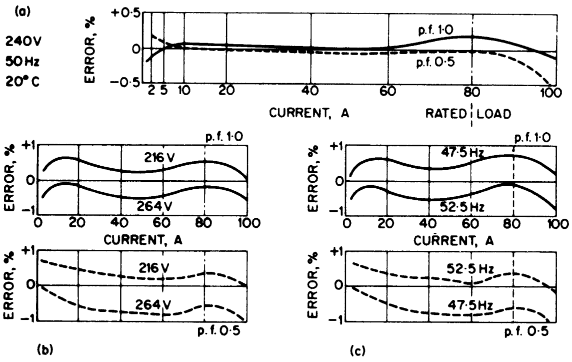


Figure 11.5 Typical performance of 240 V, 50 Hz, 20–80 A house-service meter (Ferranti Ltd)

11.6.1.2 Temperature

The temperature error of a.c. meters with no temperature compensation is negligible under the conditions normally existing in the UK. In North America it is common practice to fix meters on the outside of buildings, where they are subjected to wide temperature ranges (e.g. 80°C), which makes compensation desirable. However, even in the UK it is common to provide compensation (when required) in the form of a strip of nickel-rich alloy encircling the stator magnet. Its use is advantageous for short-duration testing, as there is an observable difference in error of a meter when cold and after a 30 min run.

11.6.1.3 Voltage and frequency

Within the usual limits of ±4% in voltage and ±0.2% in frequency, the consequent errors in a meter are negligible. Load-shedding, however, may mean substantial reductions of voltage and frequency. A significant reduction in voltage usually causes the meter to read high on all loads, as shown in *Figure 11.5*(*b*). Reduction of the supply frequency makes the meter run faster on high-power-factor loads and slower on lagging reactive loads: an increase of frequency has the opposite effect (*Figure 11.5*)(*c*). The errors are cumulative if both voltage and frequency variations occur together.

11.7 Electronic instrumentation

The rapid development of large-scale integrated (l.s.i.) circuits, as applied to analogue and especially digital circuitry, has allowed both dramatic changes in the format of measurement instrumentation by designers and in the measurement test systems available for production or control applications.

A modern instrument is often a multipurpose assembly, including a microprocessor for controlling both the sequential measurement functions and the subsequent programmed assessment of the data retained in its store; hence, these instruments can be economical in running and labour costs by avoiding skilled attention in the recording and statistical computation of the data. Instruments are often designed to be compatible with standardised interfaces, such as the IEC bus or IEEE-488 bus, to form automatic measurement systems for interactive design use of feedback digital-control applications; typically a system would include a computer-controlled multiple data logger with 30 inputs from different parameter sensors, statistical and system analysers, displays and storage. The input analogue parameters would use high-speed analogue-to-digital (A/D) conversion, so minimising the front-end analogue conditioning networks while maximising the digital functions (e.g. filtering and waveform analyses by digital rather than analogue techniques). The readout data can be assembled digitally and printed out in full, or presented through D/A converters in graphical form on *X–Y* plotters using the most appropriate functional co-ordinates to give an economical presentation of the data.

The production testing of the component subsystems in the above instruments requires complex instrumentation and logical measurements, since there are many permutations of these minute logic elements resulting in millions of high-speed digital logic operations—which precludes a complete testing cycle, owing to time and cost. Inspection testing of l.s.i. analogue or digital devices and subassemblies demands systematically programmed measurement testing sequences with varied predetermined signals; similar automatic test equipments (ATE) are becoming more general in any industrial production line which employs these modern instruments in process or quality-control operations. Some companies are specialising in computer-orientated measurement techniques to develop this area of ATE; economic projections indicate that the present high growth rate in ATE-type measurement systems will continue or increase during the next few years.

The complementary nature of automated design, production and final testing types of ATE leads logically to the interlinking of these separate functions for optimising and improving designs within the spectrum of production, materials handling, quality control and costing to provide an overall economic and technical surveillance of the product.

Some of the more important instruments have been included in the survey of electronic instrumentation in the following pages, but, owing to the extensive variety of instruments at present available and the rapid rate of development in this field, the selection must be limited to some of the more important devices.

Digital instruments usually provide an output in visual decimal or coded digit form as a discontinuous function of a smoothly changing input quantity. In practice the precision of a digital instrument can be extended by adding digits to make it better than an analogue instrument, although the final (least significant) digit must always be imprecise to ±0.5.

11.7.1 Digital voltmeters

These provide a digital display of d.c. and/or a.c. inputs, together with coded signals of the visible quantity, enabling the instrument to be coupled to recording or control systems. Depending on the measurement principle adopted, the signals are sampled at intervals over the range 2–500 ms. The basic principles are: (a) linear ramp; (b) charge balancing; (c) successive-approximation/potentiometric; (d) voltage to frequency integration; (e) dual slope; (f) multislope; and (g) some combination of the foregoing.

Modern digital voltmeters often include a multiway socket connection (e.g. BCD; IEC bus; IEEE-488 bus, etc.) for data/interactive control applications.

11.7.1.1 Linear ramp

This is a voltage/time conversion in which a linear timebase is used to determine the time taken for the internally generated voltage v_s to change by the value of the unknown voltage *V*. The block diagram (*Figure 11.6*(*b*)) shows the use of comparison networks to compare *V* with the rising

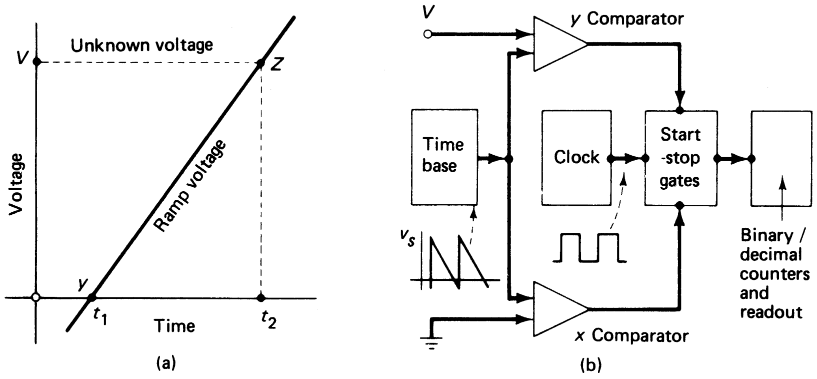


Figure 11.6 Linear ramp digital voltmeter. (a) Ramp voltage; (b) block diagram

(or falling) v_S; these networks open and close the path between the continuously running oscillator, which provides the counting pulses at a fixed clock rate, and the counter. Counting is performed by one of the binary-coded sequences; the translation networks give the visual decimal output. In addition, a binary-coded decimal output may be provided for monitoring or control purposes.

Limitations are imposed by small non-linearities in the ramp, the instability of the ramp and oscillator, imprecision of the coincidence networks at instants *y* and *z*, and the inherent lack of noise rejection. The overall uncertainty is about ±0.05% and the measurement cycle would be repeated every 200 ms for a typical four-digit display.

Linear 'staircase' ramp instruments are available in which *V* is measured by counting the number of equal voltage 'steps' required to reach it. The staircase is generated by a solid state diode pump network, and linearities and accuracies achievable are similar to those with the linear ramp.

11.7.1.2 Charge balancing

This principle[5] employs a pair of differential-input transistors used to charge a capacitor by a current proportional to the unknown d.c. voltage, and the capacitor is then discharged by a large number of small $+\delta q$ and $-\delta q$ quantities, the elemental discharges being sensed and directed by fast comparator/flip-flop circuits and the total numbers being stored. Zero is in the *middle* of the measurement range and the displayed result is proportional to the difference between the number of $+\delta q$, $-\delta q$ events.

The technique which, unlike dual-ramp methods, is inherently bipolar, claims some other advantages, such as improved linearity, more rapid recovery from overloads, higher sensitivity and reduced noise due to the averaging of thousands of zero crossings during the measurement, as well as a true *digital* auto-zero subtraction from the next measurement, as compared with the more usual capacitive-stored analogue offset p.d. being subtracted from the measured p.d.

Applications of this digital voltmeter system (which is available on a monolithic integrated circuit, Ferranti ZN 450) apart from d.c. and a.c. multimeter uses, include interfacing it *directly* with a wide range of conventional transducers such as thermocouples, strain gauges, resistance thermometers, etc.

11.7.1.3 Successive approximation

As it is based on the potentiometer principle, this class produces very high accuracy. The arrows in the block diagram of *Figure 11.7* show the signal-flow path for one version; the resistors are selected in sequence so that, with a constant current supply, the test voltage is created within the voltmeter.

Each decade of the unknown voltage is assessed in terms of a sequence of accurate stable voltages, graded in descending magnitude in accordance with a binary (or similar) counting scale. After each voltage approximation of the final result has been made and stored, the residual voltage is then automatically re-assessed against smaller standard voltages, and so on to the smallest voltage discrimination required in the result. Probably four logic decisions are needed to select the major decade value of the unknown voltage, and this process will be repeated for each lower decade in decimal sequence until, after a few milliseconds, the required voltage is stored in a coded form. This voltage is then translated for decimal display. A binary-coded sequence could be as shown in *Table 11.2*, where the numerals in **bold** type represent a logical *rejection* of that number and a progress to the next lower value. The actual sequence of logical decisions is more complicated than is suggested by the example. It is necessary to sense the initial polarity of the unknown signal, and to select the range and decimal marker for the read-out; the time for the logic networks to settle must be longer for the earlier (higher voltage) choices than for the later ones, because they must be of the highest possible accuracy; offset voltages may be added to the earlier logic choices, to be withdrawn later in the sequence; and so forth.

The total measurement and display takes about 5 ms. When noise is present in the input, the necessary insertion of filters may extend the time to about 1 s. As noise is more troublesome for the smaller residuals in the process, it is sometimes convenient to use some different techniques for the latter part. One such is the voltage–frequency principle (see below), which has a high noise rejection ratio. The

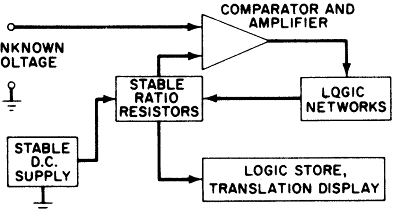


Figure 11.7 Successive-approximation digital voltmeter

Table 11.2 Voltages obtained from residual (difference) currents across high-stability resistors

Unknown analogue voltage	3	9	.	2	0	6
Logic decisions in	**8**	8		**8**	**8**	**8**
vertical binary	**4**	**4**		**4**	**4**	4
sequences and in	2	**2**		2	**2**	2
descending 'order'	1	1		**1**	**1**	**1**
Decoded decimal display	3	9	.	2	0	6

reduced accuracy of this technique can be tolerated as it applies only to the least significant figures.

11.7.1.4 Voltage–frequency

The converter (*Figure 11.8*) provides output pulses at a frequency proportional to the instantaneous unknown input voltage, and the non-uniform pulse spacing represents the variable frequency output. The decade counter accumulates the pulses for a predetermined time T and in effect measures the average frequency during this period. When T is selected to coincide with the longest period time of interfering noise (e.g. mains supply frequency), such noise averages out to zero.

Instruments must operate at high conversion frequencies if adequate discrimination is required in the final result. If a six-digit display were required within 5 ms for a range from zero to 1.000 00 V to ±0.01%, then 10^5 counts during 5 ms are called for, i.e. a 20 MHz change from zero frequency with a 0.01% voltage–frequency linearity. To reduce the frequency range, the measuring time is increased to 200 ms or higher. Even at the more practical frequency of 0.5 MHz that results, the inaccuracy of the instrument is still determined largely by the non-linearity of the voltage–frequency conversion process.

In many instruments the input network consists of an integrating operational amplifier in which the average input voltage is 'accumulated' in terms of charge on a capacitor in a given time.

11.7.1.5 Dual-slope

This instrument uses a composite technique consisting of an integration (as mentioned above) and an accurate measuring network based on the ramp technique. During the integration (*Figure 11.9*) the unknown voltage V_x is switched at time zero to the integration amplifier, and the initially uncharged capacitor C is charged to V_x in a known time T_1, which may be chosen so as to reduce noise interference. The ramp part of the process consists in replacing V_x by a *reversed* biased direct reference voltage V_r which produces a constant-current discharge of C; hence, a known linear voltage/time change occurs across C. The total voltage change to zero can be measured by the method used in the linear-ramp instrument, except that the slope is negative and that counting begins at maximum voltage. From the diagram in *Figure 11.9*(*a*) it follows that $\tan \theta_x/\tan \theta_r = (T_2 - T_1)/T_1 = V_x/V_r$ so that V_x is directly proportional to $T_2 - T_1$. The dual-slope method is seen to depend ultimately on time-base linearity and on the measurement of time difference, and is subject to the same limitations as the linear-ramp method, but with the important and fundamental quality of inherent noise rejection capability.

Noise interference (of series-mode type) is principally due to e.m.f.s at power supply frequency being induced into the d.c. source of V_x. When T_1 equals the periodic time of the interference (20 ms for 50 Hz), the charging p.d. v_c across C would follow the dotted path (*Figure 11.9*(*a*)) without changing the final required value of v_c, thus eliminating this interference.

11.7.1.6 Multislope

At time T_1 (above) the maximum dielectric absorption in C will coincide with v_c maximum; this effect degrades (a) the linearity of the run-down slope during $T_2 - T_1$ and (b) the identification of zero p.d. One improved 'multislope' technique reduces dielectric absorption in C by inserting various reference voltages during the run-up period and, after subtraction, leaves a lower p.d., v'_c, prior to run-down while storing the most significant digit of V_x during the run-up period. During run-down, C is discharged rapidly to measure the next smaller digit; the residue of v'_c, including overshoot, is assessed to give the remaining three digits in sequence using three v_r/t functions of slope +1/10, −1/100, +1/1000 compared with the initial rapid discharge. Measurement time is reduced, owing to the successive digits being accumulated during the measurement process.[6]

11.7.1.7 Mixed techniques

Several techniques can be combined in the one instrument in order to exploit to advantage the best features of each. One accurate, precise, digital voltmeter is based upon precision inductive potentiometers, successive approximation, and the dual-slope technique for the least significant figures. An uncertainty of 10 parts in 10^6 for a 3-month period is claimed, with short-term stability and a *precision* of about 2 parts in 10^6.

11.7.1.8 Digital multimeters

Any digital voltmeter can be scaled to read d.c. or a.c. voltage, current, immittance or any other physical property, provided that an appropriate transducer is inserted. Instruments scaled for alternating voltage and current

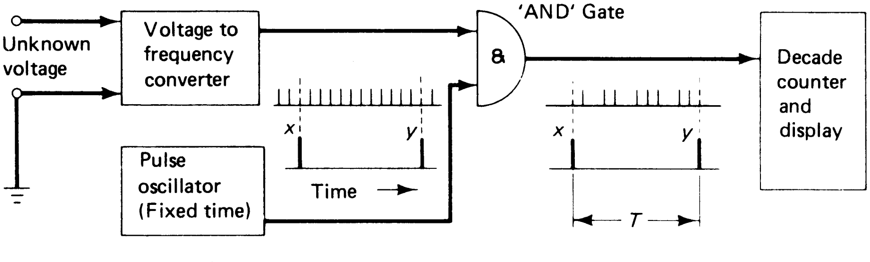
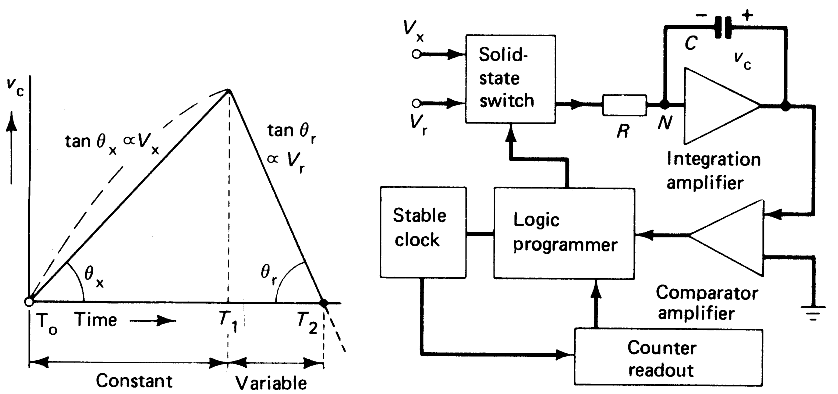

Unknown voltage
Voltage to frequency converter
'AND' Gate
&
Decade counter and display
Pulse oscillator (Fixed time)
x
y
Time
x
y
T

Figure 11.8 Voltage-to-frequency digital converter



(a) (b)

Figure 11.9 Dual-slope digital voltmeter. (a) Ramp voltage; (b) block diagram

normally incorporate one of the a.c./d.c. converter units listed in a previous paragraph, and the quality of the result is limited by the characteristics inherent in such converters. The digital part of the measurement is more accurate and precise than the analogue counterpart, but may be more expensive.

For systems application, programmed signals can be inserted into, and binary-coded or analogue measurements received from, the instrument through multiway socket connections, enabling the instrument to form an active element in a control system (e.g. IEC bus, IEEE-488 bus).

Resistance, capacitance and inductance measurements depend to some extent on the adaptability of the basic voltage-measuring process. The dual-slope technique can be easily adapted for two-, three- or four-terminal ratio measurements of resistance by using the positive and negative ramps in sequence; with other techniques separate impedance units are necessary. (See *Table 11.3*.)

11.7.1.9 Input and dynamic impedance

The high precision and small uncertainty of digital voltmeters make it essential that they have a high input impedance if these qualities are to be exploited. Low test voltages are often associated with source impedances of several hundred kilo-ohms; for example, to measure a voltage with source resistance 50 kΩ to an uncertainty of ±0.005% demands an instrument of input resistance 1 GΩ, and for a practical instrument this must be 10 GΩ if the loading error is limited to one-tenth of the total uncertainty.

The dynamic impedance will vary considerably during the measuring period, and it will always be lower than the quoted null, passive, input impedance. These changes in dynamic impedance are coincident with voltage 'spikes' which appear at the terminals owing to normal logic functions; this noise can adversely affect components connected to the terminals, e.g. Weston standard cells.

Input resistances of the order of 1–10 GΩ represent the conventional range of good-quality insulators. To these must be added the stray parallel reactance paths through unwanted capacitive coupling to various conducting and earth planes, frames, chassis, common rails, etc.

11.7.1.10 Noise limitation

The information signal exists as the p.d. between the two input leads; but each lead can have unique voltage and impedance conditions superimposed on it with respect to the basic *reference* or *ground* potential of the system, as well as another and different set of values with respect to a local *earth* reference plane.

An elementary electronic instrumentation system will have at least one ground potential and several earth connections—possibly through the (earthed) neutral of the main supply, the signal source, a read-out recorder or a cathode ray oscilloscope. Most true earth connections are at different electrical potentials with respect to each other, owing to circulation of currents (d.c. to u.h.f.) from other apparatus, through a finite earth resistance path. When multiple earth connections are made to various parts of a high-gain amplifier system, it is possible that a significant frequency spectrum of these signals will be introduced as electrical noise. It is this interference which has to be rejected by the input networks of the instrumentation, quite apart from the concomitant removal of any electrostatic/electromagnetic noise introduced by direct coupling into the signal paths. The total contamination voltage can be many times larger (say 100) than the useful information voltage level.

Electrostatic interference in input cables can be greatly reduced by 'screened' cables (which may be 80% effective as screens), and electromagnetic effects minimised by transposition of the input wires and reduction in the 'aerial loop' area of the various conductors. Any residual effects, together with the introduction of 'ground and earth-loop' currents into the system, are collectively referred to as *series* and/or *common-mode* signals.

11.7.1.11 Series and common-mode signals

Series-mode (normal) interference signals, V_{sm}, occur in series with the required information signal. Common-mode interference signals, V_{cm}, are present in both input leads with respect to the reference potential plane: the required information signal is the difference voltage between these leads. The results are expressed as *rejection ratio* (in decibels) with respect to the *effective* input error signal, V_{e}, that the interference signals produce, i.e.

$$K_{\mathrm{sm}} = 20\log(V_{\mathrm{sm}}/V_{\mathrm{e}}) \text{ and } K_{\mathrm{cm}} = 20\log(V_{\mathrm{cm}}/V_{\mathrm{e}})$$

where K is the rejection ratio. The series-mode rejection networks are *within* the amplifier, so V_{e} is measured with *zero* normal input signal as V_{e} = output interference

Table 11.3 Typical characteristics of digital multimeters

(a) Type *(b) ±Uncertainty of maximum reading* *(c) Resolution*	*Voltage*	*Current*	*Resistance*	*Voltage and frequency min ≃ 40 Hz*	*Current and frequency min ≃ 40 Hz*	*Parallel input R and C*	*Recalibration period* (months)*
(a) $3\frac{1}{2}$ digits general-purpose (mean sensing)	0.1 V–1 kV	0.1 A–1 A	1.1 kΩ–11 MΩ	0.1 V–1.0 kV	0.1 A–1.0 A	10 MΩ	3
(b)	0.3%	0.8%	0.5%	1.5% → 8% (2 kHz → 10 kHz)	2% → 3.5% (2 kHz → 5 kHz)	30 pF	
(c)	100 μV	100 μA	1 Ω!	100 μV	100 μA		
(a) $4\frac{1}{2}$ digits general purpose (mean sensing)	20 mV–1 kV	200 μA → 2 A	200 Ω → 20 MΩ	200 mV – 100 V	200 μA – 100 mA	1 MΩ	
(b)	0.03%	0.1% → 0.6%	0.02% → 0.1%	0.15% → 0.5% (10 kHz → 20 kHz)	0.3% → 0.7% (10 kHz → 20 kHz)	100 pF	12
(c)	1 μV	10 nA	10 mΩ!	10 μV	10 nA		
(a) $5\frac{1}{2}$ digits (r.m.s.)	0.1 V–1 kV	—	0.1 kΩ → 15 MΩ	*r.m.s.* 1 V–1 kV	—	*d.c.* 10^{10} Ω → 10^{7} Ω	3
(b)	0.01%	—	50 → 1000 ppm	0.1% → 9% (20 kHz → 1 MHz)	—	2 MΩ	
(a) $5\frac{1}{2}$ digits (mean)				*mean* 1 V–1 kV		100 pF	
(b)				0.2% → 0.8% (100 Hz → 250 kHz)			
(c)	1 μV	—	1 mΩ!	10 μV	—		
(a) $6\frac{1}{2}$ digits precision (r.m.s. up to 200 V)	10 mV–1 kV	—	14 Ω–14 MΩ	0.1 V–1 kV	—	1 MΩ 150 pF	6
(b)	20 ppm	—	50 ppm	0.15% → 0.9% (10 kHz → 100 kHz)	—		
(c)	1 μV	—	1 mΩ!	1 μV	—		
(a) $8\frac{1}{2}$ digits	0.1 V–1 kV	—	0.1 kΩ → 1.4 GΩ	0.1 V–1 kV r.m.s.	—	*a.c.* 1 MΩ 150 pF	3
(b)	15 ppm	—	17–46 ppm (0.1% 1 GΩ range)	0.02% → 2% 40 Hz → 1 MHz	—	*d.c.* 10 GΩ to 10 MΩ 10 MΩ (0.1 → 1 kV)	
(c)	10 nV	—	10 μΩ!	1 μV	—		

* Specifications often include 12-, 6-, and 3-month periods each with progressively smaller uncertainties of measurement down to 24-h statements for high precision instruments.

voltage ÷ gain of the amplifier appropriate to the bandwidth of V_e.

Consider the elementary case in *Figure 11.10*, where the input-lead resistances are unequal, as would occur with a transducer input. Let r be the difference resistance, C the cable capacitance, with common-mode signal and error voltages V_{cm} and V_e, respectively. Then the common-mode numerical ratio (c.m.r.) is

$$V_{cm}/V_e = V_{cm}/ri = 1/2\pi fCr$$

assuming the cable insulation to be ideal, and $X_C \gg r$. Clearly, for a common-mode statement to be complete, it must have a stated frequency range and include the resistive imbalance of the source. (It is often assumed in c.m.r. statements that $r = 1$ kΩ.)

The c.m.r. for a digital voltmeter could be typically 140 dB (corresponding to a ratio of $10^7/1$) at 50 Hz with a 1 kΩ line imbalance, and leading consequently to $C = 0.3$ pF. As the normal input cable capacitance is of the order of 100 pF/m, the situation is not feasible. The solution is to inhibit the return path of the current i by the introduction of a guard network. Typical guard and shield parameters are shown in *Figure 11.11* for a six-figure digital display on a voltmeter with ±0.005% uncertainty. Consider the magnitude of the common-mode error signal due to a 5 V, 50 Hz common-mode voltage between the shield earth E_1 and the signal earth E_2:

N–G not connected. The a.c. *common*-mode voltage drives current through the guard network and causes a change of 1.5 mV to appear across r as a *series*-mode signal; for $V_s = 1$ V this represents an error V_e of 0.15% for an instrument whose quality is ±0.005%.

N–G connected. The *common*-mode current is now limited by the shield impedance, and the resultant *series*-mode signal is 3.1 μV, an acceptably low value that will be further reduced by the noise rejection property of the measuring circuits.

11.7.1.12 *Floating-voltage measurement*

If the d.c. voltage difference to be measured has a p.d. to E_2 of 100 V, as shown, then with N–G open the change in p.d. across r will be 50 μV, as a *series*-mode error of 0.005% for

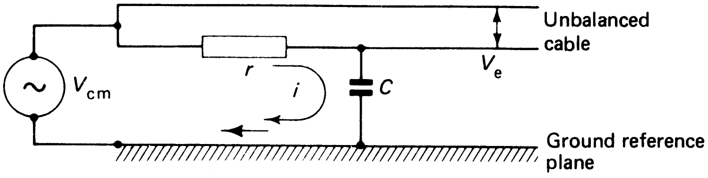


Figure 11.10 Common-mode effect in an unbalanced network

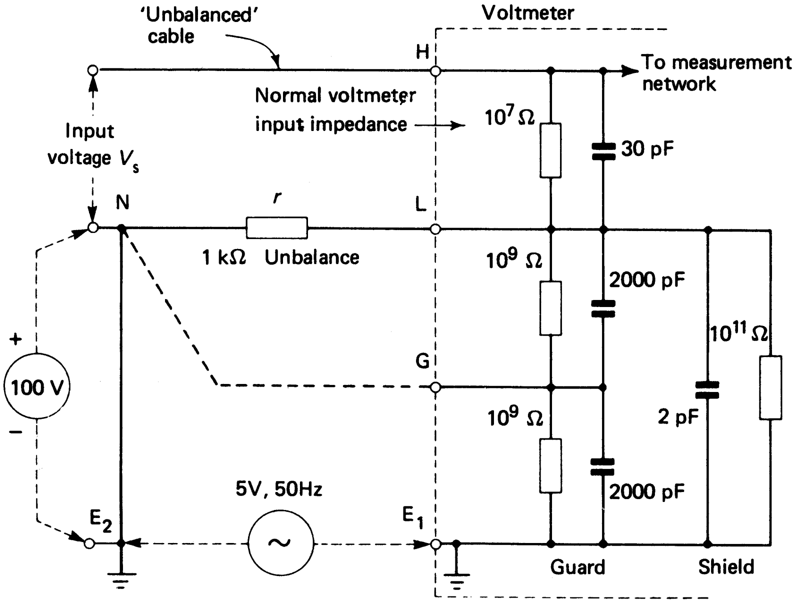


Figure 11.11 Typical guard and shield network for a digital voltmeter

a 1 V measurement. With N–G connected the change will be 1 µV, which is negligible.

The interconnection of electronic apparatus must be carefully made to avoid systematic measurement errors (and short circuits) arising from incorrect screen, ground or earth potentials.

In general, it is preferable, wherever possible, to use a *single* common reference node, which should be at zero signal reference potential to avoid leakage current through *r*. Indiscriminate interconnection of the shields and screens of adjacent components can increase *noise* currents by short-circuiting the high-impedance stray path between the screens.

11.7.1.13 Instrument selection

A precise seven-digit voltmeter, when used for a 10 V measurement, has a discrimination of ±1 part in 10^6 (i.e. ±10 µV), but has an uncertainty ('accuracy') of about ±10 parts in 10^6 (i.e. ±100 µV). The distinction is important with digital read-out devices, lest a higher quality be accorded to the number indicated than is in fact justified. The quality of any reading must be based upon the time stability of the total instrument since it was last calibrated against *external* standards, and the cumulative evidence of previous calibrations of a like kind.

Selection of a digital voltmeter from the list of types given in *Table 11.3* is based on the following considerations:

(1) No more digits than necessary, as the cost per extra digit is high.
(2) High input impedance, and the effect on likely sources of the *dynamic* impedance.
(3) Electrical noise rejection, assessed and compared with (a) the common-mode rejection ratio based on the input and guard terminal networks and (b) the actual inherent noise rejection property of the measuring principle employed.
(4) Requirements for binary-coded decimal, IEEE-488, storage and computational facilities.
(5) Versatility and suitability for use with a.c. or impedance converter units.
(6) Use with transducers (in which case (3) is the most important single factor).

11.7.1.14 Calibration

It will be seen from *Table 11.3* that digital voltmeters should be recalibrated at intervals between 3 and 12 months. Built-in self-checking facilities are normally provided to confirm satisfactory operational behaviour, but the 'accuracy' check cannot be better than that of the included standard cell or Zener diode reference voltage and will apply only to *one* range. If the user has available some low-noise stable voltage supplies (preferably batteries) and some high-resistance helical voltage dividers, it is easy to check logic sequences, resolution, and the approximate ratio between ranges. In order to demonstrate traceability to national standards, a voltage source whose voltage is traceable to national standards should be used. This can either be done by the user or, more commonly, by sending the instrument to an accredited calibration laboratory.

11.7.2 Digital wattmeters[7–9]

Digital wattmeters employ various techniques for sampling the voltage and current waveforms of the power source to produce a sequence of instantaneous power representations from which the average power is obtained. One new NPL method uses high-precision successive-approximation A/D integrated circuits with the multiplication and averaging of the digitised outputs being completed by computer.[10,11] A different 'feedback time-division multiplier' principle is employed by Yokogawa Electric Limited.[1]

The NPL digital wattmeter uses 'sample and hold' circuits to capture the two instantaneous voltage (v_x) and current (i_x) values, then it uses a novel double dual-slope multiplication technique to measure the average power from the mean of numerous instantaneous ($v_x i_x$) power measurements captured at precise intervals during repetitive waveforms. Referring to the single dual-ramp (*Figure 11.9*(*a*)) the a.c. voltage v_x (captured at T_0) is measured as $v_x T_1 = V_r(T_2 - T_1)$. If a second voltage (also captured at T_0) is proportional to i_x and integrated for $T_2 - T_1$, then, if it is reduced to zero (by V_r) during the time $T_3 - T_2$ it follows that $i_x(T_2 - T_1) = V_r(T_3 - T_2)$.

The instantaneous power (at T_0) is $v_x i_x = K(T_3 - T_2)$ and, by scaling, the mean summation of all counts such as $T_3 - T_2$ equals the average power. The prototype instrument measures power with an uncertainty of about ±0.03% f.s.d. (and ±0.01%, between 50 and 400 Hz, should be possible after further development).

The 'feedback time division multiplier' technique develops rectangular waveforms with the pulse height and width being proportional, respectively, to the instantaneous voltage (v_x) and current (i_x); the average 'area' of all such instantaneous powers ($v_x i_x$) is given (after 1.p. filtering) as a d.c. voltage measured by a DVM scaled in watts; such precision digital wattmeters, operating from 50 to 500 Hz, have ±0.05% to ±0.08% uncertainty for measurements up to about 6 kW.

11.7.3 Energy meters

A *solid state* single-phase meter is claimed to fulfil the specifications of its traditional electromechanical counterpart while providing a reduced uncertainty of measurement coupled with improved reliability at a similar cost. The instrument should provide greater flexibility for reading, tariff and data communication of all kinds, e.g. instantaneous power information, totalised energy, credit limiting and multiple tariff control. The instrument is less prone to fraudulent abuse.

11.7.4 Signal generators

Signal generators provide sinusoidal, ramp and square-wave output from fractions of 1 Hz to a few megaherz, although the more accurate and stable instruments generally cover a more restricted range. As voltage sources, the output power is low (100 mW–2 W). The total harmonic distortion (i.e. r.m.s. ratio between harmonic voltage and fundamental voltage) must be low, particularly for testing-quality amplifiers, e.g. 90 dB rejection is desirable and 60 dB (10^3:1) is normal for conventional oscillators. Frequency selection by dials may introduce ±2% uncertainty, reduced to 0.2% with digital dial selection which also provides 0.01% reproducibility. Frequency-synthesiser function generators should be used when a precise, repeatable frequency source is required. Amplitude stability and constant value over the whole frequency range is often only 1–2% for normal oscillators.

11.7.5 Electronic analysers

Analysers include an important group of instruments which operate in the frequency domain to provide specialised analyses, in terms of energy, voltage, ratio, etc., (a) to characterise networks, (b) to decompose complex signals which include analogue or digital intelligence, and (c) to identify interference effects and non-linear cross-modulations throughout complex systems.

The frequency components of a complex waveform (Fourier analysis) are selected by the sequential, or parallel, use of analogue to digital filter networks. A frequency-domain display consists of the simultaneous presentation of each separate frequency component, plotted in the x-direction to a base of frequency, with the magnitude of the y-component representing the parameter of interest (often energy or voltage scaled in decibels).

Instruments described as network, signal, spectrum or Fourier analysers as well as digital oscilloscopes have each become so versatile, with various built-in memory and computational facilities, that the distinction between their separate objectives is now imprecise. Some of the essential generic properties are discussed below.

11.7.5.1 Network analysers

Used in the design and synthesis of complex cascaded networks and systems. The instrument normally contains a swept-frequency sinusoidal source which energises the test system and, using suitable connectors or transducers to the input and output ports of the system, determines the frequency-domain characteristic in 'lumped' parameters and the transfer function equations in both magnitude and phase: high-frequency analysers often characterise networks in 'distributed' s-parameters for measurements of insertion loss, complex reflection coefficients, group delay, etc. Three or four separate instruments may be required for tests up to 40 GHz. Each instrument would have a wide dynamic range (e.g. 100 dB) for frequency and amplitude, coupled with high resolution (e.g. 0.1 Hz and 0.01 dB) and good accuracy, often derived from comparative readings or ratios made from a built-in reference measurement channel.

Such instruments have wide applications in research and development laboratories, quality control and production testing—and they are particularly versatile as an automatic measurement testing facility, particularly when provided with built-in programmable information and a statistical storage capability for assessment of the product. Data display is by $X-Y$ recorder and/or cathode-ray oscilloscope (c.r.o.) and the time-base is locked to the swept frequency.

11.7.5.2 Spectrum analysers

Spectrum analysers normally do not act as sources; they employ a swept-frequency technique which could include a very selective, narrow-bandwidth filter with intermediate frequency (i.f.) signals compared, in sequence, with those selected from the swept frequency; hence, close control of frequency stability is needed to permit a few per cent resolution within, say, a 10 Hz bandwidth filter. Instruments can be narrow range, 0.1 Hz–30 kHz, up to wide range, 1 MHz–40 GHz, according to the precision required in the results; typically, an 80 dB dynamic range can have a few hertz bandwidth and 0.5 dB amplitude accuracy. Phase information is not directly available from this technique. The instrument is useful for tests involving the output signals from networks used for frequency mixing, amplitude and frequency modulation or pulsed power generation. A steady c.r.o. display is obtained from the sequential measurements owing to digital storage of the separate results when coupled with a conventional variable-persistence cathode-ray tube (a special cathode-ray tube is not required for a digital oscilloscope): the digitised results can, in some instruments, be processed by internal programs for (a) time averaging of the input parameters, (b) probability density and cumulative distribution, and (c) a comprehensive range of statistical functions such as average power or r.m.s. spectrum, coherence, autocorrelation and cross-correlation functions, etc.

11.7.5.3 Fourier analysers

Fourier analysers use digital signalling techniques to provide facilities similar to the spectrum analysers but with more flexibility. These 'Fourier' techniques are based upon the calculation of the discrete Fourier transform using the fast Fourier transform algorithm (which calculates the magnitude and phase of each frequency component using a group of time-domain signals from the input signal variation of the sampling rate); it enables the long measurement time needed for very-low-frequency ($\ll$ 1 Hz) assessments to be completed in a shorter time than that for a conventional swept measurement—together with good resolution (by digital translation) at high frequencies. Such specialised instruments are usable only up to 100 kHz but they are particularly suitable for examining low-frequency phenomena such as vibration, noise transmission through media and the precise measurement of random signals obscured by noise, etc.

11.7.6 Data loggers

Data loggers consist of an assembly of conditioning amplifiers which accept signals from a wide range of transducers or other sources in analogue or digital form; possibly a small computer program will provide linearisation and other corrections to signals prior to computational assessments (comparable to some features of a spectrum or other class of analyser) before presenting the result in the most economical manner (such as a graphical display). The data logger, being modular in construction, is very flexible in application; it is not limited to particular control applications and by the nature of the instrument can appear in various guises.

11.8 Oscilloscopes

The cathode-ray oscilloscope (c.r.o.) is one of the most versatile instruments in engineering. It is used for diagnostic testing and monitoring of electrical and electronic systems, and (with suitable transducers) for the display of time-varying phenomena in many branches of physics and engineering. Two-dimensional functions of any pair of repetitive, transient or pulse signals can be developed on the fluorescent screen of a c.r.o.

These instruments may be classified by the manner in which the *analogue* test signals are conditioned for display as (a) the 'analogue c.r.o.', using analogue amplifiers, and (b) the 'digital storage oscilloscope', using A/D converters at the input with 'all-digit' processing.

11.8.1 Cathode-ray tube

The cathode-ray tube takes its name from Pluecker's discovery in 1859 of 'cathode rays' (a better name would be 'electron beam', for J. J. Thomson in 1897 showed that the 'rays' consist of high-speed electrons). The developments by Braun, Dufour and several other investigators enabled Zworykin in 1929 to construct the essential features of the modern cathode-ray tube, namely a small thermionic cathode, an electron lens and a means for modulating the beam intensity. The block diagram (*Figure 11.12*) shows the principal components of a modern cathode-ray oscilloscope. The tube is designed to focus and deflect the beam by means of structured electric field patterns. Electrons are accelerated to high speed as they travel from the cathode through a potential rise of perhaps 15 kV. After being focused, the beam may be deflected by two mutually perpendicular X and Y electric fields, locating the position of the fluorescing 'spot' on the screen.

11.8.1.1 Electron gun and beam deflection

The section of the tube from which the beam emerges includes the heater, cathode, grid and the first accelerating anode; these collectively form the *electron gun.* A simplified diagram of the connections is given in *Figure 11.13.* The electric field of the relatively positive anode partly penetrates through the aperture of the grid electrode and determines on the cathode a disc-shaped area (bounded by the -15 kV equipotential) within which electrons are drawn away from the cathode: outside this area they are driven back. With a sufficiently negative grid potential the area shrinks to a point and the beam current is zero. With rising grid potential, both the emitting area and the current density within it grow rapidly.

The diagram also shows the electron paths that originate at the emitting area of the cathode. It can be seen that those which leave the cathode at right angles cross the axis at a point not far from the cathode. This is a consequence of the powerful lens action of the strongly curved equipotential surfaces in the grid aperture. But, as electrons are emitted with random thermal velocities in all directions to the normal, the 'cross-over' is not a point but a small *patch*, the image of which appears on the screen when the spot is adjusted to maximum sharpness.

The conically shaped beam of divergent electrons from the electron gun have been accelerated as they rise up the steep potential gradient. The electrical potentials of the remaining nickel anodes, together with the post-deflection accelerating anode formed by the graphite coating, are selected to provide

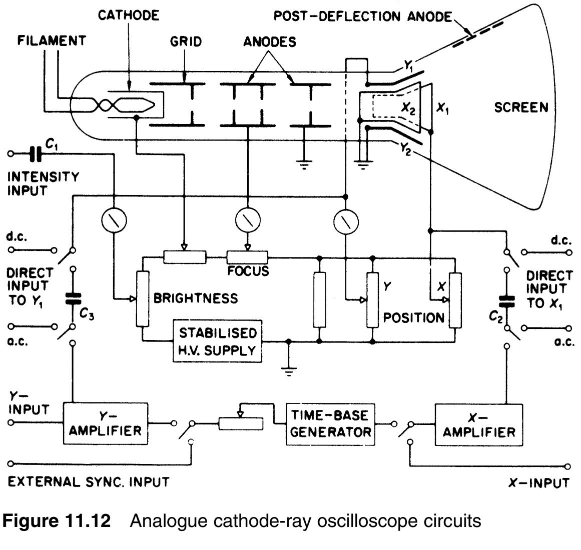


Figure 11.12 Analogue cathode-ray oscilloscope circuits

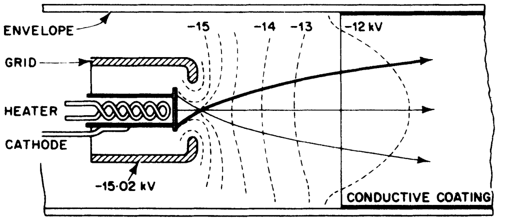


Figure 11.13 Electron beam in an electron gun

additional acceleration and, by field shaping, to refocus the beam to a spot on the screen. To achieve sharply defined traces it is essential that the focused spot should be as small as possible; the area of this image is, in part, dictated by the location and size of the 'point source' formed in front of the cathode. The rate at which electrons emerge from the cathode region is most directly controlled by adjustment of the negative potential of the grid; this is effected by the external *brightness control*. A secondary result of changes in the brightness control may be to modify, slightly, the shape of the point source. The slight de-focusing effect is corrected by the *focus control*, which adjusts an anode potential in the electron lens assembly.

11.8.1.2 Performance (electrostatic deflection)

If an electron with zero initial velocity rises through a potential difference V, acquiring a speed u, the change in its kinetic energy is $Ve = \frac{1}{2}mu^2$, so that the speed is

$$u = \sqrt{(2Ve/m)} \simeq 6 \times 10^5 \sqrt{V}\ \mathrm{m/s}$$

using accepted values of electron charge, e, and rest-mass, m. The performance is dependent to a considerable extent on the beam velocity, u.

The deflection sensitivity of the tube is increased by shaping the deflection plates so that, at maximum deflection, the beam grazes their surface (Y_1Y_2; *Figure 11.12*). The sensitivity and performance limits may be assessed from the *idealised* diagrams (and notation) in *Figure 11.14*. Experience shows that a good cathode-ray tube produces a beam-current density J_s at the centre of the luminous spot in rough agreement with that obtained on theoretical grounds by Langmuir, i.e.

$$J_s = kJ_c\theta^2 V/T \qquad (11.1)\Leftarrow$$

where J_c is the current density at the cathode, V is the total accelerating voltage, T is the absolute temperature of the cathode emitting surface, θ is one-half of the beam convergence angle, and k is a constant. If i_b is the beam current, then the apparent diameter of the spot is given approximately by $\sqrt{(i_b/J_s)}$.

The voltage V_d applied to the deflection plates produces an angular deflection δ given by

$$V_d = \tfrac{1}{2}\delta V d/L_d \qquad (11.2)$$

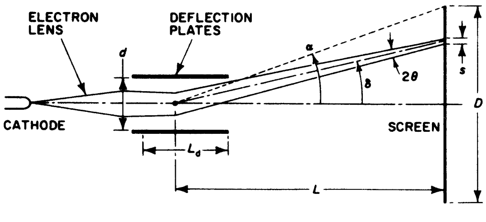


Figure 11.14 Deflection parameters

Deflection is possible, however, only up to a maximum angle α, such that

$$\alpha L_d + \theta L = \tfrac{1}{2}d \qquad (11.3)$$

From these three expressions, certain quantities fundamental to the rating of a cathode-ray tube with electric field deflection can be derived.

11.8.1.3 Definition

This is the maximum sweep of the spot on the screen, $D = 2\alpha L$, divided by the spot diameter, i.e. $N = D/s$. In television N is prescribed by the number of lines in the picture transmission. In oscillography N is between 200 and 300. Better definition can be obtained only at the expense of deflection sensitivity.

11.8.1.4 Specific deflection sensitivity

This is the inverse of the voltage that deflects the spot by a distance equal to its own diameter, i.e. the voltage which produces the smallest perceptible detail. The corresponding deflection angle is $\delta_0 = s/L$. Use of equations (11.1)–(11.3) gives for the specific deflection sensitivity

$$S = k_1 L/[(L/L_d + \alpha/\theta)\sqrt{(Vi_b)}] \qquad (11.4)\Leftarrow$$

where k_1 is a constant. Thus, the larger the tube dimensions the better the sensitivity at the smallest beam power consistent with adequate brightness of trace.

11.8.1.5 Maximum recording speed

The trace remains visible so long as its brightness exceeds a certain minimum. The brightness is proportional to the beam power density VJ, to the luminous efficiency σ of the screen, and to the time s/U during which the beam sweeps at speed U over one element of the screen. Hence,

$$U = k_2 sJV\sigma \qquad (11.5)\Leftarrow$$

Recording speeds up to 40 000 km/s have been achieved with sealed-off tubes, permitting single-stroke photography of the trace.

11.8.1.6 Maximum specified recording speed

Unless compared with the spot size s, the value of U does not itself give adequate information. The specific speed $U_s = U/s$ is more useful: it is the inverse measure of the shortest time-interval in which detail can be recorded. Combining equations (11.5) and (11.1)

$$U_s = k_2 JV\sigma = k_3\sigma(V\theta)^2 \qquad (11.6)\Leftarrow$$

The product $\sigma(V\theta)^2$ is the essential figure in the speed rating of a cathode-ray tube.

11.8.1.7 Limiting frequency

A high recording speed is useful only if the test phenomena are recorded faithfully. The cathode-ray tube fails in this respect at frequencies such that the time of passage of an electron through the deflection field is more than about one-quarter of the period of an alternating quantity. This gives the condition of maximum frequency f_m:

$$f_m = 1.5 \times 10^5 \sqrt{V}/L_d \qquad (11.7)\Leftarrow$$

which means that the smallest tube at the highest acceleration voltage has the highest frequency limit. This conflicts with the condition for deflection sensitivity. Combining the two expressions gives the product of the two:

$$(Sf_{\mathrm{m}}) = k_4/[1 + (L_{\mathrm{d}}/L)(\alpha/\theta)]\sqrt{i_{\mathrm{b}}} \qquad (11.8)\Leftarrow$$

Compromise must be made according to the specific requirement. It may be observed that the optimum current i_{b} is not zero as might appear, because for such a condition the convergence angle θ_c is also zero. In general, the most advantageous current will not differ from that for which the electron beam just fills the aperture.

Limitation of cathode-ray-tube performance by space charge is an effect important only at low accelerating voltages and relatively large beam currents. For these conditions the expressions above do not give practical ratings.

11.8.1.8 Fluorescent screen

The kinetic energy of each incident electron is converted into light owing to the exchange of energy which occurs within the phosphors coated on the inside of the screen. The result, *luminescence*, is usually described as either fluorescence or phosphorescence. *Fluorescence* is the almost instantaneous energy conversion into light, whereas *phosphorescence* is restricted to the phosphors which 'store' the converted energy and release it, as light, after a short delay. The *afterglow* can last many seconds. The respective terms are applied to the phosphors in accordance with their dominant light-emitting behaviour; however, each property is always present in a phosphor. In most cathode-ray tubes fluorescent phosphors are used with a very short afterglow (40 µs); a longer afterglow would blur a quickly changing display.

The most common fluorescent substances are silicates and sulphides, with calcium tungstate for photographic work. An excellent zinc silicate occurs naturally in the mineral willemite, but all other substances are prepared artificially. After purification, the substances are activated by the addition of very small quantities of suitable metals, and then ground to a powder of grain size 5–10 µm (for sulphides 10–50 µm).

In applying the powder to the glass several methods are used. Very finely ground silicates can be applied without a binder, by setting the powder from water, or an electrolyte. Sulphides are usually compounded with the glass wall by means of a binder such as sodium silicate; good results have been obtained also by baking them directly into the glass. In choosing the method two conflicting considerations must be taken into account. Baking or compounding gives good thermal contact, and prevents overheating of the powder particles by the electron beam. But at the same time good optical contact is established with the glass, with the result that light from the powder enters the glass also at angles exceeding the total reflection angle. This part of the light cannot escape at the viewing side, but is reflected and forms a luminous 'halo' which blurs the picture.

While silicates fluoresce usually green or blue-green, white fluorescence can be obtained by mixing cadmium sulphide and zinc cadmium sulphide in suitable proportions. The luminous efficiency of modern fluorescent powders is very high. Zinc silicate emits up to 2.5 cd/W of beam power, and some sulphides are even more effective. The luminous efficacy increases with voltage up to about 10 kV, beyond which it decreases, partly because fast electrons are not stopped by the crystal grains.

The range of phosphors is classified by 'P' numbers, as in *Table 11.4*.

The light persistence of the most widely used phosphor (P31) is 40 µs; however, there are tubes available, using this phosphor, which have both variable persistence and 'mesh storage' facilities. Adjustable persistence enables very-low-frequency signals, lasting for a minute, to be viewed as a complete trace while retaining the normal 'writing' properties of the tube. Storage, at reduced brightness, is possible for up to 8 h.

Phosphor materials are insulators. The current circuit for all the incident beam electrons is completed by an equal number of low-energy electrons being attracted along various paths through the vacuum to the positive conducting-graphite coating. The result of secondary emission of electrons is to leave the screen with a positive charge only a little lower in potential than that of the graphite coating; thus, the surplus secondary emission electrons return to the phosphor-coated screen surface. This mechanism works perfectly at the voltages usually employed in oscillography and television (3–20 kV), also in small projection-tube sealed-off high-speed oscillographs (10–25 kV). Above 25 kV the mechanism is unsatisfactory and the dispersion of charge has to be supplemented by other means.

The power concentration in the luminous spot of a cathode-ray tube is very high. In direct-vision television tubes the momentary luminance may exceed 10 cd/mm^2. A 3 mA 80 kV beam of a large-screen projection tube may be concentrated in an area of 0.1 mm^2, representing a power density of 2 kW/mm^2. (By comparison, melting tungsten radiates about 10 W/mm^2.) The width of the trace in a conventional 10 kV cathode-ray tube with a beam current of 10 µA is only 0.35 mm.

11.8.2 Deflection amplifiers (analogue c.r.o.)

Amplifiers are built into cathode-ray oscilloscope (c.r.o.) equipment to provide for time base generation and the amplification of small signals. The signals applied to the vertical (*Y*) deflecting plates are often repetitive time functions. To display them on a time base, the vertical displacements must be moved horizontally across the screen at uniform speed repeatedly at intervals related to the signal frequency, and adjusted to coincidence by means of a synchronising signal fed to the time base from the *Y* amplifier.

Table 11.4 Classification of phosphors

Phosphor	*Persistence*	*Colour of trace*	*Relative brightness*	*Application*
P2	Medium/short	Yellow/green	6.5	Low repetition rates (general oscillography)
P4	Medium	White	5.0	High-contrast displays (monochrome television)
P7	Long	White or yellow/green	4.5	Long-persistence low-speed (radar)
P11	Medium/short	Blue	2.5	High writing speeds (photography)
P31	Medium/short	Green	10	High-brightness, general use

An additional constraint may be the requirement to start the time base from the left side of the screen at the instant the Y-plate signal is zero: a synchronising trigger signal is fed to the time base circuit to initiate each sweep. Finally, the spot must return (flyback) to the origin at the end of each sweep time. Ideally, flyback should be instantaneous: in practice it takes a few score nanoseconds. During flyback the beam is normally cut off to avoid return-trace effects, and this is the practice for television tubes.

Linear time bases are described in terms either of the sweep time per centimetre measured on the horizontal centre line of the graticule, or of the frequency. The repetition frequency range is from about 0.2 Hz to 2.5 MHz.

11.8.2.1 Ramp voltage (saw-tooth) generator

If a constant current I is fed to a capacitor C, its terminal voltage increases linearly with time. An amplified version of the voltage is applied to the X plates. Various sweep times are obtained by selecting different capacitors or other constant current values.

The method shown in *Figure 11.15* employs a d.c. amplifier of high gain $-G$ with a very large negative feedback coupled through an ideal capacitor C_f. The amplifier behaves in a manner dictated by the feedback signal to node B rather than by the small direct input voltage E_s. The resultant signal v_i is small enough for node B to be considered as virtually at earth potential. Analysing the network based on the 'virtual earth' principle, then $v_i = 0$ and $i_i = 0$, the input resistance R_i being typically 1 MΩ. Then $i_s + i_f = 0$, whence

$$(E_s/R_s) = C_f(\mathrm{d}v_o/\mathrm{d}t) = 0 \text{ giving } v_o = -(1/C_f R_s)\int E_s \cdot \mathrm{d}t$$

The output voltage v_o increases linearly with time to provide the ramp function input to the X amplifier.

11.8.2.2 Variable delay

When two similar X amplifiers, A and B, are provided, B can be triggered by A; if B is set for, say, five times the frequency of A and this had a variable delay facility, then any one-fifth of the complete A waveform display can be extended to fill the screen. Both complete and expanded waveforms can be viewed simultaneously by automatic switching of A and B to the X plates during alternate sweeps.

11.8.2.3 Signal amplifiers

The X- and Y-plate voltages for maximum deflection in a modern 10 cm cathode-ray tube may lie between 30 and 200 V. As many voltage phenomena of interest have amplitudes ranging between 500 V and a few millivolts, both impedance matched attenuators and voltage amplifiers are called for. The amplifiers must be linear and should not introduce phase distortion in the working frequency range; the pulse response must be fast (e.g. 2 ns); and the drift and noise must be small (e.g. 3 μV).

Amplifiers have d.c. and a.c. inputs. The d.c. coupling is conventional, to minimise the use of capacitors (which tend to produce low-frequency phase distortion). The d.c. inputs also permit the use of a reference line. The a.c. input switch introduces a decoupling capacitor.

Two separate Y amplifiers are often provided; in conjunction either with one switched or with two separate X amplifiers, they permit the examination of two voltages simultaneously. It is important to distinguish the dual-trace oscilloscope (in which a single beam is shared by separate signal amplifiers) and the true double-beam device (with a gun producing two distinct beams).

11.8.3 Instrument selection

A general purpose c.r.o. must compromise between a wide frequency range and a high amplification sensitivity, because high gain is associated with restricted bandwidth. An economically priced c.r.o. with solid state circuitry may have a deflection sensitivity of 5 mV/cm, bandwidth 10 MHz and a 3 kV accelerating voltage for a 10 cm useful display. For greater flexibility, many modern c.r.o.s comprise a basic frame unit with plug-in amplifiers; the frame carries the tube, all necessary power outlets and basic controls; the plug-in units may have linear or special characteristics. The tube must be a high-performance unit compatible with any plug-in unit. The cost may be higher than that of a single function c.r.o., but it may nevertheless be economically advantageous.

Many classes of measurement are catered for by specially designed oscilloscopes. In the following list, the main characteristic is given first, followed by the maximum frequency, deflection sensitivity and accelerating voltage, with general application notes thereafter.

Low-frequency (30 MHz, or 2 MHz at high gain; 0.1 mV/cm–10 V/cm; 3–10 kV). General purpose oscillography, for low-frequency system applications. Usually, high X-sweep speeds are needed for transient displays and complex transducer signals. Two traces available with two-beam or chopped single-beam tube. Single-shot facility. Y-amplifier signal delay useful for 'leading-edge' display. Rise times 10–30 ns.

Medium-frequency (100 MHz; 0.1 mV/cm; 8–15 kV). General-purpose, including wide-band precision types with a rise time of 3–10 ns.

High-frequency (non-sampling) (275 MHz; 10 mV/cm; 20 kV). May include helical transmission-line deflecting

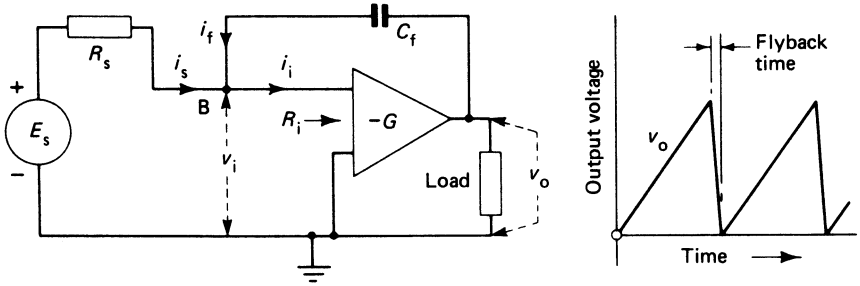


Figure 11.15 Time-base operational amplifier

plates. Rise time 1.5 ns for testing digital system and computer displays.

Variable persistence storage cathode-ray tube for h.f. transients.

Very-high frequency (sampling) (12 000 MHz; 5 mV/cm; 20 kV). The v.h.f. waveforms are 'instantaneously' measured and optically stored, *n* times at *n* successively retarded instants during the period of each of *n* successive waveforms (with *n* typically 1000). The *n* spots accumulate in the correct time sequence as a single waveform derived from the *n* waveforms sampled.

Large-screen (20 MHz; 60 mV/cm; 28 kV). Screen area up to 0.25 m^2. Useful for industrial monitoring, classroom display and computer graphics. (Magnetic deflection gives a shorter overall length of tube, probably with improved focusing towards the edges but with a lower maximum frequency.)

11.8.3.1 Other features

With all c.r.o.s, the X–Y cartesian coordinate display can be augmented by Z-modulation. This superimposes a beam-intensity control on the $X-Y$ signals to produce the illusion of three dimensions. The beam control may brighten one spot only, or gradually 'shape' the picture.

Oscilloscopes are supplemented by such accessories as special probes for matching and/or attenuating the signal to the signal/cable interface. Anti-microphony cables are available which minimise cable-generated noise and capacitance change. Many types of high-speed camera can be used, including single-shot cameras that can employ ultraviolet techniques to 'brighten' a record from a trace of low visual intensity.

11.8.4 Operational use

To avoid degradation of the signal source, a c.r.o. has a high input impedance (1 MΩ, 20 pF) and a good common-mode rejection (50 dB). The quality can be considerably impaired by impedance mismatch and incorrect signal/earth (or ground) connections.

11.8.4.1 Impedance matching

Voltage signals are often presented to a c.r.o. through wires or coaxial cables terminated on crocodile clips. Although this may be adequate at low frequencies, it can behave like a badly matched transmission line, causing attenuation distortion and phase distortion. The earth screen of a coaxial cable is not fully effective, and in consequence the cable can pick up local radiated interference. In a normal cable the central conductor is not bonded to the insulation, and flexure of the cable may distort the signal waveform because of changes in the capacitance distribution, or generated noise. The cable capacitance of about 50 pF/m is additive to the c.r.o. input capacitance, and will distort the signal, particularly if the frequency and source impedance are high.

These adverse effects can be mitigated by use of fully screened probes and cables of low and constant capacitance associated with low loss. The probes may be *active* (including an impedance matching unity-gain amplifier) or *passive* (including resistance with adjustable capacitance compensation); attenuators are incorporated in each to reduce both the signal level and the capacitive loading of the source, the compensation being most effective for high-impedance sources. The active probe can be used to match the source to the standard 50 Ω input impedance of a high-frequency c.r.o. (100 MHz and above), an important matter in that a quarter wavelength at 100 MHz is about 75 cm, and severe standing-wave effects may otherwise occur to invalidate the measurements.

11.8.4.2 Earths and grounds

Ground planes (e.g. extended screens) and earth planes (e.g. trunking) constitute low-impedance paths through which interfering signals can circulate, of a wide variety of frequencies and waveforms. The arbitrary connection of screens to earth and ground terminals can have the unwelcome result of increasing noise, due to earth-loop currents. The latter must not enter the input signal lines of either the system or the c.r.o.; and the interference can be minimised by using a single ground/earth connection to serve the whole equipment, if this is possible. The common connection should be made at the zero-potential input signal lead.

11.8.5 Calibration

Operational specifications will include performance statements, typically: voltage accuracy, ±3%; time-base frequency ±3%; Y-amplifier linearity, ±3%, phase shift, 1°; pulse response, $\not>5\%$ after 2 ns; sweep delay accuracy, ±2%. The calibration facility provided usually consists of a 1 or 2 kHz square waveform with a 3 μs rise time, with both the frequency and voltage amplitude within ±1%.

The practical application of the c.r.o. as a test instrument includes (a) the general diagnostic display of waveforms, and (b) the testing of a product during development or manufacture in relation to its accuracy and quality control. For (a) the built-in calibration facility is adequate; but for (b) it is prudent (or essential) to make regular checks of the actual parameters of the self-calibration facility and of the basic characteristics of the instrument. Such tests would include assessments of the accuracy on alternating voltages, and the ratio and phase angle of the amplifiers and attenuators over the frequency range. The most important and critical test for a c.r.o., which will inevitably be used with modern digital networks, is its response to high frequency square waveforms. The rise time is quite easily measured, but the initial resonant frequency of the 'flat' part of the waveform and its rate of attenuation are important in assessing the actual duration of this part.

The quality and usefulness of these tests is ultimately related to the readability of the display, which would be no better than ±0.3% in a 10 cm trace. The quality of the standard reference devices would, for example, be ±0.1% for a variable-frequency alternating voltage source; hence, there would be a total measurement uncertainty of about 0.4% in the ±1% waveform. It follows that the actual c.r.o. voltage source must be correct to within 0.6% if it is to be used with confidence as a ±1% source. It is a fundamental requirement that the quality of all the reference standards be known from recent linearity and traceability tests.

11.8.6 Applications

The principal application is for diagnostic testing. In addition, the c.r.o. is used to monitor the transfer characteristics of devices and systems, as well as for numerical measurements. Among special uses are to be found the television camera, electron diffraction camera and electron microscope.

Single-deflection measurements The tube functions with a single pair of deflecting plates as a voltmeter or ammeter, with the advantages of being free from damping, unaffected by change of frequency or temperature or over-deflection, and imposing only a minute load on the test circuit. Examples are the monitoring of signals in broadcast and recording studios, modulation checks, indication of motor peak starting currents, thickness measurement, null detection and the measurement of current and voltage in networks of very low power.

Differential measurements Similar or related phenomena are compared by use of both pairs of deflecting plates, the resulting Lissajous figure being observed. In this way phase difference can be measured, frequencies accurately checked against a standard, armature windings tested, modulation indicated by a stationary waveform, and the distortion measured in receivers, amplifiers and electro-acoustic equipment. *Figure 11.16* is a Lissajous figure comparing two frequencies by counting the horizontal and vertical loops, giving the ratio 4/5. The diagram in *Figure 11.17* shows the connections for testing the distortion of an amplifier: the oscillator is set to a known frequency, the two pairs of deflectors connected, respectively, to the input and output (with suitable attenuators). A straightline display indicates absence of distortion, while curvature or looping indicates distortion.

Repetitive time base Using a linear time base enables sustained waveforms to be displayed for machines, transformers and rectifiers, rapid operations to be timed, and surge phenomena to be shown with the aid of a recurrent-surge generator.

Single sweep Non-repetitive transients, such as those arising by lightning or switching, are most readily traced by use of the continuously evacuated tube, but sensitive sealed-off tubes can also be employed. In either case a non-repetitive time base is used consisting of the voltage of a capacitor discharged through a resistor and triggered by a gap. The transient discharge is normally obtained by use of a surge generator for power system plant testing, but other single transients, such as the small voltages of cardiac origin, may also require to be recorded.

Independent bases The c.r.o. is well suited to the display of quantities related by some variable other than time. The current due to an impressed voltage can be displayed on a frequency base as a resonance curve, radio receivers aligned, and B/H loops taken for steel samples. Transistor curve tracers are useful applications capable of giving a complete family of characteristics in a single display.

Modified time bases There are several modified forms of time base. If two voltages of the same amplitude and frequency are displaced in time phase by a quarter-period and applied, respectively, to the X and Y plates as in *Figure 11.18*, they will produce a circular trace. One application, in decade digital logic circuits, involves a Z modulation of the c.r.o. grid. To check that a frequency divider section is working correctly, the period of the circular time base is made 10 times that of the next higher count rate section. With the latter used to modulate, 10 equal bright and dark peripheral sectors should appear on the trace.

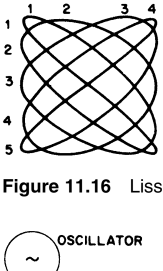


Figure 11.16 Lissajous figure

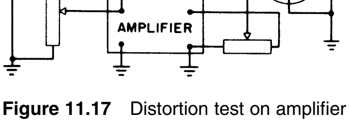


Figure 11.17 Distortion test on amplifier

11.8.7 Digital storage oscilloscopes

Digital storage oscilloscopes (d.s.o.) (*Figure 11.19*) offer advantages over the analogue c.r.o. in some circumstances:

(1) they allow long-term display of a transient quantity;
(2) it is easy to produce a hard copy of the display;
(3) they allow computation and signal processing within the d.s.o.;
(4) they allow easy transfer of data to a computer; and
(5) they use cheap cathode-ray-tube construction because the deflection is not at signal frequency but much lower than the maximum signal frequency.

The d.s.o. has become available for laboratory and industrial use because of the development of relatively cheap, accurate and fast A/D converters.

11.8.8 Digital oscilloscope characteristics

11.8.8.1 Sampling rate

Clearly the sampling rate must be high enough to give a faithful representation of the applied signal. Nyquist's theorem states that a periodic signal must be sampled at more than twice the highest frequency component of the signal. In practice, because of the finite time available, a sample rate somewhat higher than this is necessary. A sample rate of 4 per cycle at oscilloscope bandwidth would be typical. Therefore, for single shot, a digital oscilloscope specification will give a bandwidth of 100 MHz or a sample rate of 400 megasamples/s. If the display is a dot display, then something like 20 samples per cycle is necessary if the eye is to interpret the display easily. Usually some form of interpolation is provided to aid interpretation. The simplest method is to draw straight lines between successive sample parts. Other interpolation methods are used such as sine interpolation, designed for fairly pure sine waves. This interpolation will give erroneous displays for non-sinusoidal inputs.

All this means that great care must be taken in interpreting a display when the input approaches the oscilloscope bandwidth, particularly if it has significant high frequency components such as a square wave. This is also true of course for an analogue c.r.o.

The d.s.o. in repetitive mode can build up the display over a number of sections of the displayed waveform, and so the sampling-rate problems are not so severe. Specifications will give bandwidth, single-shot bandwidth and sampling rate. If single-shot mode is to be used, then care must be taken to

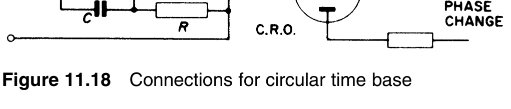


Figure 11.18 Connections for circular time base

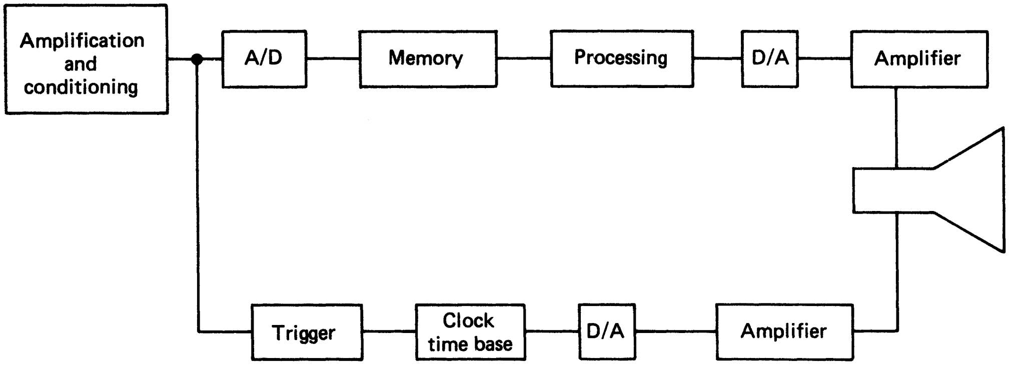


Figure 11.19 The digital storage oscilloscope

look for the adequacy of the single-shot bandwidth or the sampling rate.

For an analogue c.r.o., a rule of thumb relates the rise time of the c.r.o. to its bandwidth rise time (s) × bandwidth (Hz) = 0.35. So, a 100 MHz oscilloscope has a rise time of 3.5 ns. A signal rise time approaching this cannot be faithfully displayed, as the fastest displayed rise time is that of the instrument itself. Usually the fastest rise time displayable is 8 times the c.r.o. rise time. So the fastest signal rise time = 2.8 ÷ bandwidth. For a 100 MHz c.r.o. this would be 28 ns.

A similar restriction occurs for the d.s.o. A fast signal pulse on an expanded time scale is shown in *Figure 11.20*. The worst case giving the slowest rise time is when one of the samples is in the middle of the edge. The time from 10% to 90% of the rise (the rise time) is given by (80/100) × 2 sample periods (= 1.6 sample periods).

So for a 100 megasample/s d.s.o., the sampling interval is 10 ns and the displayed edge, regardless of the rise time of the applied signal, is at worst 16 ns. Again, in order to faithfully display the signal, and not the characteristics of the d.s.o., the fastest signal rise time should be about 8 times this limit, say 10 sample intervals. So, as a rule of thumb:

$$\text{Fastest signal rise time} = 10 \text{ sample intervals} = \frac{10}{\text{d.s.o. sampling rate}}$$

e.g. for 100 megasamples/s d.s.o. the fastest signal rise time is 100 ns. For a signal rise time faster than this it is likely that the performance of the d.s.o. will show itself in the display of the signal.

11.8.8.2 Vertical resolution

Vertical resolution is another important parameter of the d.s.o. An important contribution to this is the number of bits of the A/D converter. The more bits, the greater the resolution. The minimum discernible level is

$$\frac{\text{Input voltage}}{2^n - 1}$$

where n is the number of bits of the A/D.

For example, for an input voltage of 1 V and an 8-bit A/D, the minimum discernible level is

$$\frac{1}{2^8 - 1} = \frac{1}{255}\ \text{V}\ (\simeq 4\,\text{mV})$$

This value is sometimes given as a percentage and in this case would be ±0.2%.

So on a 1-V signal, noise of less than 4 mV will not be displayed. Increasing the number of bits in the A/D converter will help, but internally generated noise limits the advantage achieved. This leads to the concept of effective bits of the d.s.o., which is performance of a hypothetical idealised A/D converter which gives the same vertical resolution of the actual A/D converter with its noise limitation.

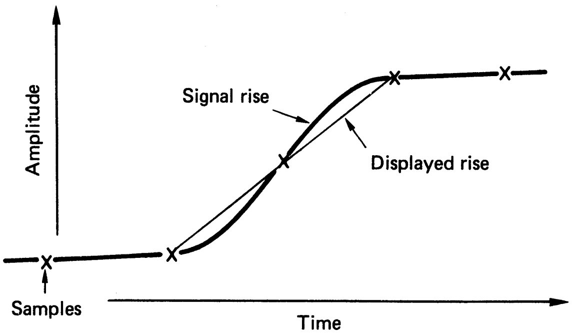


Figure 11.20 The effect of sampling rate on displayed rise time

11.9 Potentiometers and bridges

11.9.1 D.c. potentiometers

D.c. potentiometers may be used to measure a direct voltage by comparison with the known ratio of the e.m.f. of a Weston standard cell or electronic solid state reference device. The methods of achieving the linear ratio between the unknown p.d. and the standard-cell e.m.f. are by successive approximation to the balance condition (a) using groups of matched resistors, (b) using opposed m.m.f.s between matched windings and current-ratio resistors (comparator potentiometer), and (c) pulse-width modulation (ratio of time interval). The d.c. potentiometer is a ratio device of high precision with the uncertainty of the standard voltage source limiting the ultimate accuracy.

11.9.1.1 Weston standard cell

The highest quality saturated mercury/cadmium-mercury cell has an e.m.f. of about 1.018 620 V at 20°C. The net e.m.f./temperature coefficient is about − 40 μV/°C, arising from the two different coefficients of the limbs, each of the order of 350 μV/°C. It is, therefore, important to avoid a temperature differential across the cell by mounting it in oil or air within a constant temperature enclosure (e.g. for 20°C ±10 m°C, the uncertainty can be only ±0.4 μV).

Standard cells are stable, the e.m.f. falling by, perhaps, 1 μV per year. For accurate work, *regular* confirmation of the e.m.f. should be made at an accredited calibration laboratory.

The internal resistance is 600–800 Ω. The current drawn from a cell should ideally be zero: in practice it should be limited to a few nanoamperes for a few seconds.

Should a cell be compared with an electronic solid state voltage reference (below), the cell must be presented and removed while the electronic source is switched on.

11.9.1.2 Electronic solid state e.m.f. reference

These devices use the constant voltage property of Zener diodes to deliver highly stable d.c. voltages, usually at 1 V and 10 V levels for direct voltage meter calibrations, and 1.018 61 V to simulate a Weston cell for d.c. potentiometric standardisation. Up to four diodes (each stabilising at about 6.3 V) are used, having been carefully selected for stability, low noise and low temperature/voltage coefficient properties. After further ageing, these temperature-stabilised monolithic devices may be used either separately with a carefully designed operational amplifier, or connected in cascade in specialised networks to achieve the required output p.d. by potential division using highly stable resistors.

It would appear that these devices are excellent transportable standards of voltage (which should be reliable to within 2 μV/year) and they are much more convenient than Weston cells in this respect. Except for the most extreme requirements of accuracy, they could be used as laboratory standards, provided that at least three separate units are available for intercomparison. With more experience and/or minor refinement to these devices, they could shortly replace the Weston cell completely, since at present the Weston cell can be used as a transportable standard to less than 0.5 μV, but only if very great care is exercised.

Typical electronic devices can deliver, say, 5 mA, 10 V from 2 mΩ and 1.0186 V from 2kΩ.

11.9.1.3 Potentiometer principle (using groups of matched resistors)

In *Figure 11.21* AB is a manganin wire of length a little greater than 1 m, through which the current I from a 2 V secondary cell may be varied by means of R. The current is adjusted so that the p.d. between points F and D, 1.0186 m apart, balances the e.m.f. of the Weston standard cell, the balance condition being obtained by use of the switch S and galvanometer G. The volt drop along the wire is now 1 mV per mm length. The switch S may now be set to apply the unknown voltage V_x, and a new balance point H found. Then the length DH in mm represents the unknown voltage in millivolts.

This very simple arrangement is inconvenient in practice. It suffers also from the disadvantage that the precision of the voltage measurement depends upon the precision with which the point of contact of the slider is known. For example, if the length DH is known only to the nearest millimetre, the precision of the voltage measurement would, in general, be less than 1 in 1000. Much greater precision than this is often required, and to obtain it (and to secure a more compact form of apparatus) some elaboration and refinement in the circuit elements is required. Many practical forms of d.c. potentiometer have been devised.

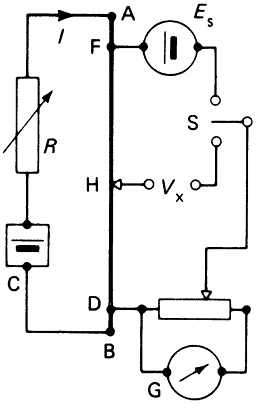


Figure 11.21 Simple potentiometer

11.9.1.4 Practical potentiometer

The potentiometer in *Figure 11.22* is supplied across AB from a 2 V secondary cell through a rheostat R. Resistors R_2, R_3 and R_4 are connected in series between A and B, but the conditions first considered are with R_4 short circuited by the switch at P_1. The resistor R_3 is divided into equal sections with 19 tappings brought out to studs, and there is an additional stud connected to a tapping on R_2.

One terminal of the standard cell is connected to a tapping on the parallel resistor R_1, and the other terminal may be connected (when the switch S is in the left-hand position

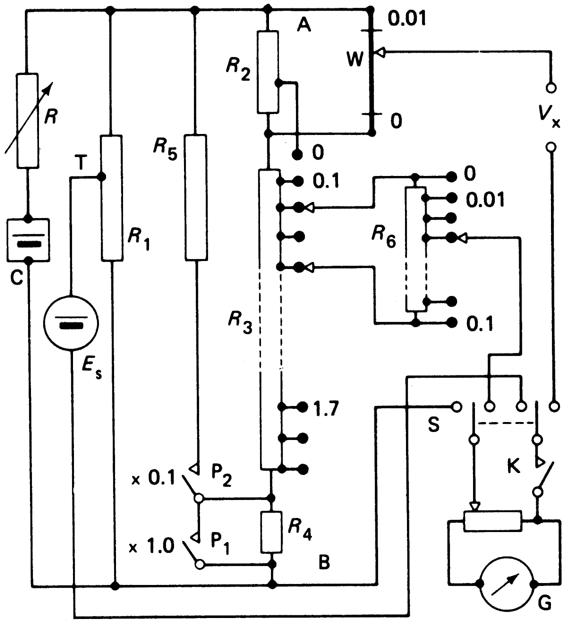


Figure 11.22 Practical potentiometer

and the key K is depressed) to the point B through galvanometer G. A balance may be obtained with the standard cell by adjusting the rheostat R, and the resistance values are such that the p.d. between successive studs on R_3 is then precisely 0.1 V, and the voltage between A and B a little over 1.8 V. In parallel with the shunting resistor R_2 is the slide wire W. The tapping on R_2 enables a true zero, and small negative readings, to be obtained on W. By suitable choice of the resistance of R_2 and W the voltage range covered by the whole slide wire is a little over 0.01 V.

R_6 is a tapped resistor with 10 equal sections and 11 studs. The total resistance of R_6 is made equal to the resistance of two sections of R_3, and R_6 always spans two sections by means of sliding contacts on the R_3 switch as shown. The total resistance of R_6 and two sections of R_3 in combination is therefore always equal to a single section of R_3 and the potential difference between successive tappings on R_6 is then 0.01 V.

The unknown voltage V_x is applied to terminals connected through the galvanometer switch S to the sliding contact on W and to the tapping on the R_6 switch. If the slide-wire dial has 100 divisions, each division represents 0.0001 V. Thus, for the switches in the positions shown, V_x might be 0.2376 V. The maximum V_x readable, assuming that there are some graduations provided beyond the 0.01 position on the slider dial, is $1.7 + 0.1 + 0.0100 = 1.8100$ V.

The resistors R_4 and R_5 are introduced into the circuit by opening switch P_1 and closing P_2; this reduces the p.d.s in the potentiometer circuits to one-tenth of those marked. One division on the slider dial would then correspond with 0.000 01 V or 10 μV.

Medium- to high-precision potentiometers include a 'standard cell setter' so that the actual e.m.f. of the standard cell can be used for standardisation and, subsequently, for rechecking the standardisation during use *without* altering the measuring dials: this is more convenient, although less accurate than standardising against the dials. The contact T could be the slider of a 10-turn voltage divider with an indicated range, e.g. of 1.018 250 to 1.018 750 V in 1-μV steps.

D.c. potentiometers of very high precision have a ratio non-linearity of ±5 parts in 10^6. A stable current of 50 mA must be provided by an electronic controller, and a high-gain low-noise detector is necessary: this may be a photocell galvanometer amplifier with a display sensitivity of 5000 mm/μV in a 10-Ω circuit.

11.9.1.5 Comparator potentiometer (using m.m.f. balance)

The comparator potentiometer is one of a group of precision bridges employing mutual-coupled inductive ratio arms. The bridges exploit the high m.m.f. linearity and discrimination of a constant current in a variable number of turns.

A simplified network is shown in *Figure 11.23*. The measuring loop M is magnetically coupled to the balancing loop B by the turns N_m and N_b, respectively, on a magnetic core, which also carries a feedback winding with N_f turns. The feedback network includes an a.c. modulated sensor which actuates an a.c./d.c. electronic control network C_b to change the direct current I_b until a zero m.m.f. condition is achieved in the core.

11.9.1.6 Operation and standardisation

If I_m and N_b are constant, then for an automatically maintained m.m.f. balance $N_m I_m = N_b I_b$ it follows that linear changes in N_m result in corresponding linear changes in I_b. As N_m has in effect a discrimination and linearity of about 1 in 10^7, these properties are imposed also on I_b. To convert this linear current scale to voltage across R_b, the instrument has to be standardised against a standard cell. The cell, of e.m.f. E_r, is connected (through S1) in opposition to the p.d. across R_m, and balance achieved by use of the adjustable direct-current controller C_m, which is then left to maintain this current level.

The equivalent cell e.m.f. $I_m R_m$ is now connected (through S2) in opposition to $I_b R_b$, and N_m turns are selected to be numerically equal to E_r. The m.m.f is balanced automatically by the feedback control system and by a small adjustment to N_b so that $I_m R_m = I_b R_b = E_r$, represented numerically by N_m. With switch S set to S_0 the unknown p.d. V_x is presented through S_x and galvanometer G_1 in opposition to $I_b R_b$ and balance achieved by manual adjustment of N_m. Then N_m is numerically the unknown voltage.

Instruments of this kind are probably the most linear d.c. potentiometers available. They incorporate additional windings that enable the linearity to be checked by the operator. While the actual values of R_m and R_b are not of first importance, it is vital that they should be constant. Resistors are incorporated for subdivision of I_m for lower decades of N_m in order to simulate the 10^7 turns implied by a discrimination of 1 in 10^7, but they do not contribute unduly to the overall non-linearity, which is about $\pm 0.3 \times 10^{-6}$ of full scale. The system uncertainty must include that of the standard cell, at least ±1–2 μV.

11.9.1.7 Pulse width modulation potentiometer[12]

This method subdivides a known stable d.c. voltage in terms of a time-period ratio. If a standard voltage source E_r is repeatedly connected to a low-pass filter through a very-high-speed semiconductor switch for a time t_1, and then disconnected for an additional time t_2, the high-frequency rectangular, chopped waveforms will, after smoothing, give an output $E_0 = E_r \cdot t_1/(t_1 + t_2)$ with $t_1 + t_2 = T$, a constant period. The time *ratio* can be precisely set using a variable time-interval counter for t_1 and a fixed-period counter for T, with each counter being driven from a common highly stable crystal-controlled oscillator. An unknown voltage V_x, in opposition to E_0 through a galvanometer, can be measured by variation of E_0 (through the voltage-scaled setting of the t_1 counter) until galvanometer zero balance is achieved. The e.m.f. E_r of the solid state source of steady voltage must be known accurately and the source must be capable of delivering small current surges into the low-pass filter without any significant change in voltage. Conversely, V_x and E_r could be interchanged when $V_x > E_r$.

One d.c. variable-voltage source uses the accurately determined voltage as input to constant gain d.c. amplifiers (designed for decade steps) to give a wide-range d.c. voltage standard source with ±50 p.p.m. uncertainty (3-month stability) up to 1.2 kV with a 25 → 10 mA (at 1 kV) current capability and 0.5 → 3.0 s settling times.

11.9.1.8 Applications

The d.c. potentiometer may be used to determine temperature by measuring the thermo-e.m.f.s in calibrated thermocouples. In conjunction with standard four-terminal resistors and shunts it is applied to the calibration of voltmeters, ammeters and wattmeters, and for the comparison of resistors. In *Figure 11.24* the values V_x, I_x, R_x and P_x are unknowns, and E_p is measured by the potentiometer. Resistors $R_1, \ldots, R_7$ have known values. At (*a*) a voltage divider is used to measure V_x, and at (*b*) a shunt is employed to measure I_x. At (*c*) R_x is compared with R_4 by means of

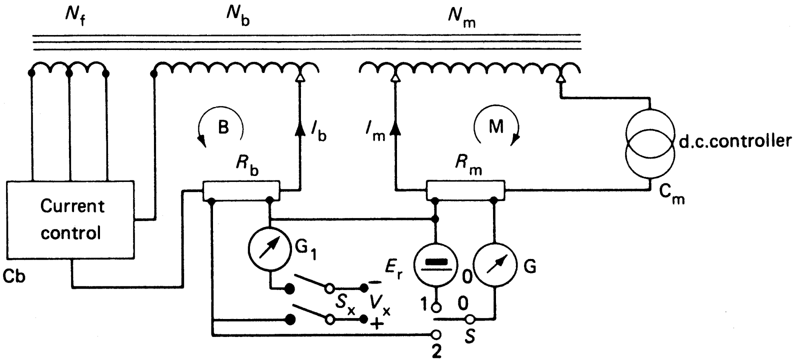


Figure 11.23 D.c. comparator potentiometer

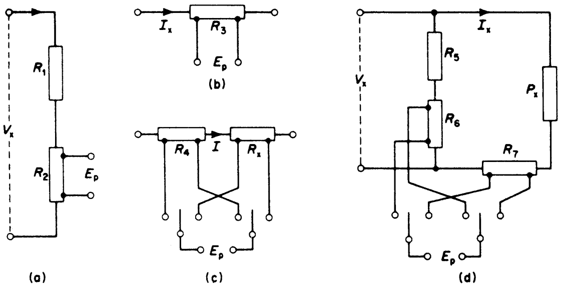


Figure 11.24 D.c. potentiometer measurement of voltage current, resistance and power

their respective p.d.s. The network (*d*) enables the power $P_x = V_x I_x$ to be determined from V_x and I_x. The resistance of four-terminal resistors is known between the potential (inner) terminals, and *not* the current (outer) terminals, which minimises errors due to non-uniform current densities and contact resistances.

The National Physical Laboratory (NPL) designed 'Wilkins' resistor is a high-stability, four-terminal, a.c./d.c. resistor available in various decade and other values (Tinsley Co., London) which has an exceptionally low time constant for accurate a.c./d.c. comparator networks, etc.

11.9.2 A.c. potentiometers

The a.c. potentiometer may be used to measure the magnitude of an alternating voltage and its phase relative to a datum waveform. The measurements are normally restricted to sinusoidal waveforms, as the presence of harmonics makes a null balance impossible; and operation is usually confined to a single frequency. One recent instrument (Yorke) is capable of measurements up to audiofrequency, with an uncertainty approaching ±0.05%. The general principles of the Larsen and Gall potentiometers are set out below.

11.9.2.1 Larsen potentiometer

The primary winding of a variable mutual inductor is connected in series with a tapped resistor to a source of voltage *V*. The unknown voltage V_x is normally derived from the same source, because identity of frequency is essential. The supply current *I* is controlled by resistor R_1 and indicated on an ammeter. The inductor secondary e.m.f. in series with a tapped fraction of the volt drop across *R* is balanced against the unknown V_x (*Figure 11.25*) by means of the vibration galvanometer G, the sensitivity of which can be varied by resistor R_2. If *M* and *R* are the mutual inductance and resistance values at balance, then

$$V_x = I\sqrt{|R^2 + \omega^2 M^2|} \text{ and } \phi = \arctan(\omega M/R)$$

where ϕ is the phase angle between V_x and *I*. In practice it may be necessary to reverse either or both of the component voltages to secure balance, and reversing switches are incorporated for this purpose. The mutual inductor should be astatic, to avoid pick-up errors.

11.9.2.2 Gall potentiometer

The Gall potentiometer (*Figure 11.26*) provides two quadrature voltage components, summed and balanced against the unknown V_x. The supply voltage *V* is applied to the primaries of isolating transformers T_1 and T_2. The secondary of T_1 supplies a current, approximately in phase with *V*, to R_1 which consists of a tapped resistor and slide wire in series. The current is adjusted by resistor R_3 and read on an ammeter A, and passes through the primary of a fixed mutual inductor *M*. Any desired value of in-phase voltage (within the available range) is obtained from tappings on R_1. The primary of transformer T_2 has a series resistor R_5 and a variable capacitor *C*, and may be so adjusted that the current in R_2 is in quadrature with that in R_1. The exact phase angle, and the calibration of the quadrature circuit, are achieved by balancing the secondary e.m.f. of the mutual inductor *M* against a fraction of the p.d. across R_2. Thus, variable voltages, phase and quadrature, can be obtained from R_1 and R_2, respectively. These, in series, are balanced against the unknown V_x through the vibration

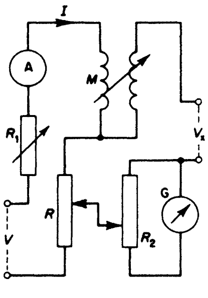


Figure 11.25 Larsen a.c. potentiometer

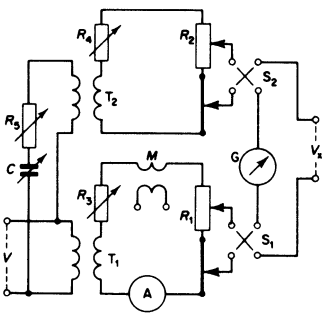


Figure 11.26 Gall coordinate a.c. potentiometer

galvanometer G. The reversing switches S_1 and S_2 facilitate balance and provide for a phase angle over a full 360°.

11.9.3 D.c. bridge networks

A bridge network has as its distinctive feature a 'bridge' connection between two nodes, with a null detector to sense the balance condition of zero p.d. between the nodes.

11.9.3.1 Wheatstone bridge

The Wheatstone bridge is an arrangement in very wide use for the determination of one unknown resistance in terms of three known resistances. The network is shown in *Figure 11.27*, where R_1, R_2, R_3 and R_4 are resistors connected at the nodes a and b through a reversing switch S to a d.c. supply. The galvanometer G, with a shunting resistor to control its sensitivity, and the key K are connected to the nodes c and d as shown. R_1 and R_3 may be set at known fixed values. R_2 is a variable resistor and R_4 is the unknown resistance to be measured. The bridge is balanced by adjusting the value of R_2 until the deflection of the galvanometer, set for maximum sensitivity, is brought to zero. The condition for balance is easily seen to be $R_4/R_3 = R_2/R_1$ and is independent of the voltage applied to the bridge: whence the unknown is $R_4 = (R_3/R_1)R_2$. The bridge may be built up as a single unit with, for example, R_1 and R_3 each able to be set at one of the resistance values 1, 10, 100 or 1000 Ω. The bridge ratio R_3/R_1 may therefore have a range of values from 1000/1 to 1/1000. In this case R_2 might conveniently be adjustable, by means of dial switches, in steps of 1 Ω from 1 to 11 110 Ω.

The bridge resistors are wound from manganin wire, since manganin is an alloy with a low temperature coefficient of resistance and a low thermoelectric e.m.f. to copper. Any thermoelectric e.m.f.s in the bridge connections which did not on average cancel themselves out could give a false result. Errors from this cause are eliminated by taking a balance for each position of the changeover switch S and using the mean of two observed values of R_2.

High-precision Wheatstone bridges are capable of measuring resistances between 1 Ω and 100 MΩ with uncertainties which vary between 5 and 100 in 10^6 of the reading over this range of values. Modern resistors are very stable devices and they should not change in value by more than a few parts in 10^6 per year; however, it is prudent to check the linearity of resistive bridges regularly, particularly if they are being used to the limit of the original specification. Some bridges have trimming facilities available with the principal resistors, although corrections can always be applied which although tedious, can avoid imposing new resistance drifts on the bridge as a consequence of changing *stable* resistors.

Various forms of high-quality bridges are used for measuring the changes which occur in resistive transducers. The Smith and Mueller bridges, used for measurements with precision platinum resistance thermometers, can yield 0.001°C discrimination in readings. For strain gauge transducers there are portable bridges that can detect changes of 0.05% or less.

11.9.3.2 Kelvin double bridge

The Kelvin double bridge is an adaptation of the Wheatstone bridge which may be used for the accurate measurement of very low resistances, such as four-terminal shunts or short lengths of cable. In the network (*Figure 11.28*) R_x is the unknown and R_s is a standard of approximately the same ohmic value. The two are connected in series in a heavy-current circuit (e.g. 50 A for $R_s = 10$ mΩ, giving 0.5 V drop and adequate sensitivity). Suitable values for R_1, R_2, R_3 and R_4 are in each case a few hundred ohms. To use the bridge, R_1 and R_2 are made equal; then R_3 and R_4 are adjusted (keeping $R_3 = R_4$) until balance is achieved. Then the balance condition is $R_x/R_s = R_4/R_2 = R_3/R_1$. Provided that it is short and of adequate cross-section, the connection between R_x and R_s does not affect the measurement.

Good-quality commercial bridges claim an uncertainty of about ±0.05% for unknowns between 1 Ω and a few micro-ohms. Near the latter level a small constant resistance becomes a limitation. As $R_1, \ldots, R_4$ constitute a Wheatstone bridge network, a Kelvin bridge is often extended to measure two-terminal resistances up to 1 MΩ with an uncertainty of ±0.02%.

11.9.3.3 Digital d.c. low resistance instruments

Digital milliohmmeter/micro-ohmmeter instruments have built-in d.c. supplies; they are auto-ranging and make four-terminal measurements by internal comparator-ratio techniques to give, e.g. $4\frac{1}{2}$-digit displays in the case of the Tinsley 5878 micro-ohmmeter with 0.1 μΩ best resolution and uncertainty ±(0.1% → 0.3%) of reading. Also, a IEEE 488 bus attachment is available for use in automated test systems. Thermal e.m.f. balance is incorporated as well as lead compensation.

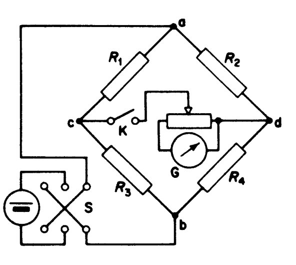


Figure 11.27 Wheatstone bridge

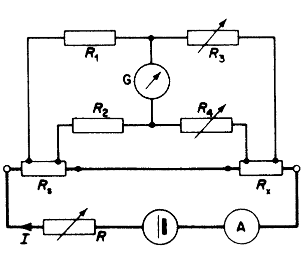


Figure 11.28 Kelvin double bridge

11.9.3.4 *Comparator resistance bridge*

The comparator resistance bridge is a variant of the comparator potentiometer (*Figure 11.23*). In this case the network connections and adjustments are such that, at balance,

$$I_mR_m = I_bR_b \text{ and } I_mN_m = I_bN_b, \text{ whence } R_m = N_m(R_b/N_b)$$

where R_b is the known standard four-terminal resistor and N_b has been adjusted to make R_b/N_b an exact decade value. The unknown four-terminal resistance R_m is found in terms of a number of turns N_m and a decade factor.

The two resistors can be compared at different current levels, so that, for example, the rated current can be used with the test resistor, while the standard resistor can be used at a current well below the level at which self-heating effects would change its resistance value. The actual comparison uncertainty with this class of bridge is less than 1 in 10^6, to which must be added the basic uncertainty in R_b, probably 2 in 10^6 if it is a class 'S' (or 'Wilkins') standard four-terminal resistor *known* to be stable.

11.9.4 A.c. bridge networks

A.c. network parameters (impedance, admittance, phase angle, loss angle, etc.) can be measured in the following ways:

(1) Single-purpose or multipurpose a.c. bridge networks, with four or six arms or with inductive ratio arms.
(2) Analogue and digital networks developed for special applications and consisting of frequency-selective or resonant or filter networks; they are associated usually with measurements at radiofrequencies and high audiofrequencies.
(3) Digital multimeter instruments for measuring the modulus of the unknown parameter (but not other qualities).

The treatment in this subsection is confined largely to (1).

11.9.4.1 *Balance conditions*

In the four-arm bridge of *Figure 11.29*, the arms comprise *impedance operators* Z_1, Z_2, Z_s and Z_x. The network is supplied at nodes a and b from a signal generator, and a sensitive null detector is connected between nodes c and d. If Z_1 and Z_2 (fixed standards) and Z_s (variable standard) are so adjusted that the p.d. between c and d is zero, then the unknown impedance Z_x depends on whether Z_1 and Z_2 are adjacent as in *Figure 11.29*(*a*), or in opposite arms as in *Figure 11.29*(*b*); in either case the balance condition is that the product of pairs of opposite arms is equal, giving

(*a*) *ratio bridge*: $Z_x = Z_s(Z_1/Z_2)$;
(*b*) *product bridge*: $Z_x = Y_s(Z_1/Z_2)$

where $Y_s = 1/Z_s$. These are *complex* expressions, so that the equality involves both magnitude and phase angle; alternatively, both phase ('real') and quadrature ('imaginary') components. Thus, equation (*a*) above can be written in terms of *magnitudes* in either of the following ways:

$$Z_x\angle\phi_x = Z_s(Z_1/Z_2)\angle(\phi_s + \phi_1 - \phi_2)$$
$$R_x + jX_x = (R_s + jX_s)(R_1 + jX_1)/(R_2 + jX_2)$$

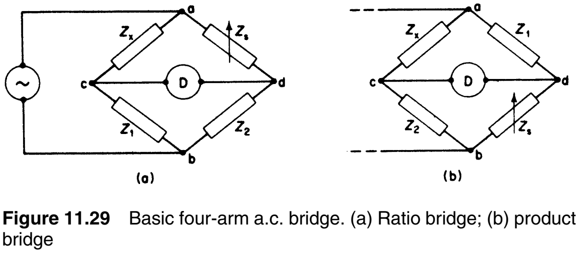


Figure 11.29 Basic four-arm a.c. bridge. (a) Ratio bridge; (b) product bridge

The bridge arms will include intentional or stray inductance and capacitance, and the corresponding reactances are functions of frequency. Most practical bridges are so chosen that only two or three final adjustments are necessary to obtain the optimum null condition of the detector.

11.9.4.2 *Signal source*

The bridge supply can be from a screened power-frequency instrument transformer, but more probably from a fixed- or variable-frequency signal generator (see Section 11.7.4). A voltage up to 30 V may be required, but a few volts is normally adequate, especially for radiofrequency bridges. The maximum power of a signal generator is of the order of 1000 mW, so that to avoid overload and distortion it is important that the appropriate voltage be set and a reasonable match be obtained between the generator and the effective input impedance of the bridge network.

11.9.4.3 *Detector*

A moving-coil vibration galvanometer may be used at discrete frequencies over the range of 10 Hz–1.2 kHz. The galvanometer has a taut suspension mechanically tuned to the operating frequency. The method is sensitive, and effectively filters harmonics, but it is often more convenient to employ the electronic detectors included in most commercial bridges. These include tuned-resonance amplifier networks or broad band high-gain electronic amplifiers; both give a rectified out-of-balance display on a d.c. instrument, with 'magic eye' or c.r.o. as possible alternatives. In some portable instruments it was the practice to use head-telephone sets, which are highly sensitive in the audio range of 0.8–1.0 kHz.

11.9.4.4 *Bridge-arm components*

Precision a.c. resistance boxes can be designed with very low time constants. They are available with switched or plugged groups of ten resistors in decade values between 0.1 Ω and 1 MΩ. Careful construction and screening enable some of these devices to be used at frequencies up to 1.2 MHz before the residual reactance becomes significant. An overall uncertainty of ±0.05% is possible, but 0.1–1.0% is more conventional. (See also Section 11.9.1.8.)

Multi-dial switched mica capacitors are available with good stability and low loss tangent. Single- and double-screen capacitors have respectively, three and four terminals, and care is needed to avoid unwanted stray capacitance, the screens being maintained at the proper potentials with respect to the equipment and the environment.

Continuously variable calibrated air-dielectric capacitors with ranges between 20 and 1200 pF are used for small adjustment, and ranges of fixed or variable mutual inductors are available. For high-voltage bridges the standard capacitors are of fixed value, and designed with air or compressed gas as dielectric for 300 kV and above.

11.9.4.5 *Low-frequency and audiofrequency bridges*

A selection is given of the more important bridges, many of which form the basis of commercial instruments; for example, the inductance bridges of Maxwell and Hay, the modified de Sauty bridges for capacitance, the Schering bridge for capacitance and loss tangent at low and high voltages, and the Wien bridge for frequency and a.c. resistance. These bridges are commonly made to give results with a ±1% uncertainty, but can be designed for ±0.1%.

All bridges should be checked occasionally to show that the reference standard components have not changed and that the range and ratio division has not drifted; also, if the bridge is frequency sensitive, the frequency of the oscillator and the drift rate should be investigated. Naturally, superior standard components must be available which have recent 'traceability' certification, in addition to accurately known decade-ratio measurement techniques. Where bridge measurements are implicit in contract specifications, it may be prudent or necessary to justify the claims by referring the instrument to a calibration service standards laboratory.

11.9.4.6 Mutual inductance

Figure 11.30(*a*) shows the simple Felici–Campbell bridge for the measurement of an unknown mutual inductance M_x in terms of a variable standard M. For balance on the null detector D, then $M_x = M$, on the assumption that the inductors are perfect, i.e. that the secondary e.m.f. is in phase quadrature with the primary current. Each, however, has a small in-phase component that can be represented by an equivalent resistance σ. The modification in *Figure 11.30*(*b*) shows Hartshorn's arrangement, which enables the impurities to be measured: at balance by adjustment of M and r, the conditions are $M_x = M$ and $\sigma_x = r + \sigma$.

11.9.4.7 Inductance

Inductance may be measured by Wien's modification of the Maxwell bridge (*Figure 11.31*(*a*)). If the unknown inductor has an inductance L_x and an equivalent series resistance R_x, balance gives

$$L_x = R_2R_3C_1; \quad R_x = R_2R_3/R_1; \quad Q_x = \omega L_x/R_x = \omega C_1R_1$$

The advantage is that inductance is measured in terms of a high-quality and almost loss-free capacitor. The bridge can measure a wide range of inductance values with Q factors less than 10. High Q factors require excessively large values of R_1, and by creating the same effective phase characteristic by means of a low-valued resistor R_1 in series with C_1, the Hay bridge of *Figure 11.31*(*b*) is obtained. The balance conditions for it are

$$Q_x = 1/\omega C_1R_1; \quad R_x = R_2R_3/R_1(1 + Q_x^2);$$
$$L_x = R_2R_3C_1/(1 + 1/Q_x^2)$$

A disadvantage is the need to measure the frequency.

A commercial instrument using the Maxwell and Hay bridges has built-in supplies at frequencies of 50 Hz, 1 kHz and 10 kHz, and an inductance range from 0.3 μH to 21 kH with an accuracy within ±1%. The loss resistance is known to about ±5%, and it is possible to introduce a d.c. bias.

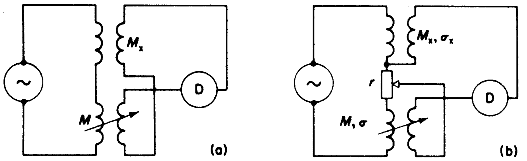


Figure 11.30 Mutual inductance bridges. (a) Felici–Campbell; (b) Hartshorn

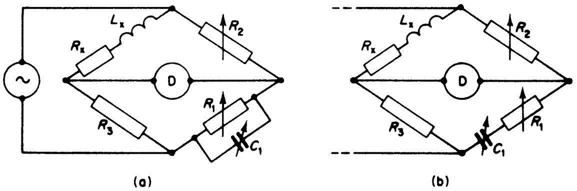


Figure 11.31 Inductance bridges. (a) Maxwell–Wein; (b) Hay

11.9.4.8 Capacitance at low voltage

Relatively pure capacitors can be measured by the de Sauty bridge (*Figure 11.32*(*a*)), the balance condition for which is $C_x = C_s(R_4/R_3)$. Imperfect capacitors can be compared by the arrangement shown in *Figure 11.22*(*b*), in which the imperfection is represented by a series resistance r. To obtain balance, adjustment of all four resistors R is necessary:

$$C_x = C_s(R_4/R_3) = C_s[(R_2 + r_s)/(R_1 + r_x)]$$

For the loss tangent,

$$\tan\delta_x = \tan\delta_s + \omega C_s[R_2 - (R_1R_4/R_3)]$$

Portable commercial capacitance testers, with battery-driven 1 kHz oscillators and rectified d.c. or headphone detectors, use a modified de Sauty bridge network so that loss tangent can be estimated and the capacitance measured to an uncertainty of about ±0.25%.

11.9.4.9 Capacitance at high voltage

Dielectric tests at high voltages, particularly of the loss tangent, can be made with the Schering bridge (*Figure 11.33*). The bridge is in wide use for precision measurements on solid and liquid dielectrics, and for insulation testing of cables, high-voltage machine windings and capacitor bushings. The capacitance of the high-voltage standard capacitor C_s is calculable, and with an air or gas dielectric can be assumed to have zero loss. A typical standard capacitor for 150 kV is shown in *Figure 11.34*; for higher voltages the clearances necessary for avoiding corona and breakdown are large, but the dimensions can be reduced by a construction in which the dielectric gas is under pressure.

The test capacitor in *Figure 11.33* is represented by C_x and a series loss resistance r_x. R_3 is a variable resistor, and the fourth arm comprises a variable low-voltage capacitor C_4 in parallel with a resistor R_4 (which may be fixed). Then, for balance

$$C_x = C_s(R_4/R_3) \quad \text{and} \quad \tan\delta_x = \omega C_4R_4$$

The loss tangent is a valuable and sensitive indication of the quality of the test insulation. If periodical tests reveal a gradual increase in the loss tangent, then deterioration is

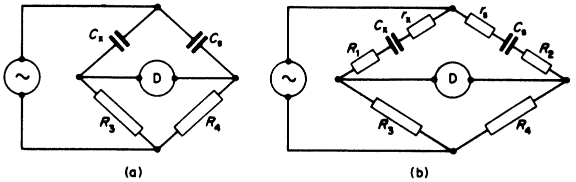


Figure 11.32 Capacitance bridges. (a) De Sauty; (b) modified de Sauty

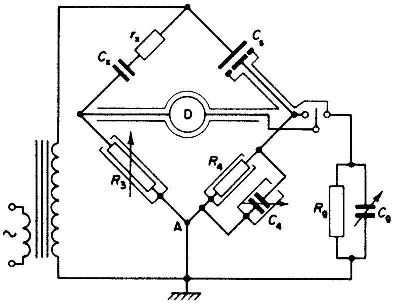


Figure 11.33 Schering bridge

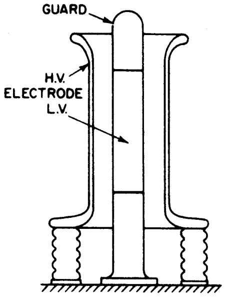


Figure 11.34 Air capacitor

occurring, the loss power is increasing and breakdown in service is probable.

To avoid electric field coupling between bridge components—in particular, between the high-voltage electrodes and the connecting leads—that would affect accuracy, components R_3, R_4 and C_4 are enclosed in metal screens connected to the earthed point A. Again, to avoid errors due to intercapacitive currents between the centre low-voltage electrode and the guard electrode of C_s, these electrodes should be brought to the same potential by aid of the auxiliary branches C_g and R_g. Balance is achieved with the detector switched to this combination, in addition to the main balance. If the test capacitor C_x is a cable with an earthed sheath, or a capacitor bushing on site and not readily insulated from earth, then the bridge node A must be isolated: careful screening is now even more necessary. *Figure 11.35* shows the arrangement for a portable Schering bridge equipment for testing an earthed bushing. It will be seen that an earthed screen isolates the bridge from the transformer primary, and a second screen connected to point A keeps capacitive currents from the high-voltage connections out of the bridge arms.

Discharge bridge This has been developed to detect the onset of void breakdown in cables and bushings. The network (*Figure 11.36*) can be used to test an unearthed component—for example, the bushing C_x; the arrangement resembles that of the Schering bridge. The stray capacitances of the guards and screening of C_s act as the capacitance C_4, which, however, at the discharge frequencies concerned, is greater than that required for balance. Hence, it is necessary to provide the variable capacitor C_3. The bridge output is fed to a rectifier milliammeter D through a filter and amplifier designed to pass, e.g. the band 10–20 kHz, so that the indication is a measure of the discharge current. On the supply side of the bridge is connected some artificial source of discharge voltage, such as the neon tube N, which serves initially to balance the detector D.

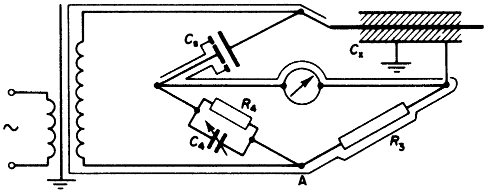


Figure 11.35 Schering bridge and earthed bushing

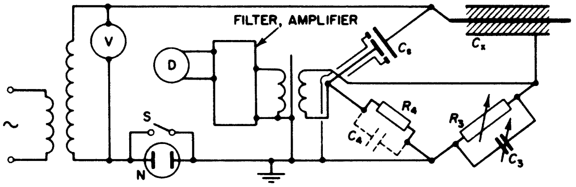


Figure 11.36 Discharge bridge

11.9.4.10 *Wagner earth*

Capacitance between bridge arms and earth in the Schering bridge causes disturbing currents to flow so that false balance conditions may result. If, as in *Figure 11.37*, the bridge supply is not earthed but the nodes c and d are at earth potential, then the detector is also at earth potential and no stray capacitance currents can flow in it. Further, all branch capacitances to earth are stabilised. The Wagner earth method secures this condition by the use of the additional impedances Z_5 and Z_6, the junction of which is solidly earthed. Z_5 and Z_6 must be of the same type as either Z_1 and Z_2, or Z_3 and Z_4. If balance is achieved for both positions of the switch, then nodes c and d must have the same potential as the earthed junction, which is zero. Stray capacitance at a and b is innocuous as it merely 'loads' the supply.

The Wagner earth is usually applied to commercial Schering bridges, but the technique is applicable to any bridge and can be particularly helpful at higher frequencies.

11.9.4.11 *Inductive ratio-arm bridge*

The principle is explained by reference to the simplified diagram in *Figure 11.38*(*a*), the essential features being the coupled windings of, respectively, N_x and N_s turns. If the voltage per turn is v, then the two windings have the voltages $V_x = N_x v$ and $V_s = N_s v$, respectively. Let $V_x = V_s$; then the current I_x through the unknown capacitive admittance Y_x can be made equal to the current I_s by variation of the parallel standards (conductance G_s and capacitance C_s),

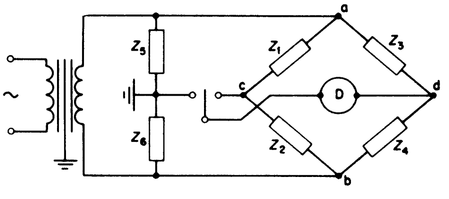


Figure 11.37 Wagner earth arrangement

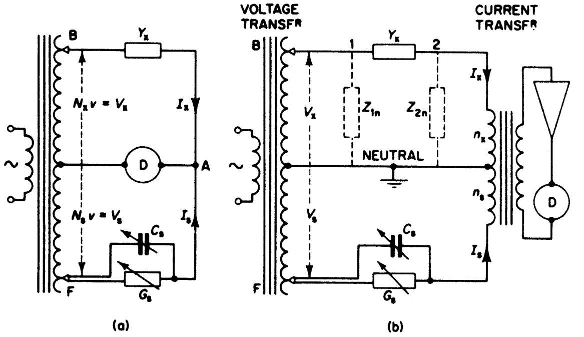


Figure 11.38 Inductive ratio-arm bridge. (a) Basic principle; (b) with isolated detector

until the detector shows a null reading. For this condition the standard admittances represent the *parallel* equivalent network values of Y_x. If Y_x is inductive, the C_s connection at F is transferred to B to avoid the use of standard inductors, which are very inferior to standard capacitors.

The voltage linearity of the input transformer is of fundamental importance to this class of bridge. With modern techniques and magnetic core materials, it is possible to wind toroidal cores to produce an extremely uniform reactive distribution (leakage inductive reactance and turn-to-turn capacitance) with a resultant non-linearity of about 1 in 10^7 in the magnitude of the mutually induced e.m.f. per turn.

In one well-known bridge (Wayne Kerr) the primary winding of a current transformer is inserted at A and the detector removed to the secondary winding as in *Figure 11.38(b)*. The two parts of the primary winding (n_s and n_x turns) are wound in opposite senses and the detector can indicate the zero m.m.f. condition in the windings.

In general, N_s differs from N_x and n_s from n_x, so that $I_x n_x = I_s n_s$ for balance; further, the primary of the current transformer is at the ground (or earth) potential of the neutral if the trivial resistance drop of the windings is ignored, so that $I_x = (N_x v) Y_x$ and $I_s = (N_s v) Y_s$. Combining these results gives

$$Y_x = Y_s(N_s/N_x)(n_s/n_x)$$

The double turns ratio product can be used to permit *one* accurate standard resistor or capacitor to be switched to any N_s turns of value 1–10 to simulate *ten* accurate standard components by the precise selection of turns. Moreover, wide-range multiplication of the simulated standard is achieved by decade tappings on N_x and n_x.

An important advantage with this class of bridge is the possibility of connecting the neutral to earth; then small leakage currents from Y_x to earth are almost completely eliminated, since the current transformer is at earth potential and any *small* leakage currents from the high-voltage side of Y_x merely cause a slight *uniform* reduction in v but does not alter the turns ratio. Hence, determinations such as the following are possible:

(1) One-port low-capacitance measurement in the presence of a large shunt capacitance due to connecting leads.
(2) One-port *in situ* impedance measurement, the effect of the remainder of the network (shown dotted in *Figure 11.38*) is neutralised.
(3) Three-terminal impedance transfer functions for correctly terminated networks.

The results of measurements appear in *parallel admittance* form, and may require elementary computation to render them into *series impedance* form, with due regard to frequency. With medium to high values of impedance, resistance and capacitance are unaffected, but for inductance measurement (involving a 'resonance' balance) a reciprocal ω^2 correction is needed. For low-impedance networks, resistance and inductance are unchanged but capacitance readings require correction.

11.9.4.12 Characteristics

Universal bridges of the inductive ratio-arm type measure over a wide range of values, at an angular frequency of, e.g., 10 krad/s from a built-in oscillator, or at other frequencies up to 10 MHz using external signal generators and tuned electronic detectors. The measurement uncertainty can be within ±0.1–1.0%, provided that the standard component values are trimmed against superior standards and that the actual frequency is measured when frequency sensitive parameters are measured.

The Wayne Kerr precision bridge operates at 1591.55 Hz ($\omega = 10$ krad/s) and the six-figure display has a minimum uncertainty of 0.01%. The instrument uses an electronic null-seeking process to assist the *automatic* balancing procedure. With external source the bridge has a frequency range of 0.2–20 kHz with manual balancing.

Admittance bridges are commercially available to operate up to 100 MHz. At this frequency lumped parameter measurements are being displaced by distributed-parameter network characteristics.

11.9.4.13 Current comparator bridge

The a.c. current comparator instrument uses current transformers only in the load network. It is based on the principle, of m.m.f. (i.e. ampere-turn) balance on an ideal magnetic circuit, in which the current ratio is the inverse turns ratio.

The two basic current coils, P and S in *Figure 11.39*, are wound in opposing senses on a common magnetic core with a high initial permeability. The null m.m.f. condition is sensed by winding M and manual selection of the turns ratio N_p/N_s. The ideal balance condition can be closely approached in practice by shielding the detector from magnetic leakage by a laminated magnetic screen and by a compensation winding W_c. The shield must not form a short-circuited turn; but it does aid magnetic energy transfer between windings P and S. This is advantageous in current-transformer calibration, as it enables an external burden B to be supported under balance conditions. Any change in the capacitive error due to B can be neutralised by W_c connected on one side of the separate centre-tapped transformer W_b. By careful design the capacitance distribution is uniform.

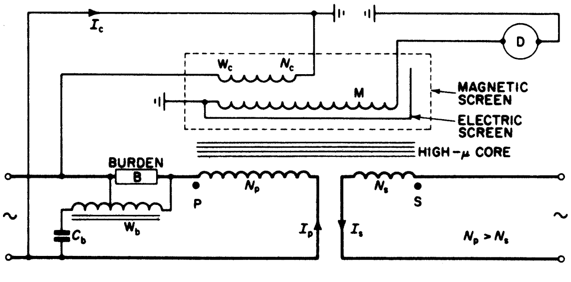


Figure 11.39 Current comparator with capacitive and burden compensation

The internal p.d. of the major winding can be neutralised by a parallel compensation winding W_c wound within the magnetic shield, so reducing capacitive currents, as it can be held at earth or reference ground potential.

The commercial range of comparators include d.c. potentiometers, direct voltage and current ratios, d.c. and a.c. four-terminal resistance comparison, voltage- and current-transformer errors, and high-voltage capacitor comparison. All the instruments derived from d.c. and a.c. comparators possess very linear properties, excellent discrimination and good stability. When certified they can be used as comparative reference standards.

11.9.4.14 H.V. capacitance comparator bridge

Standard capacitors used with high-voltage bridges (such as the Schering and comparator types) use compressed gas construction up to a maximum of 1000 pF, although 100 pF is more common. Measurement of such capacitors is based on the principle that a stable alternating voltage to two capacitors produces a current ratio equal to the capacitance ratio. The variable turns ratio of the comparator bridge can be used to balance an unknown capacitor up to 1000 times as large as the standard. By cascading, the range can be extended still further.

11.9.4.15 Substitution

In any active network the equivalent parameters of a component may be measured if variable reference standards can be either substituted for the component or connected in parallel with it—provided always that the standards can be adjusted so that no final change occurs in the voltage, current, phase or harmonic content of the waveforms in the network. The substitution principle can be applied to any type of bridge, or other class of network; resonant networks often provide the best discrimination between the two conditions. It is important that the stray capacitance, inductance and resistance paths, to earth and between components, should be unchanged by the act of substitution; hence, for the substitution test there should be (ideally) no change in the geometrical layout of leads and components. This can be easily achieved with low-loss changeover switches or coaxial connectors.

For the measurement of radiofrequency components, where shunt capacitance effects are prominent, the substitution principle is particularly valuable. Typical applications are discussed below.

11.9.4.16 Parallel-T network

Figure 11.40(*a*) shows a general parallel-T network, with an a.c. source as input and a detector across the output terminals. The balance condition (of zero output) in terms of the impedance operators is

$$Z_1 + Z_2 + (Z_1Z_2/Z_3) + Z_4 + Z_5 + (Z_4Z_5/Z_6) = 0$$

Suppose that Z_1 and Z_2 are pure reactances of like sign, e.g. $-jX_1$ and $-jX_2$, and that Z_3 is purely resistive, then

$$Z_1Z_2/Z_3 = -(X_1X_2/R_3)\Leftarrow$$

which is equivalent to a negative resistance and capable of balancing out a positive resistance in Z_4 or Z_5. An example is shown in *Figure 11.40*(*b*). The source is a modulated signal generator and the detector is, in effect, a radio receiver tuned to the 'carrier' frequency. Balance is achieved by variation of C_3 and C_6 with the test terminals open-circuited. If an unknown admittance $G_x + jB_x$ is now connected across the test terminals and the disturbed balance restored by altering C_3 and C_6 to $C_3'\Leftarrow$and C_6', respectively, then

$$B_x = \omega(C_3 - C_3') \text{ and } G_x = \omega^2C^2(C_6' - C_6)P/C_4$$

Bridge screening is simplified because the common source and detector terminal can be earthed. The use of a variable resistor as a balance arm can often be avoided, a considerable simplification for radiofrequency measurement.

11.9.4.17 Bridged-T network

This can be used at frequencies up to about 50 MHz. *Figure 11.41*(*a*) shows the schematic arrangement, and *Figure 11.41*(*b*) illustrates a common application. If the unknown impedance $Z_x = R + j\omega L$ is such that R^2 is negligible compared with ω^2L^2, then for balance

$$1/\omega^2LC = 2 \quad \text{and} \quad R = S/4$$

whence R and L can be obtained if the angular frequency $\omega\varsigma$ is known.

11.10 Measuring and protection transformers

Measuring (instrument) transformers are used primarily for changing currents and voltages in power networks to values more suited to the range of conventional indicators. Protection transformers are employed in systems of fault protection. Both types are dealt with in BS 3938 (*Current transformers*) (IEC 185) and BS 3941 (*Voltage transformers*) (IEC 186 and 186A).

11.10.1 Current transformers

Air-cooled current transformers (c.t.) are used in circuits of voltage up to 660 V and currents up to 75 kA. The *bar primary* form employs the actual cable or bus-bar of the main circuit as the primary winding. The core is built of high-grade steel laminations in rectangular or circular shape. The secondary has the appropriate number of complete turns uniformly around the annular core or on all four sides of

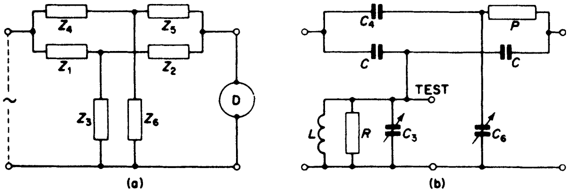


Figure 11.40 Parallel-T networks

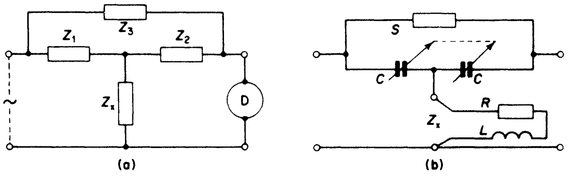


Figure 11.41 Bridged-T networks

the rectangular core. The assembly of core and secondary encloses the bar primary, which is equivalent to a single turn.

For lower primary currents it is necessary to provide a conventional primary winding. This gives a better primary/secondary coupling and a greater accuracy.

Current transformers must be insulated to withstand the service voltage, which may be up to 400 kV. Oil immersion may be necessary, in which case the core and coils are clamped to the top cover to facilitate removal as a complete unit.

11.10.1.1 Burden and output

Primary currents have preferred values between 1 A and 75 kA, with rated secondary currents of 5 A and 1 A (also 2 A). The *rated burden* is the ohmic impedance Z_T of the secondary circuit when carrying the rated current I_s at a stated power factor. The *rated output* is the volt–ampere product $(I_s Z_T)I_s$. The rated output, the limiting value for which accuracy statements apply, appears in selected preferred values between 1.5 and 30 VA. The normal operating mode of a c.t. is as a low-impedance device; hence, a short-circuiting switch is provided for the secondary winding for use when a low-impedance load (e.g. an ammeter) is not connected to the secondary terminals.

11.10.1.2 Errors

In an ideal c.t. the primary/secondary current ratio is precisely the same as the secondary/primary turns ratio, and the currents produce equal m.m.f.s in exact antiphase. In a practical c.t. the current ratio diverges from the turns ratio, and the phase angle differs by a small defect from opposition. These *ratio* and *phase angle errors* arise from that component of the primary current required to magnetise the core and provide for core loss, and from the e.m.f. necessary to circulate the secondary current through its burden. In transformers intended for accurate measurement and for metering, these errors must be small.

11.10.1.3 Classification

The limits of error define the 'class' of transformer. BS 3938 defines six classes for measuring and three for protective c.t.

Measuring c.t. The classes in descending order of accuracy are designated AL, AM, BM, CM, C and D. The limits of ratio error vary from ± 0.1% for AL to ± 5% for D; and the phase error from ± 5 min in AL to ± 120 min (2°) in types CM and C, with no limits quoted for type D.

Protective c.t. The classes are termed S, T and X. The limits of ratio error for S and T are, respectively, ± 3 and 5%, and the corresponding phase limits are 3°⇐and 6°⇐approximately. Limits for type X are not stated explicitly. In testing S, T and X classes it is necessary to distinguish between low- and high-impedance types.

11.10.1.4 Measurement

Errors are normally measured by comparison with a c.t. of higher class. By convention, a phase error is positive when the secondary current phasor leads that of the primary current.

11.10.1.5 Arnold method

In *Figure 11.42* C is the c.t. under test and B is a standard of the same nominal ratio and having known, very small errors. The working burden of the test c.t. is Z. The network is such that the difference between the secondary currents I_b and I_c flows in the non-reactive resistor R and is measured by means of the mutual inductor M and slide wire r fed by the auxiliary c.t. F (conveniently of ratio 5/5 A). Phase balance is achieved by adjusting M (positive or negative), and ratio balance by selection of r around the centre-tap of P. From the phasor diagram, the components of the difference current I are approximately $I_b - I_c$ and $-jI_c(\beta_c - \beta_b)$ because the phase angles are actually very small. At balance the condition is given by $I_b(r - j\omega M) = RI$. Equating the 'real' parts gives $I_c/I_b = 1 - (r/R)$; whence in terms of the current ratios $k_c = I_p/I_c$ and $k_b = I_p/I_b$, where the primary current is common to both c.t.s,

$$k_c = k_b(I_c/I_b) = k_b/[1 - (r/R)] \simeq k_b[1 + (r/R)]\Leftarrow$$

provided that $r \ll R$. Equating the 'imaginary' parts gives

$$\beta_c - \beta_b = (\omega M/R)(I_b/I_c) \simeq \omega M/R$$

The bridge readings give the errors directly by comparison with those of the reference transformer.

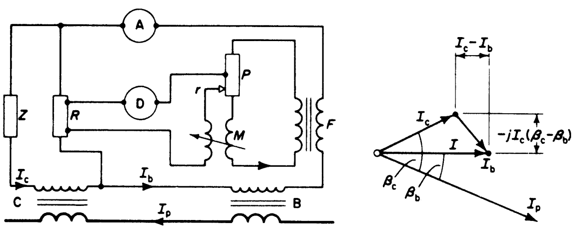


Figure 11.42 Arnold method for measurement of current-transformer errors

11.10.1.6 Kusters method

This is a comparator method (*Figure 11.43*) in which C is the c.t. under test with its rated burden Z, and B is an a.c. comparator with the same nominal ratio as C but possessing negligible errors. The errors of C can be deduced directly from the balance settings of the arms of the parallel $G - C_1$ network.

11.10.2 Voltage transformers

Magnetically coupled voltage transformers (v.t.) resemble power transformers in basic principle, and have recommended primary voltages up to $396/\sqrt{3}$ kV. Oil-immersion

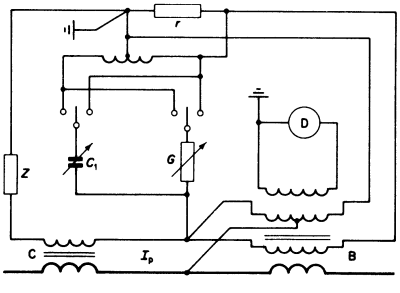


Figure 11.43 Kusters method for measurement of current-transformer errors

is necessary for these levels. The preferred secondary voltages lie between $110/\sqrt{3}$ and 220 V/ph.

11.10.2.1 Burden and output

The preferred rated output burdens lie between 10 and 200 VA/ph, and the rated burden is normally the limit for which stated accuracy limits apply.

11.10.2.2 Errors

Five classes are listed in descending order of accuracy, namely AL, A, B, C and D. The voltage ratio error limit varies between 0.25 and 5% and applies for *small* voltage changes (±10%) around the rated voltage; and the phase error limits are ±10 min to ±60 min from class AL to class C. The phase error of D is not defined.

For dual purpose v.t. (i.e. for both measuring and protective application) additional classes E and F are defined which quote the permitted limits of error for classes A, B and C when used at voltages between 0.05 and 1.9 times rated voltage. The v.t. is then denoted by both relevant class letters. The extended error limits are ±3–5% and ±2–5%.

11.10.2.3 Measurement

By convention, a phase error is positive when the secondary voltage phasor V_s leads the primary voltage phasor V_p. The ratio error is $(k_rV_s - V_p)/V_p$, where k_r is the nominal rated transformation ratio. The rated output is stated in volt-amperes at unity power factor for the specified class accuracy.

11.10.2.4 Dannatt method

This deduces the error in terms of network parameters (*Figure 11.44*). The v.t. primary is fed from a supply to which is also connected the standard air capacitor C_s in series with a low-valued resistor R_s. The v.t. secondary is loaded with the rated burden Z in parallel with a series-parallel network comprising high-valued resistors R_1 and R_2, a capacitor C_1 and the primary of a mutual inductor M. The secondary voltage of the mutual inductor is balanced against the volt drop across R_s. To ensure that the guard and guarded electrodes of C_s are held at the same potential, a variable resistance R_g is connected so that balance is obtained with either position of the switch S. The balance condition is such that the ratio and phase errors are given very closely by

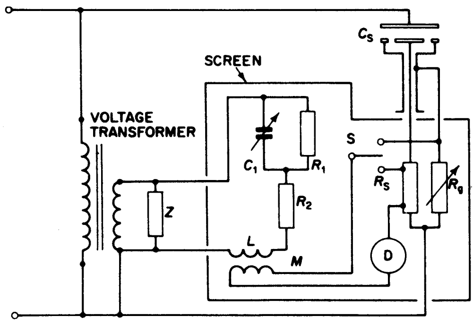


Figure 11.44 Dannatt method for measurement of voltage-transformer errors

$$M/C_sR_s(R_1+R_2) \quad \text{and} \quad \omega(L - C_1R_1^2)/(R_1+R_2)$$

where L is the self-inductance of the mutual inductor primary.

11.10.2.5 Kusters method

The a.c. comparator bridge can be applied. The capacitance ratio of two gas-filled high-voltage capacitors supplied by a common secondary voltage is measured with the currents passing through the separate balancing windings. If one high-voltage capacitor and its series-connected balancing winding are now fed from the primary supply of the v.t., then the voltage ratio can be found by rebalancing the comparator when the two capacitors are connected, respectively, to the primary and secondary voltages.

11.10.2.6 Capacitor-divider voltage transformer

For high-voltage transformers for 100 kV and above, a more economical and satisfactory voltage division for measurement and protective purposes is given by use of the capacitor-divider system (*Figure 11.45*). The reduced intermediate voltage appears across the protecting gap G as the input to a tuned transformer. Changes in the secondary burden Z, which would adversely affect the errors, are minimised by adjusting the transformer leakage reactance and/or that of a series inductor L to resonate with the input capacitance $(C_1 + C_2)$ effectively across the primary. There is a consequent sensitivity to frequency.

The classification is the same as for magnetic transformers, except that AL does not apply and there is an additional limitation on frequency. The transient performance is naturally significantly different from that of the normal v.t.

11.11 Magnetic measurements

Because of the non-homogeneity of bulk ferromagnetic materials, magnetic metrology is relatively inaccurate and imprecise. The fields of interest lie in the basic physics of magnetism, the distribution of the magnetic field, the assessment of magnetic parameters, and the measurement of core loss.

Physical basis The origin of magnetic phenomena lies in the statistical quantum-mechanical behaviour of electrons and particles. Additional information can be obtained by photography of domain formation.

Field distribution Two-dimensional field distributions in air-gaps are readily traced by current-field analogues such as conducting paper or the electrolytic tank. The tank can also be employed for restricted three-dimensional fields, and tests around models or actual equipments can be

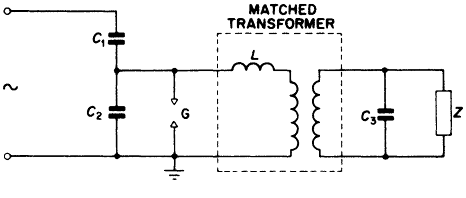


Figure 11.45 Capacitor-divider voltage transformer

based on various types of small sensor working into calibrated instruments (fluxmeter or ballistic galvanometer) or a Hall effect device.

Parameters The 'static' magnetisation characteristic, the relative permeability and the hysteresis loop are the principal parameters of interest.

Core loss The loss in a core is a matter of considerable technological importance. It is, however, very difficult to measure directly the losses in such a way that the results can be directly applied to machine constructions other than transformers.

11.11.1 Instruments

The most simple fundamental standard of magnetic flux density is based on the accurately known dimensions of a long, uniformly wound solenoid carrying a known current. Alternatively, a uniformly wound toroid can be used, although in this case the flux density will not be quite uniform across the core section. The measurement of flux and flux density can be carried out by means of the ballistic galvanometer or fluxmeter for the former, and the Hall effect instrument for the latter.

11.11.1.1 Ballistic galvanometer

The moving coil of a galvanometer with a long periodic time gives a first swing proportional to the time integral of the current through it (i.e. the charge) provided that the duration is very short. The charge results from a time integral of voltage (i.e. magnetic linkage) impressed on a known resistance. The moving coil is connected in series with a search coil and a resistor, and is immersed in the magnetic field due to a known current. When this current is *rapidly* reversed, the total change of linkage is presented to the galvanometer as an impulsive charge, and the first deflection of the instrument is used for calibration against a known (or calculable) linkage change, or for measuring an unknown linkage. The main limitations to accuracy are the observation of the scale and the uncertainty in the current measurement. A typical high sensitivity ballistic galvanometer has a period of 20 s, a moving-coil resistance of 850 Ω, and a sensitivity of 8.5 m/μC on a scale distant 1 m from the coil using a lamp and mirror technique.

11.11.1.2 Fluxmeter

The fluxmeter is a permanent-magnet instrument with a moving coil of low inertia and negligible control torque. Its damping is made high by use of a relatively thick aluminium former, so that the period is long. It is immaterial whether the time taken to change the linkage in a search coil connected to the moving coil is long or short, and in consequence the instrument is useful in iron testing, where the time taken for a flux to collapse or reverse may be several seconds. The deflection is read from the initial position of a pointer on a quadrant scale when the pointer reaches its maximum deflection; after this the pointer drifts slowly back to a zero position. A typical full-scale deflection would be given by a change of 10 μWb-t.

Strong air-gap flux densities can be measured by an alternative method in which a small coil is rotated at a high and known speed, the induced e.m.f. being proportional to the local flux density.

11.11.1.3 Hall effect instrument

The Hall effect applies to conductors generally, but its application is normally associated with semiconductor materials. These should be of low resistance so that, even for small signals, the thermal (Johnson) noise effects will be low. Bulk materials, indium arsenide and indium antimonide, have low resistances and good output voltage coefficients, with the InSb voltage larger than that of InAs; however, InAs is usually selected, as its temperature coefficient is one-tenth that of InSb. Thin-film InSb is used for switching applications where the higher output e.m.f. is the prime consideration. The Hall effect response is very fast, being usable up to the megaherz range. Bulk material InAs elements are more effective than thin-film elements, owing to the much lower resistances obtainable.

In commercial instruments the current I_c in the sensing element can be a direct current chopped at audio frequency; the resultant audiofrequency Hall voltage E_H, after linear amplification and demodulation, can be displayed on a taut-band d.c. moving coil indicator against a calibrated scale of flux density. Typical ranges have full-scale deflections between 0.1 mT and 5 T for steady fields. Pulsating fields can be measured with d.c. instruments at frequencies up to 500 Hz with a cathode-ray-oscilloscope detector. Instruments for alternating fields only are available for frequencies of about 30 kHz, with ranges similar to those of d.c. instruments.

An instrument and probe can be checked by use of stable reference magnets, which may have uncertainties of 0.5–1.0%. The overall uncertainty of a scale reading is typically ± ~~3~~% of the full-scale deflection for the range.

The sensing elements are protected by epoxy glass-fibre or similar enclosures, and are available in a variety of forms for probing transverse, coaxial or tangential fields. The flat form is most common, typically with the dimensions 0.5 mm × ~~4~~ mm × ~~25~~ mm. Incremental measurements, using two matched probes and backing-off networks, enable perturbations as small as 0.1 μT to be observed either in the presence of, say, a given 0.1 T field or as a difference between two separate 0.1 T field systems. Hall effect instruments have many obvious applications where the magnitude and direction of the field distribution is required such as air gaps of machines and instruments, magnetron magnets, mass spectrometer fields, residual interference fields, etc.: in addition, the Hall-effect principle can be used in a variety of transducer applications by making I_c and B proportional to separate physical functions, and measuring the instantaneous or average results of this product. One example is a wide band, 50 kHz wattmeter in which I_c and B are made, respectively, proportional to the scalar values of the load current and voltage.

11.11.2 Magnetic parameters

The d.c. magnetisation curve, hysteresis loop and relative permeability can be obtained by use of a fluxmeter or calibrated ballistic galvanometer, using a method substantially the same as for calibration. The magnetic test material constitutes the core of the toroid, or is built in strips to form a hollow square. The a.c. cyclic magnetisation loop differs from the 'static' loop, and is still further modified when a d.c. bias m.m.f. is superposed. The a.c. loop can be viewed on a c.r.o. displaying the B and H functions on the Y and X axes, respectively. The H function can be obtained from the volt drop across a small resistor carrying the magnetising current, and the B function from the time integral of the search-coil voltage obtained with the aid of an operational amplifier.

11.11.2.1 Core loss

In a low-frequency test the loop area of the *B*/*H* relation is directly proportional to the core loss per cycle of magnetisation, which provides a simple comparative test for specimen material. A more convenient method employs low power factor wattmeters, but care is needed to exclude instrument and connection losses. Selected samples of sheet steel intended for the cores of inductors and power transformers are prepared for testing in the form of strips (e.g. 30 cm × 3 cm) assembled in bundles and butted or interleaved at the corners to form a hollow square. The sides are embraced by magnetising and search coils of known dimensions. Input power and current, and mean and r.m.s. voltages, are read at various frequencies and voltages to assess the overall power loss within the material and (if required) an approximate indication of the eddy and hysteresis loss components. The Lloyd–Fisher and Epstein squares are commonly used forms, the latter being referred to in BS 601 (*Sheet steel*).

The magnetic properties of bars, forgings and castings are derived from galvanometer or fluxmeter measurements (BS 2454). A.c. potentiometer and bridge methods are also used, especially for low-loss materials at frequencies in the audio range. With ferrite and powder cores, as with *Q*-value tests of a coil with and without a slug core, r.f. resonance methods are convenient.

Production and quality control test equipment for core-loss and coil-turns instruments include a 'coil-turns tester' (Tinsley 5812D) with a four-digit display of the number of turns in non-uniform and uniform windings, by using an inductive comparative measurement against a standard winding: the same range of 'magnetic type' instrumentation includes a 'shorted-turns tester' which will detect single and multiple shorted circuited turns in coils of any shape or number of windings; while a 'core tester' can check transformer core characteristics of EI stacks, C cores and toroids including hysteresis loop output for a cathode-ray-oscilloscope display.

11.11.3 Bridge methods

Two examples of bridges for audiofrequency tests are given. The non-linear magnetic characteristics of materials introduce harmonics at higher flux densities, imposing limitations because the bridge must be balanced at fundamental frequency.

11.11.3.1 Campbell method

The specimen, shown as a toroid, is uniformly wound with N_2 secondary turns, overlaid with N_1 primary turns (*Figure 11.46*(*a*)). The magnetising current I_1 passes through the primary of the mutual inductor *M* and the resistor *r*. At balance indicated on detector D, the power factor and power loss are given by

$$\cos\phi = r/\sqrt{(r^2+\omega^2M^2)} \quad \text{and} \quad P = I_1^2 r(N_1/N_2)$$

The corresponding magnetising force and flux density are found from

$$H - (I_1N_1/l)\sin\phi \quad \text{and} \quad B_{\mathrm{m}} = V/4faN_2$$

where *l* is the mean length of the magnetic path in the toroid, *a* is its cross-sectional area, and *V* is the *mean* voltage as measured on the high-impedance *average* voltmeter V.

11.11.3.2 Modified Owen method

The toroidal specimen (*Figure 11.46*(*b*)), has N_1 a.c. magnetising turns, and an additional winding N_2 through which a steady polarising d.c. excitation can be superposed. The d.c. supply is taken from a battery through a high-valued inductor to minimise induced alternating current from N_1. The Wagner earth arrangement $R_5C_5C_6$ may be added to eliminate earth capacitance errors from the main bridge arms by balancing in both positions of the switch S. The high-impedance voltmeter V measures the r.m.s. voltage *V* across R_3. The a.c. magnetising force H_a due to the current V/R_3 in N_1, the d.c. magnetising force H_d due to I_d in N_2, the power loss in the specimen (including the I^2R loss in the winding N_1 which can be allowed for separately) and the peak flux density B_m are given by

$$H_{\mathrm{a}} = VN_1/R_3l; \quad H_{\mathrm{d}} = I_{\mathrm{d}}N_2/l; \quad P = V^2(C_4/C_3R_3)$$
$$B_{\mathrm{m}} = \sqrt{2}V(C_4/\omega C_2aN_1)\sqrt{(1+\omega^2C_2^2R_2^2)}$$

where *a* is the cross-sectional area of the specimen, and it is assumed that all time-varying quantities have sinusoidal waveform and angular frequency ω.

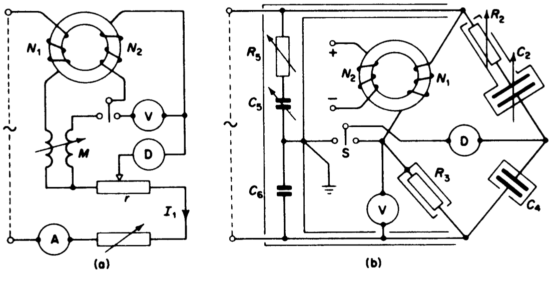


Figure 11.46 Bridges for magnetic measurements. (a) Campbell; (b) Owen

11.12 Transducers

In instrumentation systems a transducer is a device that can sense changes of one physical kind and transpose them systematically into a different physical kind compatible with a signal processing system. The compatible signals considered here are generally electrical or magnetic. Transducers can sense most non-electrical quantities (e.g. humidity, pressure, temperature and force) and, so far as the electrical response is concerned, can be classified as 'active' or 'passive', as follows:

Passive. The output response produces proportional changes in a passive network parameter such as resistance, inductance and capacitance.

Active. Transducers act as generators, the class including piezoelectric, magnetoelectric and electrochemical devices.

The 'Code for Temperature Measurement' is included in BS 1041. The sensors and instruments for the purpose are classified according to temperature range:

(1) Specialised thermometers for measurements near absolute zero. One such, based on the magnetic susceptance of certain paramagnetic salts, is suitable for temperatures below 1.5 K; another, employing an acoustic resonant-cavity technique, can be used up to 50 K. The associated

instruments are a mutual inductance bridge for the former, and electronic measurement of length for the latter.

(2) Ge, Si and GaAs p–n junction diode and carbon resistor thermometers are used with d.c. potentiometers and Wheatstone bridges for temperatures up to 100 K.
(3) Vapour pressure thermometers are used for standard readings below 5 K and for general measurements up to 370 K.
(4) Thermistor (resistance) and quartz (resonance) methods, with Wheatstone bridge and electronic counter, respectively, cover the range 5–550 K.
(5) Electrical resistance sensors (usually of Pt) are used with d.c. bridges throughout the range 14–1337 K.
(6) Thermocouple e.m.f.s are measured by d.c. potentiometer. The method, widely employed in industry, has ranges between 100 and 3000 K with various combinations of metals.
(7) The expansion of mercury in glass or steel capillary tubes is applied to measurements from 230 to 750 K. The range is 80–300 K if the mercury is replaced by toluene.
(8) Radiation and optical thermometers employ thermopiles, photodiodes or photomultipliers for measurements up to 5000 K, using d.c. potentiometric or optical balance methods. These provide the only practical methods for very high temperatures.

Mercury-in-glass and optical thermometers are indicating instruments only, but the remainder can be made to furnish graphical records. The electrical resistance and thermoelectrical thermometers are especially suitable for multipoint recording as for heated solid surfaces, points in a mass, or inaccessible places in electrical machines.

11.12.1 Resistive transducers for temperature measurement

A resistive sensor based on a substantially linear resistance/temperature coefficient should have small thermal capacity for rapid response and avoidance of local temperature gradient. Hence, materials of high resistivity (giving small volume) and temperature coefficient are desirable. Thermistors are made to fulfil these requirements.

11.12.1.1 Pure-metal sensors

Pure conductors such as Pt, Ni and W have low resistance/temperature coefficients but these are stable and fairly linear. Pt is generally adopted for precision measurements and, in particular, for the definitive experiment within the range 100–903.5 K of the International Practical Temperature Scale. Platinum and other metals can be deposited on ceramics to produce metal-film resistors suitable (if individually calibrated) for temperature measurement.

11.12.1.2 Platinum resistance thermometer

To a reasonable approximation the resistance R of a platinum wire at temperature θ (in degrees Celsius) in terms of its resistance R_0 at 0°C is

$$R = R_0(1 + k\theta) \quad \text{where} \quad k = (R_{100} - R_0)/100R_0$$

The temperature θ_p for a given value R is

$$\theta_p = 100(R - R_0)/(R_{100} - R_0)$$

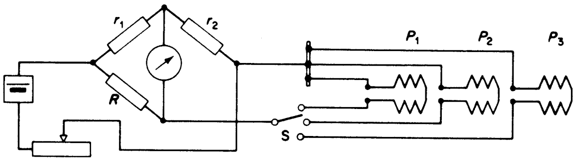


Figure 11.47 Temperature-sensor network

and is known as the *platinum temperature*. Conversion to true temperature is found from the difference formula

$$\theta - \theta_p = \delta[(\theta/100)^2 - (\theta/100)]$$

The value of δ depends upon the purity of the metal. It is obtained by measurement of the resistance of the sensor at 0, 100 and 444.67°C, the last being the boiling point of sulphur. The value of δ is typically 1.5.

A practical sensor usually consists of a coil of pure platinum wire wound on a mica or steatite frame, the coil being protected by a tube of steel or refractory material. The resistance measurement is carried out by connecting the coil in a Wheatstone bridge network or to a potentiometer. In the former case the bridge usually has equal ratio-arms and a pair of compensating leads is connected in the fourth arm. These leads run in parallel to the actual leads from the sensor coil and compensate for their resistance changes. If the initial resistance and coefficient k of the sensor are large, the compensating leads may not be required. *Figure 11.47* shows the method of connecting three such coils in turn in a Wheatstone bridge network, so as to measure the temperatures at three different locations. In this case the bridge is not balanced; the out-of-balance current through the galvanometer gives the temperature directly. Initial setting is done by adjustment of the battery current until some definite deflection is obtained, when a standard resistance replaces the sensors in the bridge circuit.

11.12.1.3 Thermocouple sensors

Thermocouple sensors are active transducers exploiting the Seebeck effect developed between two dissimilar metals when two junctions are at different temperatures. The International Practical Temperature Scale is defined in terms of a (0.9 Pt + 0.1 Rh)/Pt thermocouple over the range 903.5–1336 K. Readings between 100 and 3000 K are possible with Au/CoCu at the lower and W/Re at the higher end. Conventional materials for intermediate ranges include

copper/constantan (670 K)
iron/constantan (1030 K).
chrome/constantan (1270 K)
chromel/alumel (1640 K)

the figures giving the upper limits. The thermo-e.m.f. varies between 10 and 80 μV/K. To make an instrument direct-reading for the temperature at one junction, the second ('cold') junction must be kept at a reference temperature. The cold junction can be maintained at 0°C by immersion within chipped, melting ice in a thermos flask, or by using a commercial ice-point apparatus working upon the Peltier effect. When high temperatures are being measured, the terminals of the detector can sometimes be used as the 'cold' junction, but compensating leads should be connected between the thermocouple and the detector. The most accurate method for measuring the e.m.f. is with a d.c. potentiometer.

The e.m.f. e for a hot-junction temperature θς(in degrees Celsius) with the cold junction at 0°C is of the form log $(e) = \nless A \log(\theta) + B$, where the constants A and B have, for example, the values 1.14 and 1.36 for copper/constantan, with e in microvolts.

The small thermal capacity of thermocouples makes them suitable for the measurement of rapidly changing temperatures and for the temperatures at particular points in a piece of apparatus. One useful application is the measurement of surface temperatures, in which case the thermocouple consists of the two metals in the form of a flat flexible strip with a welded junction at the centre, this strip being applied to the surface under test.

11.12.2 Thermistors

Thermistors are semiconductor sensors which are made from the sintered compounds of metallic oxides of Cu, Mn, Ni and Co formed into beads, rings and discs. A high resistivity is achieved with a large resistance/temperature coefficient. The units can have resistances at 20°C from 1 Ω to several mega-ohms. The thermal-inertia time-constant is not more than 1 s for small beads, but rather more for discs and coated specimens. The outstanding property is the negative resistance/temperature coefficient of several parts in 10^2 per degree Celsius. Thermistors can be used over the range 200–550 K. One consequence of the current/voltage characteristic is self-heating if the thermistor carries appreciable current.

Wheatstone bridge networks are widely used, with the thermistor in one arm and the out-of-balance detector current as an indication of temperature. Although the characteristics are strongly non-linear, it is possible to obtain thermistors in matched pairs, which can be applied to differential temperature measurement. Another application, not directly a temperature measurement, is to the indication of air flow in pipe, or as anemometers: one thermistor is embedded in a metal block, acting as a thermal reservoir, the other senses the air speed as a cooling effect.

11.12.3 p–n Junctions

Certain semiconductor diodes are suitable for temperature measurement over a wide range; Ge, for example, is useful below 35 K. One differential thermometer has two matched p–n diodes which, when linearly amplified, give a discrimination better than 0.0001°C. Such instruments have several indirect applications, such as sensing thermal gradients in 'constant-temperature' enclosures or vats, monitoring load changes, and displaying the input–output liquid temperature differences in fluid pumps.

11.12.4 Pyrometers

Radiation thermometers respond to the total radiation (heat and light) of a hot body, while optical thermometers make use only of the visible radiation. Both forms of *pyrometer* are specially suited to the measurement of very high temperatures, because they do not involve contact with the source of heat.

Radiation thermometers Both the Fery variable-focus and the Foster fixed-focus types comprise a tube containing at the closed end a concave mirror which focuses the radiation on to a sensitive thermocouple. In use the tube is 'sighted' on to the hot body. The sighting distance is not critical, provided that the image formed is large enough to cover the thermocouple surface. Calibration is by direct sighting on to a body of known surface temperature.

Optical thermometers The commonest form is the disappearing-filament type, which consists of a telescope containing a lamp, the filament brightness of which can be so adjusted by circuit resistance that, when viewed against the background of the hot body, the filament vanishes. The lamp current passes through an ammeter scaled in temperature. The telescope eyepiece contains a monochromatic glass filter to utilise the phenomenon that the light of any one wavelength emitted by the hot body depends on its temperature. Although calibration is based on the assumption that the hot body is a uniform radiator, departure from this condition involves less error than in the radiation pyrometer.

11.12.5 Pressure

The piezoelectric effect is the separation of electronic charge within a material when applied pressure deforms the crystal structure. Conversely, the application of charge to the crystal will cause changes in the dimension of the crystal. Quartz is the only *natural* crystal in general use as a sensor. It is important that the crystal be prepared by slicing along the plane most sensitive for the particular application. Quartz is used in oscillators as the stable frequency-reference element. The self-resonant frequency is temperature-dependent, so constant temperature (e.g. 35°C) ovens are required for very stable oscillators: this application of quartz is not as a sensor, but the same temperature dependence of resonant frequency is applied for very-low-temperature measurement (below 35 K) where the temperature change can be interpreted by changes in frequency. Numerous ceramic crystals have been developed based upon barium titanate with controlled added impurities; the piezoelectric properties have to be applied by a special polarising treatment during the cooling period following the sintering process in a kiln.

Active four-arm strain gauge bridges, diffused into a single crystal silicon diaphragm, form the basic sensor for a wide range of sensitive pressure elements which are encapsulated in solid state transducers/transmitters (e.g. Druck Limited, Groby, Leicestershire, England). Minimum to maximum pressures can be measured, from venous or arterial values in physiology, to aerospace applications in engines and satellites, and to high-pressure industrial processes. The sensors have 10 V d.c. or a.c. input, with 15 mV → 10 V outputs, 0.1% non-linearity working into a $4\frac{1}{2}$-digit multichannel indicator/BCD-recorder which, when coupled with a $5\frac{1}{2}$-digit, 0.04% calibrator, provides traceability through a low-cost primary standard.

11.12.6 Acceleration

Accelerometers can be used for velocity measurement by integration of the output signal. The acceleration force of a mass is made to increase or decrease the spring pressure on a ceramic or quartz crystal to produce proportional piezoelectric effect. The device is insulated and hermetically sealed, with care taken to avoid loss of signal through parallel insulation paths of resistance comparable with that of the crystal itself (e.g. 1000 MΩ). Conditioning of the signal involves matching to the much lower input impedance of conventional amplifiers by use of an emitter–follower network. 'Integrated circuit' forms of the network can be located actually within the transducer unit.

The signal amplifier can be a voltage amplifier or a charge amplifier (i.e. an operational amplifier with capacitive feedback). The latter is preferred because the equivalent feedback capacitive effect is large and dominates the input capacitance associated with varying lengths of coaxial cable.

Other accelerometers are based either on changes in the reluctance of differential transformers or on variations in the resistance of a strain gauge. Slow changes in acceleration can be sensed by connecting the seismic mass to a wiper on a resistance voltage divider.

The upper frequency of the output is limited by the natural frequency of the transducer, while the minimum is almost zero. If the lowest useful frequency of an a.c. electronic amplifier is 20 Hz, the analysis of very-low-frequency signals may be achieved by first recording them on a precision frequency-modulated tape-recorder, which is then replayed at high speed for amplification and analysis.

Constant bandwidth frequency analysers are convenient for stable periodic complex waveforms. Constant per cent bandwidth types suit cases such as machines in which there is some small fluctuation in the nominal periodic behaviour. Truly random vibrations of a stochastic nature may more profitably be analysed in terms of the power spectral density function. Analysers are available for measuring many properties of random signals—e.g. 'probability density function' devices based on instantaneous signal amplitudes (see Section 11.7.5).

11.12.7 Strain gauges

Electrical strain gauges are devices employed primarily for detecting and measuring small dimensional variations in the surfaces to which they are attached, particularly where direct measurement is difficult. The gauge essentially converts mechanical displacement into a change in some electrical quantity (usually resistance). The essential feature is that strain shall be communicated to the gauge without fatigue or inertia effects.

11.12.7.1 Resistance-wire, p–n junction and carbon gauges

A grid of resistance wire is cemented between two insulating films. The grid is usually smaller than 30 mm × 15 mm, and has a resistance in the range 60–2000 Ω. When the gauge is cemented to the surface under test, the change ΔR in total resistance R due to a displacement ΔL in a length L is converted into strain ϵ_ς by means of the gauge factor

$$k_g = (\Delta R/R)(L/\Delta L) = (\Delta R/R)/\epsilon_\varsigma$$

If F represents the force and E the elastic modulus (i.e. the stress/strain ratio), then

$$F = k_g EaR/\Delta R$$

where a is the area normal to the direction of the applied force. Resistance-wire strain gauges have been in use for many years. Recent developments include metal-foil strain gauges based upon printed circuit and photoetching techniques: these can be made smaller. Thin-film techniques have been developed to apply the gauges directly to very small surface areas.

Semiconductor *silicon junction* strain gauges are perhaps 30 times more sensitive than the resistive gauge; however, their stress and temperature ranges are more restricted, and they are more difficult to 'cement' in position. The inherent non-linearity of the response by sensors can be offset by special bridge techniques using compensation networks.

Gauges consisting of a layer of *carbon*, which responds to strain in a manner similar to that of a carbon microphone, have a gauge factor typically of 20, considerably higher than for metals but much less than for p–n junctions. The inherent disadvantage is sensitivity to temperature and humidity.

11.12.7.2 Strain-gauge measurement techniques

Strain An 'active' gauge is cemented to the workpiece, and a similar 'compensating' gauge is left free. The gauges form two arms of a bridge (*Figure 11.48*(*a*)), which is inherently self-compensating for temperature. Balance is achieved initially by adjusting R_1, the load is applied and the bridge rebalanced. For multiple-gauge arrangements, 100-channel instruments with zero-balancing facilities are available, giving a printed read-out. Alternatively, after initial balance, the residual bridge voltage is measured, or applied to a cathode-ray-oscilloscope for recording. Dynamic strain may be measured by the circuit of (*Figure 11.48*(*b*)), usable with a flat-response amplifier over the range 0.5–1000 Hz. As in any structural stress analysis it is important to know the time relation as well as the amplitude of stress, the amplifier must have either a negligble phase shift or one directly proportional to frequency over the required working range.

Torque For a circular-section shaft in torsion, the principal stresses lie at 45° to the axis. If four gauges are applied as in *Figure 11.49*, and connected to a bridge through suitable slip-ring gear (usually silver ring surfaces and silver graphite brushes), the bridge unbalance is four times that of a single gauge. The arrangement is inherently temperature compensated. For engine testing it is normal to make up a special length of shaft with gauges and slip-rings, to be fitted between the engine and the brake or driven unit. With automatic recorders it is possible to record simultaneous data such as torque, pressure variation, temperature, etc., to give all phenomena on a time base display.

Thrust Various other measurements of mechanical power transmission can be measured. The arrangement in *Figure 11.50* is for thrust. Four axial (A) and four circumferential (C) gauges are disposed symmetrically and connected as indicated in a bridge network, with r as balancing adjustment at zero load. The monitoring system (which can be used with other strain-gauge applications) consists of a self-balancing potentiometer with chart (Ch), digital (Dg) and indicator (Id) readouts. The speed of a self-balancing potentiometer is restricted to about one reading per second.

Principal stresses The direction and magnitude of principal stresses can be determined from a *rosette* of three or

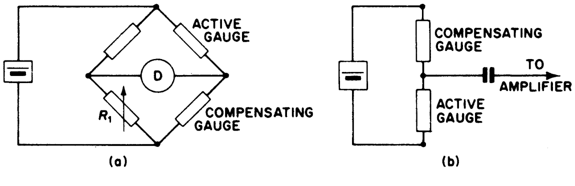


Figure 11.48 Strain-gauge networks

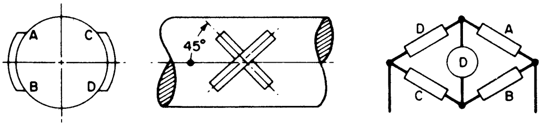


Figure 11.49 Strain-gauge torque measurement

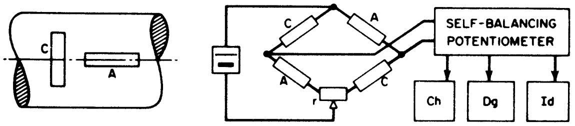


Figure 11.50 Strain-gauge thrust measurement

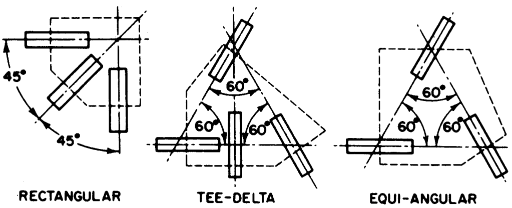


Figure 11.51 Typical strain-gauge rosettes

more strain gauges arranged with suitable orientations (*Figure 11.51*). Commercial rosettes are available in several geometrical formations. As it is unlikely that the principal strain will be on the axis of any one gauge, note must be taken of the sensitivity of the gauges in a direction at right angles to the axis if accurate results are required. The correction is not likely to exceed 3%.

11.12.8 Magnetostrictive transducers

Magnetostriction is the dimensional change that occurs in a ferromagnetic material when subjected to magnetisation, and an inverse effect of change in magnetisation when the physical dimensions are altered by the application of external force. The dimensional changes are very small (e.g. 30 parts in 10^6), but if the magnetic field alternates at a frequency corresponding to one of the natural frequencies of vibration of the material in some form, such as a bar, resonance increases the vibration amplitude considerably to perhaps 1 part in 10^3.

Materials vary widely in the magnitude and sign of the magnetostrictive effect, which also depends on the magnetic field intensity. Pure iron may have positive or negative magnetostriction: the addition of nickel makes the effect positive at all frequencies. With 30% Ni the longitudinal magnetostriction falls to zero, a property utilised in invar–nickel alloys to give very low coefficients of thermal expansion, because the latter is neutralised by the magnetostrictive contraction. In nickel alloys there is a peak dimensional change when the Ni content is about 45% and at about 63% the most sensitive magnetostrictive condition is reached.

11.12.8.1 Sensors

For the purpose of detecting surface and submarine ships, magnetostriction transducers have been employed using under-water supersonic vibrations. The compression stresses generated by the transmitter are reflected by the submerged object and are picked up by a detector utilising the inverse effect—i.e. change of magnetic properties under mechanical stress.

The presence of flaws, air pockets or impurities can similarly be detected in some opaque substances (e.g. rubber) which should normally be homogeneous. The phase displacement between the incident and reflected waves can be interpreted to indicate the position of the obstruction.

Liquid level can be controlled by a sensor probe just above the required level and connected to an oscillator network. As the level of the liquid reaches the probe, the oscillation ceases because of the greatly increased damping. For the measurement of mechanical stress (*Figure 11.52*) the bridge is balanced when the sensor and compensating elements are equally stressed, the compensating element eliminating thermal errors. If the stress in the sensor element is increased, its permeability is reduced, and the out-of-balance current is indicated on the detector D, which can be calibrated in stress units. If the stress is rapidly varying, D can be replaced by a cathode-ray oscilloscope.

11.12.9 Reactance sensors

Many transducers use the properties of inductive and capacitive reactance. The inductance of an iron-cored inductor varies very rapidly with the length of an included air gap. The basic network in *Figure 11.53* employs a bridge with inductors 1 and 2, one with a fixed gap and the other with a gap variable in accordance with some physical quantity such as pressure, strain, thickness, acceleration, etc. The detector D is calibrated appropriately for direct indication. Reliable readings of displacement down to about 3×10^{-5} mm are readily secured with a properly designed instrument. If the displacement fluctuates, it will modulate the supply waveform, and the modulation waveform can be filtered out from the carrier.

Capacitive reactance can also be used, but as the magnitude of its change with displacement is very small, the capacitor sensor is made to be part of the capacitance in an

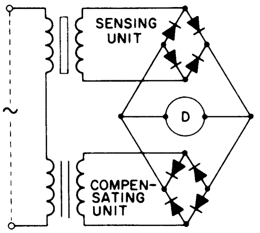


Figure 11.52 Magnetorestriction stress sensor circuit

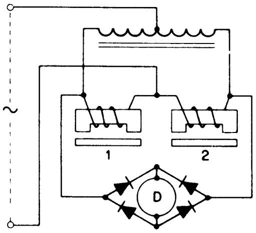


Figure 11.53 Reactance bridge electric guage

oscillatory circuit, the frequency of which can be related to the displacement. One type of capacitor pressure sensor consists of two flexible parallel plates with a 'vacuum' dielectric. A cylindrical capacitor suitable for sensing high pressures has as electrodes an inner deforming cylinder and an outer ring section: dimensional changes in the radial gap occur when the axial force on the inner cylinder causes a change in its diameter. The change of capacitance is well within the range of commercial straingauge bridges.

11.12.9.1 Angular velocity sensors

Measurements can be read from the output of a permanent-magnet d.c. tachometer for speeds up to 3000 rev/min. The instrument is direct-reading, but a proportional direct output voltage can be used as an electrical transducer. For low speeds the commutator ripple can be excessive and an a.c. tachometer with full-wave rectification will provide a more uniform output, particularly if the instrument is of the multi-tooth variable-reluctance type. These instruments act as loads on the source and they are unable to assess the fine detail of the angular velocity such as superimposed small amplitude oscillations, or non-uniform accelerations and retardations due to changes in load. Simple optical solutions to the problem involve the use of a stroboscope or photosensor coupled with digital counters, and each method avoids any loading of the machine.

11.12.10 Stroboscope

The stroboscope generates high-intensity impulsive flashes of light at controllable repetition frequency. If the repetition frequency corresponds to the time of one revolution of a rotating object or to one complete excursion of a reciprocating object (or any multiple thereof), the object appears to be at rest. If the period is a submultiple $1/n$, then the object appears stationary in n different positions. It is desirable for the repetition frequency to exceed 30 Hz in order to avoid visual flicker.

From the flash frequency and observation, the character of velocity perturbations can be viewed and counted and dynamic distortion effects seen. Electrical stroboscopes may involve special neon lamps connected to the secondary side of an induction coil, the primary of which is interrupted by a driven tuning fork. In modern instruments the lamp is a xenon gas discharge tube emitting white light. The flashing frequency is adjusted by means of a transistorised multivibrator-shaper circuit producing pulses of constant energy and fast rise time. Each trigger pulse initiates the discharge of a capacitor to create the flash. Capacitor recharging limits the upper frequency to about 1.5 kHz with a 200 lux illumination level. High frequencies up to 10 kHz can be obtained from a Z-modulated cathode-ray tube as a low-level light source.

11.12.11 Photosensors

In this context, 'light' usually means radiation at wavelengths covering the visible spectrum from the near infrared to the near ultraviolet. Light transducers can be identified by three basic forms of physical behaviour: photovoltaic, photoresistive/photojunction and photoemissive.

11.12.11.1 Photovoltaic

Photovoltaic sensors are p–n junction devices with the normal barrier layer potential present between the two materials. When illuminated, the higher-energy protons raise electrons from the valence into the conduction band to create electron–hole pairs. The consequent continual charge separation is enough to drive current through a resistive load without any external source.

Selenium cell This has a linear current/illumination characteristic for load resistances up to 100 Ω, but the relation is progressively more non-linear for higher resistance, and reaches an open-circuit e.m.f. of 0.6 V. The peak spectral response is in the centre of the visible band (0.57 μm), but the energy efficiency is only 0.5%.

Silicon p–n junction cell The energy conversion is 10–15%, which is adequate for solar cell use in space vehicles. As the time response (a few microseconds) is very fast, silicon photocell arrays are used for reading punched-card and tape, and for optical tracking. The output of a single sensor is typically 70 μA for 5000 lux, the current being constant up to 200 mV. When they are reverse-biased, silicon photosensors behave as photoresistive devices.

11.12.11.2 Photoresistive

A photoresistive sensor consists of a single homogeneous semiconductor material such as n-type cadmium sulphide or cadmium selenide. The material is doped to permit of a large electron charge amplification by the selective absorption of holes, and a spectral response that can simulate that of the eye. This type of sensor is well known as a photographic exposure meter. The CdS sensor is not quite linear over wide illumination ranges, and its response time (100 ms), is slow compared with that (10 ms) of the CdSe type.

The sensors can be used in series with a source and a relay, the latter operating when illumination reduces the sensor resistance to about 1 kΩ. The power dissipation by the sensor can be minimised if the slow resistance changes are used to alter the state of a conventional Schmitt trigger circuit (*Figure 11.54*). The required illumination level is set by R_1; when the illumination is reduced, the cell resistance rises to switch on T_1, which causes T_2 to switch off and T_3 then saturates. The relay operates when the Zener diode conducts at about 12 V. The sensor is useful for alarm circuits, street-lighting control and low-rate counting.

11.12.11.3 Photojunction

Normal diodes and transistors are light-shielded. In the photojunction devices controlled light is admitted to enhance the light effect.

Photodiodes These are normally of silicon unless the greater infra-red response of germanium is needed. Time constants of 10–100 ns are normal. Silicon planar p-type/intrinsic/n-type photodiodes can respond to laser pulses of 1 ns, and are usable at very low light level equivalent to a

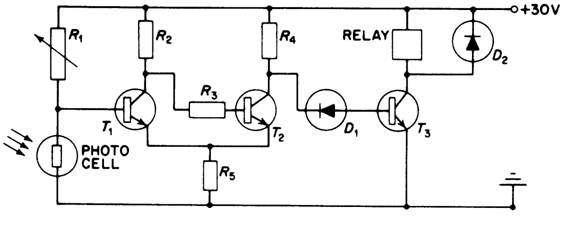


Figure 11.54 Photoresistive cell in Schmitt trigger circuit

current of about 100 pA. Special low-noise transistor amplifiers are required.

Phototransistors The normal form is the n–p–n silicon planar transistor, with conventional bias except for the reverse-biased photodiode between base and collector. It behaves as a common-collector transistor amplifier with a photocurrent generator between base and collector. The maximum cut-off frequency is 50 kHz. A derivative of the field effect transistor, the 'photofet', is more sensitive and the gain–bandwidth product is higher.

11.12.11.4 Photoemissive

Photoemissive sensors are vacuum or gas-filled phototubes. The emission of electrons from the cathode occurs when light falls on it, to be collected by a positive anode. Vacuum phototubes have been superseded for light measurement by photodiodes. In the *photomultiplier* tube the vacuum emission is enhanced by secondary emission at a succession of anodes to give an electron multiplication of 10^8 or 10^9, providing significant output for flashes of very short duration. The photomultiplier is used as a *scintillation* sensor by viewing the weak light emissions that occur in selected phosphors exposed to α, β- or γ-radiation. These scintillation 'counters' have wide application in spectrophotometry, flying-spot scanning, photon counting and whenever a very weak light signal is to be amplified. An alternative device is the Bendix magnetic multiplier. This consists of a single layer of resistive film with the emissive cathode at one end; by suitable accelerating p.d.s applied along the film, together with shaped magnetic fields, the electrons are multiplied by secondary emission and bounced along the strip to accumulate on an anode to give an overall electron gain of 10^9. With different emissive materials, ultraviolet rays and X-rays can be amplified for subsequent analysis.

Gas filled phototubes are low-frequency devices but have a high signal/noise ratio. They are used for film sound-track sensing, and have been applied to monitor natural gas flames (which emit considerable ultraviolet radiation) to avoid danger from unburnt gas.

11.12.11.5 Selection

Photovoltaic cells do not require an external supply, but they are generally less sensitive than photoresistive types, which do. Photodiodes can be selected for a range of optimum spectral sensitivities and are fast-operating, particularly so for the silicon planar p–i–n type. Phototransistors provide amplification at the cost of a relatively slow response (minimum 20 μs). Photomultiplier tubes provide high amplification, even for single photons.

Integrated-circuit chips about 4 mm^2 in area, produced with progressively more complex networks, can include scores of photodiode arrays, with amplifiers, logic counting networks, scanning circuits, etc., to give such outputs as digital pulse rates proportional to illumination, logic for character recognition and punched-card reading, and as replacements for photomultipliers. These devices find a wide range of application in transducer instrumentation for control and communication.

11.12.12 Nuclear radiation sensors

Radiation sensors are required in particle physics research, for monitoring reactions in nuclear power stations, as gauges for industrial processes, and for measurements employing radioactive isotopes. The emanations used for industrial processes include α, β- and γ-rays with X-rays for special applications such as the examination of welds; the sensor characteristics largely determine the radiation to be processed. The devices include the following.

Low-pressure gas ionisation chambers to measure α-particles by accumulating the electrons released by their collisions with gas molecules, and without subsequent multiplication.

Proportional counters, similar to the foregoing but with a higher accelerating voltage (500–800 V) to give cumulative gas amplification (Townsend avalanche) and an electron gain of 10^5–10^6. The use is for counting α-particles in the presence of β-particles and γ-rays.

Geiger counters operating at 800–1500 V, or at low voltage, for β- and γ-ray detection. The complete ionisation must be quenched between counts, and the dead time restricts the counting rate to about 1000/s.

Scintillation sensors (or 'counters') for α, β- or γ-detection. There are also semiconductor sensors for the same application and for X-rays.

α-, β- and γ-Ray sensors require to be followed by electronic pulse amplifiers and counters. A counter may accumulate a total/fractional scaled count, or it may give a rate-meter display. These outputs, together with pulse height analysers, are used to assess the energy spectra of the received signals. Coincidence counters have multiple input sensors so arranged that only the required type of emanation is measured during any selected period.

The detectors can form the bases for industrial gauges and monitors when used with radioisotope sources. The selective absorption of materials enables β-radiations through paper to indicate the uniformity of the product; liquid-level gauges give correcting signals when the rise of level absorbs the radiation; radioisotope tracers with short half-lives can be introduced into liquid and gas channels to indicate flow rate, uniformity of mixing and the detection of leakages.

11.13 Data recording

Chart recorders responsive to signals at frequencies up to a maximum of 25 kHz include: (1) a.c. ultraviolet, (2) $X-Y$, and (3) analogue and digital strip-chart forms. All provide records of related phenomena that need to be preserved in graphical form.

The visual discrimination of the best-quality trace, when stated as a fraction of the chart width, should not be significantly different from the accuracy and linearity of the display; this condition dictates the width of the chart paper. Many different types of treated chart paper are needed in order to be compatible with the various writing systems mentioned below. Paper is provided in rolls or overlapping folds, the latter being very convenient for quick retrieval of data.

Conventional pens for ink-feed methods use felt, ball-point or nylon disposable tips. The ink cartridges are easily replaced, and the supply is drawn to the tip by capillary action. The nylon pen is suitable for higher writing speeds; to retain a uniform trace, a variable pressure is automatically applied to the ink cartridge. In multiple-display applications, the possibility of easy trace separation by colour is a unique convenience of ink writing.

Many alternative writing methods have been developed in an attempt to overcome the inertia (response time) limitation of conventional ink writing, as well as to improve the quality of the trace. These include: (a) a heated stylus and sensitive plastic-coated paper; (2) mechanical pressure on chemically treated paper; (3) ultraviolet light beams directed on to photographic paper, the trace appearing and becoming fixed within a few seconds of exposure to natural light; (4) the passage of current through treated metallised paper causing the chemical reduction process to develop a black imprint; and (5) electrostatic copying, in which a flat capacitor is formed from conductive paper with a dielectric coating—a very small area can be charged, and particles are attracted to the charged area to provide the visible trace, which may require treatment to make it permanent.

References

1 BRITISH STANDARDS INSTITUTION, *BS 5233 Glossary of terms used in metrology*, Milton Keynes (1986)
2 KIBBLE, B. P., ROBINSON, I. A. and BELISS, J. H., *Metrologia* 27, 173–192 (1990)
3 http://www.npl.co.uk/npl/cem/dclf/index.html
4 GALAKHOVA, O. P., HARKNESS, S., HERMACH, F. L., HIRAYAMA, H., MARTIN, P., ROZDESTVENSKA, T. H. and WILLIAMS, E. S., *IEEE Transactions on Instrumentation and Measurement*, **IM-29**, 396–399 (1980)
5 EMMENS, T., *IEE Electronics and Power*, 166 (February 1981)
6 JONES, L. T., RESSMEYER, J. J. and CLARK, C. A., *Hewlett-Packard Journal*, **32**(4), 23 (1981)
7 IKEDA, Y., *NPL DES Memorandum No. 21* (1976)
8 TURGEL, R. S., *NBS Technical Note No. 870* (June 1975)
9 TURGEL, R. S., *IEEE Transactions on Instrumentation and Measurement*, **IM-23**, 337–341 (1974)
10 STOCKTON J. R., *IEE Electronics Letters*, **13**(14) 406–407 (1977)
11 KNIGHT, R. B. D. and STOCKTON, J. R., *NPL Reports DES60* (1981)
12 STOCKTON, J. R. and CLARKE, F. J. J., *DES71* (August 1981)

12

Industrial Instrumentation

E A Parr MSc, CEng, MIEE, MInstMC
CoSteel Sheerness

Contents

12.1 Introduction 12/3
  12.1.1 Definition of terms 12/3
  12.1.2 Range, accuracy and error 12/3
  12.1.3 Dynamic effects 12/3
  12.1.4 Signals and standards 12/4
  12.1.5 P&ID symbols 12/4

12.2 Temperature 12/6
  12.2.1 General 12/6
  12.2.2 Thermocouples 12/6
  12.2.3 Resistance thermometers 12/8
  12.2.4 Pyrometers 12/8

12.3 Flow 12/9
  12.3.1 General 12/9
  12.3.2 Differential pressure flowmeters 12/9
  12.3.3 Turbine flowmeters 12/10
  12.3.4 Vortex shedding flowmeters 12/12
  12.3.5 Electromagnetic flowmeters 12/13
  12.3.6 Ultrasonic flowmeters 12/13
  12.3.7 Hot wire anenometer 12/14
  12.3.8 Mass flowmeters 12/14

12.4 Pressure 12/16
  12.4.1 General 12/16
  12.4.2 Manometers 12/16
  12.4.3 Elastic sensing elements 12/16
  12.4.4 Piezo elements 12/17
  12.4.5 Force balance systems 12/17
  12.4.6 Vacuum measurement 12/17
  12.4.7 Installation notes 12/19

12.5 Level transducers 12/20
  12.5.1 Float based systems 12/20
  12.5.2 Pressure based systems 12/20
  12.5.3 Electrical probes 12/21
  12.5.4 Ultrasonic transducers 12/22
  12.5.5 Nucleonic methods 12/22
  12.5.6 Level switches 12/23

12.6 Position transducers 12/24
  12.6.1 Introduction 12/24
  12.6.2 The potentiometer 12/24
  12.6.3 Synchros and resolvers 12/24
  12.6.4 Linear variable differential transformer (LVDT) 12/27
  12.6.5 Shaft encoders 12/27
  12.6.6 Variable capacitance transducers 12/27
  12.6.7 Laser distance measurement 12/27
  12.6.8 Proximity switches and photocells 12/28

12.7 Velocity and acceleration 12/29
  12.7.1 Introduction 12/29
  12.7.2 Velocity 12/29
  12.7.3 Accelerometers and vibration transducers 12/30

12.8 Strain gauges, loadcells and weighing 12/31
  12.8.1 Introduction 12/31
  12.8.2 Stress and strain 12/31
  12.8.3 Strain gauges 12/32
  12.8.4 Bridge circuits 12/33
  12.8.5 Load cells 12/34
  12.8.6 Weighing systems 12/35

12.9 Fieldbus systems 12/35

12.10 Installation notes 12/39

12.1 Introduction

12.1.1 Definition of terms

Accurate measurement of process variables such as flow, pressure and temperature is an essential part of any industrial process. This chapter describes methods of measuring common process variables.

Like most technologies, instrumentation has a range of common terms with precise meanings.

Measured variable and *process variable* are both terms for the physical quantity that is to be measured on the plant (e.g. the level in tank 15). The *measured value* is the actual value in engineering units (e.g. the level is 1252 mm).

A *primary element* or *sensor* is the device which converts the measured value into a form suitable for further conversion into an instrumentation signal, i.e. a sensor connects directly to the plant. An orifice plate is a typical sensor. A *transducer* is a device which converts a signal from one quantity to another (e.g. a PT100 temperature transducer converts a temperature to a resistance). A *transmitter* is a transducer which gives a standard instrumentation signal (e.g. 4–20 mA) as an output signal, i.e. it converts from the process measured value to a signal which can be used elsewhere.

12.1.2 Range, accuracy and error

The *measuring span*, *measuring interval* and *range* are terms which describe the difference between the lower and upper limits that can be measured, (e.g. a pressure transducer which can measure from 30 to 120 bar has a range of 90 bar). The *rangeability* or *turndown* is the ratio between the upper limit and lower limits where the specified accuracy can be obtained. Assuming the accuracy is maintained across the range, the pressure transmitter above has a turndown of 4:1. Orifice plates and other differential flow meters lose accuracy at low flows, and their turndown is less than the theoretical measuring range would imply.

The *error* is a measurement of the difference between the measured value and the true value. The *accuracy* is the maximum error which can occur between the process variable and the measured value when the transducer is operating under specified conditions. Error can be expressed in many ways. The commonest are absolute value (e.g.+/−2°C for a temperature measurement), as a percentage of the actual value, or as a percentage of full scale. Errors can occur for several reasons; calibration error, manufacturing tolerances and environmental effects are common.

Many devices have an inherent coarseness in their measuring capabilities. A wire wound potentiometer, for example, can only change its resistance in small steps and digital devices such as encoders inherently measure in discrete steps. The term *resolution* is used to define the smallest steps in which a reading can be made.

In many applications the accuracy of a measurement is less important than its consistency. The consistency of a measurement is defined by the terms *repeatability* and *hysteresis*.

Repeatability is defined as the difference in readings obtained when the same measuring point is approached several times from the same direction.

Hysteresis occurs when the measured value depends on the direction of approach as *Figure 12.1*. Mechanical backlash or stiction are common causes of hysteresis.

The accuracy of a transducer will be adversely affected by environmental changes, particularly temperature cycling, and will degrade with time. Both of these effects will be seen as a *zero shift* or a change of sensitivity (known as a *span error*).

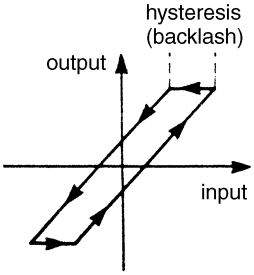


Figure 12.1 Hysteresis, also known as backlash. The output signal is different for an increasing or decreasing input signal

12.1.3 Dynamic effects

A sensor cannot respond instantly to changes in the measured process variable. Commonly the sensor will change as a first order lag as *Figure 12.2* and represented by the equation:

$$\mathrm{T}\frac{\mathrm{d}x}{\mathrm{d}t}+x=a$$

Here *T* is called the time constant. For a step change in input, the output reaches 63% of the final value in time *T*. *Table 12.1* shows the change at later times.

It follows that a significant delay may occur for a dynamically changing input signal.

A second order response occurs when the transducer is analogous to a mechanical spring/viscous damper. The response of such a system to a step input of height *a* is given by the second order equation:

$$\frac{\mathrm{d}^2x}{\mathrm{d}t^2}+2b\omega_\mathrm{n}\frac{\mathrm{d}x}{\mathrm{d}t}+\omega_\mathrm{n}^2x=a$$

where *b* is the damping factor and ω_n the natural frequency. The final steady state value of *x* is given by:

$$x=\frac{a}{\omega_\mathrm{n}^2}$$

The step response depends on both *b* and ω_n, the former determining the overshoot and the latter the speed of

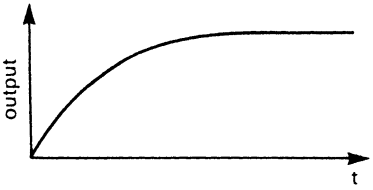


Figure 12.2 Response of a first order lag to a step input signal

Table 12.1

Time	*% Final value*
T	63
2T	86
3T	95
4T	98
5T	99

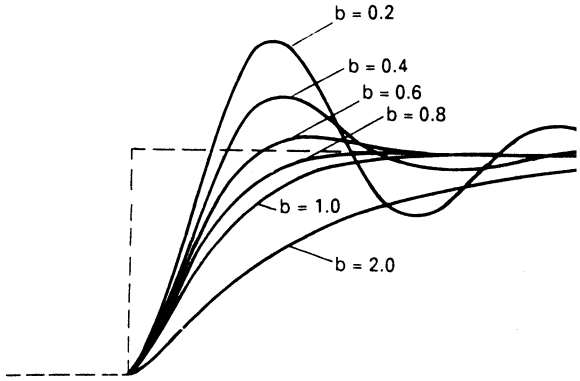


Figure 12.3 Response of a second order lag to a step input signal for various values of damping factor

response as shown on *Figure 12.3*. For values of $b < 1$ damped oscillations occur. The case where $b = 1$ is called *critical damping*. For $b > 1$, the system behaves as two first order lags in series.

Intuitively one would assume that $b = 1$ is the ideal value. This may not always be true. If an overshoot to a step input signal can be tolerated a lower value of b will give a faster response and settling time within a specified error band. The signal enters the error band then overshoots to a peak which is just within the error band as shown on *Figure 12.4*. Many instruments have a damping factor of 0.7 which gives the fastest response time to enter, and stay within, a 5% settling band. *Table 12.2* shows optimum damping factors for various settling bands. The settling time is in units of $1/\omega_n$.

Table 12.2

Settling band	*Optimum 'b'*	*Settling time*
20%	0.45	1.8
15%	0.55	2
10%	0.6	2.3
5%	0.7	2.8
2%	0.8	3.5

12.1.4 Signals and standards

The signals from most primary sensors are very small and in an inconvenient form for direct processing. Commercial transmitters are designed to give a standard output signal for transmission to control and display devices.

The commonest electrical standard is the 4–20 mA loop. As its name implies this uses a variable current with 4 mA representing one end of the signal range and 20 mA the other. This use of a current (rather than a voltage) gives excellent noise immunity as common mode noise has no effect and errors from different earth potentials around the plant are avoided. Because the signal is transmitted as a current, rather than a voltage, line resistance has no effect.

Several display or control devices can be connected in series provided the total loop resistance does not rise above some value specified for the transducer (typically 250 to 1 kΩ).

Transducers using 4–20 mA can be current sourcing (with their own power supply) as *Figure 12.5(a)* or designed for two wire operation with the signal line carrying both the supply and the signal as *Figure 12.5(b)*. Many commercial controllers incorporate a 24 V power supply designed specifically for powering two wire transducers.

The use of the offset zero of 4 mA allows many sensor or cable faults to be detected. A cable break or sensor fault will cause the loop current to fall to zero. A cable short circuit will probably cause the loop current to rise above 20 mA. Both of these fault conditions can be easily detected by the display or control device.

12.1.5 P&ID symbols

Instruments and controllers are usually part of a large control system. The system is generally represented by a *Piping & Instrumentation drawing* (or P&ID) which shows the devices, their locations and the method of interconnection. The description *Process & Instrumentation diagram* is also used by some sources. The letters P&ID should not be confused with a PID controller described in Chapter 13. The basic symbols and a typical example are shown on *Figure 12.6*.

The devices are represented by circles called balloons. These contain a unique tag which has two parts. The first is two or more letters describing the function. The second

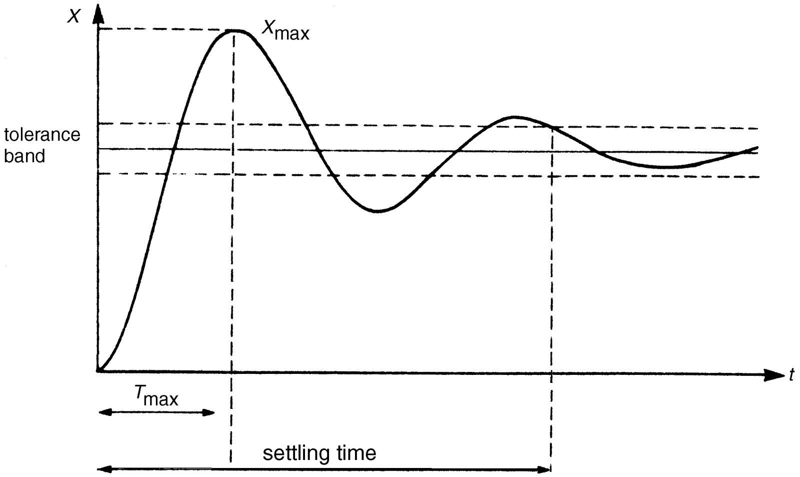


Figure 12.4 Overshoot on a second order system and definition of settling time

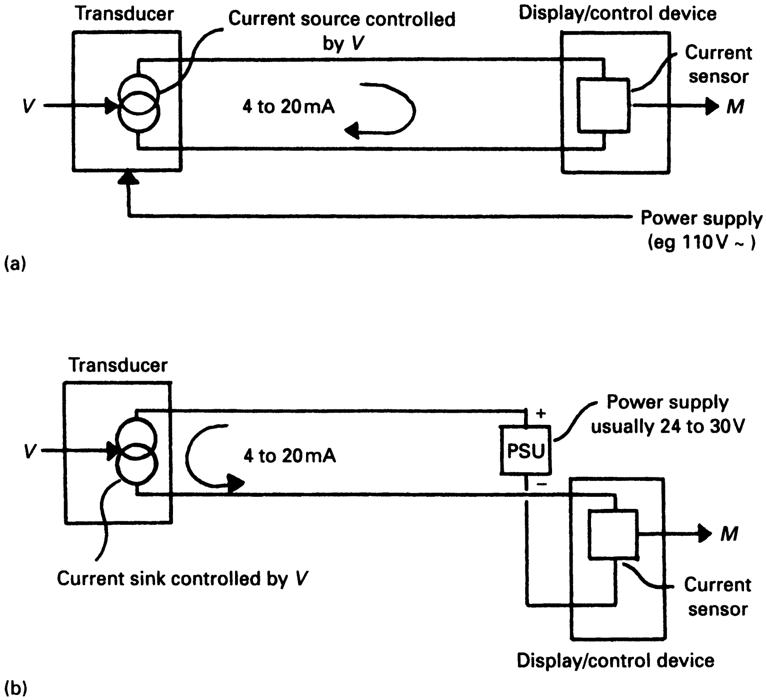

Transducer
Current source controlled by V
Display/control device
Current sensor
V
4 to 20 mA
M
Power supply (eg 110 V ~)
(a)
Transducer
Power supply usually 24 to 30 V
+
PSU
−
V
4 to 20 mA
M
Current sink controlled by V
Current sensor
Display/control device
(b)

Figure 12.5 Current loop circuits. (a) Current sourcing with a self powered transducer; (b) Current sinking with a loop powered (two wire) transducer

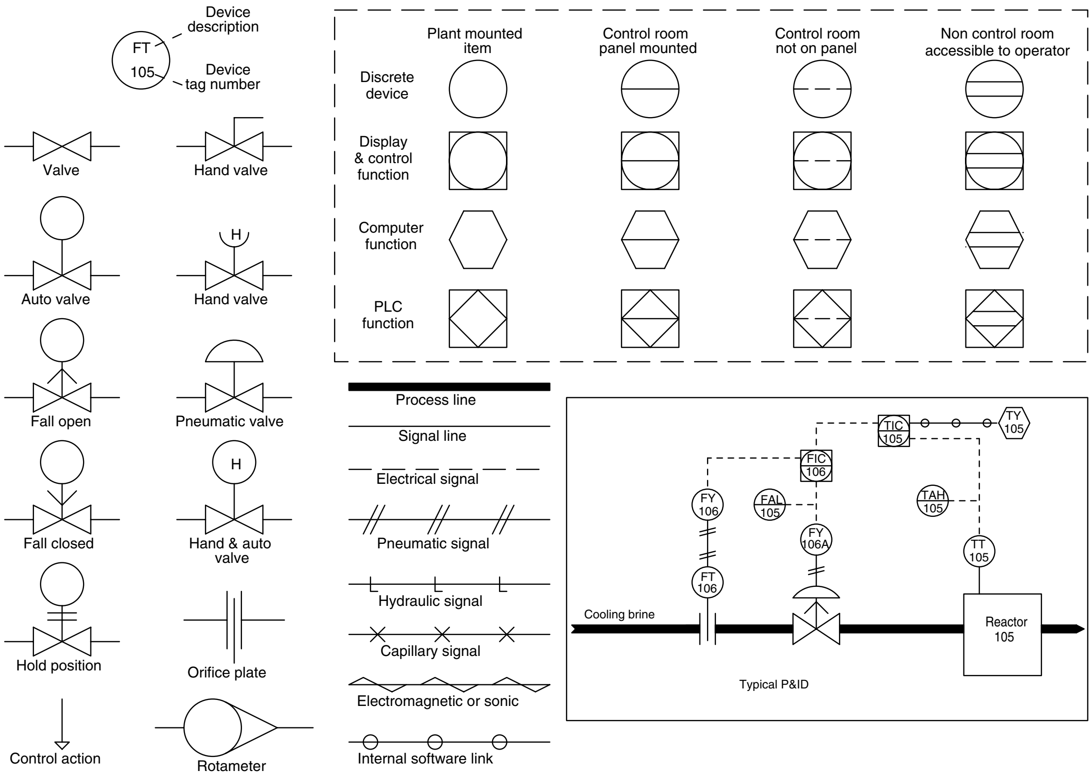

FT
105
Device description
Device tag number
Valve
Hand valve
Auto valve
H
Hand valve
Fall open
Pneumatic valve
Fall closed
H
Hand & auto valve
Hold position
Orifice plate
Control action
Rotameter
Plant mounted item
Control room panel mounted
Control room not on panel
Non control room accessible to operator
Discrete device
Display & control function
Computer function
PLC function
Process line
Signal line
Electrical signal
Pneumatic signal
Hydraulic signal
Capillary signal
Electromagnetic or sonic
Internal software link
TIC 105
TY 105
FIC 106
FY 106
FAL 105
TAH 105
FY 106A
TT 105
FT 106
Cooling brine
Reactor 105
Typical P&ID

Figure 12.6 Common Piping & Instrumentation Diagram (P&ID) symbols and simple schematic

part is a number which uniquely identifies the device, for example FE127 is flow sensor number 127.

The meanings attached to the letters are:

	First letter	*Second and subsequent letters*
A	Analysis	Alarm (often followed by H for high or L for low)
B	Burner	
C	Conductivity	Control function
D	Density	
E	Voltage or misc. electrical	Primary element (sensor)
F	Flow	
G	Gauging	Glass sight tube (e.g. level)
H	Hand	High (with A = alarm)
I	Current	Indicator
J	Power	
K	Time	Control station
L	Level	Light. Also Low (with A = alarm)
M	Moisture	
O		Orifice
P	Pressure	Point
Q	Concentration or Quantity	Integration (e.g. flow to volume)
R	Radioactivity	Recorder
S	Speed	Switch or contact
T	Temperature	Parameter to signal conversion
U	Multivariable	Multifunction
V	Viscosity	Valve
W	Weight, force	Well
X		Signal to signal conversion
Y		Relay
Z	Position	Drive

It is good engineering practice for plant devices to have their P&ID tag physically attached to them to aid maintenance.

12.2 Temperature

12.2.1 General

Accurate knowledge and control of temperature is required in the majority of industrial processes.

12.2.2 Thermocouples

If two dissimilar metals are joined together as shown in *Figure 12.7*(*a*) and one junction is maintained at a high temperature with respect to the other, a current will flow which is a function of the two temperatures. This current, known as the *Peltier effect*, is the basis of a temperature sensor called a thermocouple. In practice, it is more convenient to measure the voltage difference between the two wires rather than current. The voltage, typically a few mV, is again a function of the temperatures at the meter and the measuring junction.

In practice the meter will be remote from the measuring point. If normal copper cables were used to link the meter and the thermocouple, the temperature of these joints would not be known and further voltages, and hence errors, would be introduced. The thermocouple cables must therefore be run back to the meter. Two forms of cable are used; *extension cables* which are essentially identical to the thermocouple cable or *compensating cables* which match the thermocouple characteristics over a limited temperature range. Compensating cables are much cheaper than extension cables.

Because the indication is a function of the temperature at both ends of the cable, correction must be made for the local meter temperature. A common method, called *Cold Junction Compensation,* measures the local temperature by some other method, (such as a resistance thermometer) and adds in a correction as *Figure 12.7*(*b*).

Thermocouples are non-linear devices and the voltage can be represented by an equation of the form:

$$V = \mathrm{a} + \mathrm{b}T + \mathrm{c}T^2 + \mathrm{d}T^3 + \mathrm{e}T^4 + \cdots \Leftarrow$$

where a, b, c, d etc. are constants (not necessarily positive) and T is the temperature. Linearising circuits must be provided in the meter if readings are to be taken over an extended range.

Although a thermocouple can be made from any dissimilar metals, common combinations have evolved with

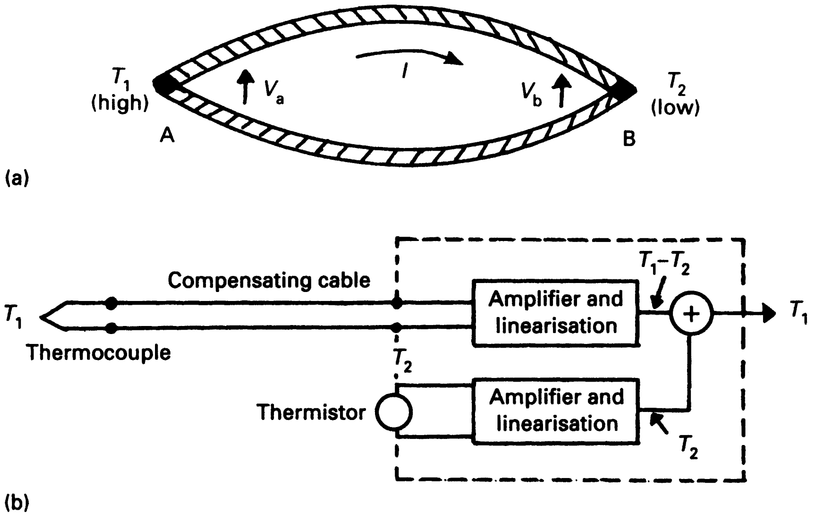


Figure 12.7 Thermocouple circuits. (a) Basic thermocouple; (b) Cold junction compensation

Table 12.3 Common thermocouple types

Type	*+ve Material*	*−ve Material*	*μV/°C*	*Usable Range*	*Comments*
E	Chromel 90% Nickel, 10% Chromium	Constantan 57% Copper 43% Nickel	68.00	0–800°C	Highest output thermocouple
T	Copper	Constantan	46.00	−187 to 300°C	Used for cryogenics and mildy oxidising or reducing atmospheres. Often used for boiler flues
K	Chromel	Alumel	42.00	0 to 1100°C	General purpose. Widely used
J	Iron	Constantan	46.00	20 to 700°C	Used with reducing atmospheres. Tends to rust
R	Platinum with 13% Rhodium	Platinum	8.00	0 to 1600°C	High temperatures (e.g. iron foundries and steel making). Used in UK in preference to type 'S'
S	Platinum with 10% Rhodium	Platinum	8.00	0 to1600°C	As Type R. Used outside the UK
V	Copper	Copper/Nickel	—		Compensating cable for type K to 80°C
U	Copper	Copper/Nickel	—		Compensating cable for types 'R' and 'S' to 50°C

μV/°C is typical over range.
Alumel is an alloy comprising 94% Nickel, 3% Manganese, 2% Aluminium and 1% Silicon.

well documented characteristics. These are identified by single letter codes. *Table 12.3* gives details of common thermocouple types.

Thermocouple tables give the thermocouple voltage at various temperatures when the cold junction is kept at a defined reference temperature, usually 0°C. A typical table for a type K thermocouple is shown on *Table 12.4*. This has entries in 10°C steps, practical tables are in 1°C steps.

Thermocouple tables are used in two circumstances: for checking the voltage from a thermocouple with a millivoltmeter or injecting a test voltage from a millivolt source to test an indicator or controller. Because a thermocouple is essentially a differential device with a non-linear response, in each case the ambient temperature must be known. In the examples below the type K thermocouple from *Table 12.4* is used.

To interpret the voltage from a thermocouple there are four steps:

(1) Measure the ambient temperature and read the corresponding voltage from the thermocouple table. For an ambient temperature of 20°C and reference voltage of 0°C a type K thermocouple this gives 0.798 mV
(2) Measure the thermocouple voltage. Let us assume it is 12.521 mV
(3) Add the two voltages; 12.521 + 0.798 = 13.319 mV
(4) Read the temperature corresponding to the sum. The thermocouple table shows that the voltage corresponding to 330°C is 13.040 mV and interpolation with 41.9 μV/°C gives a temperature of just over 336°C.

To determine the correct injection voltage there are three steps:

(1) As before measure the ambient temperature at the instrument or controller terminals and read the corresponding voltage from the tables. As before we will assume an ambient of 20°C which gives a voltage of 0.798 mV
(2) Find the table voltage corresponding to the required test temperature, say 750°C which gives 31.213 mV
(3) Subtract the ambient voltage from the test temperature voltage. The result of 30.415 mV is the required injection voltage. As before the μV/°C slope can be used to work out voltages for temperatures between the 10°C steps of *Table 12.4*.

Table 12.4 Voltages for a type K thermocouple, voltages in micro volts with reference junction at 0°C

Deg C	*0*	*−10*	*−20*	*−30*	*−40*	*−50*	*−60*	*−70*	*−80*	*−90*	μV/°C
−200	−5891	−6035	−6158	−6262	−6344	−6404	−6441	−6458			
−100	−3554	−3852	−4138	−4411	−4669	−4913	−5141	−5354	−5550	−5730	23.4
0	0	−392	−778	−1156	−1527	−1889	−2243	−2587	−2920	−3243	35.5
Deg C	*0*	*10*	*20*	*30*	*40*	*50*	*60*	*70*	*80*	*90*	
0	0	397	798	1203	1612	2023	2436	2851	3267	3682	41.0
100	4096	4509	4920	5382	5735	6138	6540	6941	7340	7739	40.4
200	8138	8539	8940	9343	9747	10153	10561	10971	11382	11795	40.7
300	12209	12209	12624	13040	13874	14293	14713	15133	15554	15975	41.9
400	16397	16820	17243	17667	18091	18516	18941	19366	19792	20218	42.5
500	20644	21071	21497	21924	22350	22776	23203	23629	24055	24480	42.6
600	24905	25330	25755	26179	26602	27025	27447	27869	28289	28710	42.2
700	29129	29548	29965	30382	30798	31213	31628	32041	32453	32865	41.5
800	33275	33685	34093	34501	34908	35313	35718	36121	36524	36925	40.5
900	37326	37725	38124	38522	38918	39314	39708	40101	40494	40885	39.5
1000	41276	41665	42053	42440	42826	46211	43595	43978	44359	44740	38.4
1100	45119	45497	45873	46249	46623	46995	47367	47737	48105	48473	37.2
1200	48838	49202	49565	49926	50286	50644	51000	51355	51708	52060	35.7
1300	52410	52759	53106	53451	53795	54138	55479	54819			

Note that if the local input terminals at the meter are shorted together, the local ambient temperature (from the cold junction compensation) should be displayed. This is a useful quick check.

The tables show that the voltage from a thermocouple is small, typically less than 10 mV. High gain, high stability amplifiers with good common mode rejection are required and care taken in the installation to avoid noise.

In critical applications a high resistance voltage source is connected across the thermocouple so that in the event of a cable break the meter will indicate a high temperature.

12.2.3 Resistance thermometers

If a wire has resistance R_0 at 0°C its resistance will increase with increasing temperature giving a resistance R_t at temperature T given by:

$$R_t = R_0(1 + aT + bT^2 + cT^3 + \cdots) \Leftarrow$$

where a, b, c etc. are constants. These are not necessarily positive.

This change in resistance can be used to measure temperature. Platinum is widely used because the relationship between resistance and temperature is fairly linear. A standard device is made from a coil of wire with a resistance of 100 Ω at 0°C, giving rise to the common name of a *Pt100 sensor*. These can be used over the temperature range −200°C to 800°C. At 100°C a Pt100 sensor has a resistance of 138.5 Ω, and the 38.5 Ω change from its resistance at 0°C is called the *fundamental interval.*

The current through the sensor must be kept low to avoid heating effects. Further errors can be introduced by the resistance of the cabling to the sensor as shown on *Figure 12.8(a)*. Errors from the cabling resistance can be overcome by the use of the three and four wire connections of *Figure 12.8(b, c* and *d)*. Three wire connections are usually used with a bridge circuit and four wire connections with a constant current source.

A *thermistor* is a more sensitive device. This is a semiconductor crystal whose resistance changes dramatically with temperature. Devices are obtainable which decrease or increase resistance for increasing temperature. The former (decreasing) is more common.

The relationship is very non linear, and is given by:

$$R = R_0 \exp\left[B\left(\frac{1}{T} - \frac{1}{T_0}\right)\right]$$

where R_0 is the defined resistance at temperature T_0, R the resistance at temperature T and B is a constant called the *characteristic temperature*. A typical device will go from 300 kΩ at 0°C to 5 kΩ at 100°C.

Although very non-linear, they can be used for measurement over a limited range. Their high sensitivity and low cost makes them very useful for temperature switching circuits where a signal is required if a temperature goes above or below some preset value. Electric motors often have thermistors embedded in the windings to give early warning of motor overload.

12.2.4 Pyrometers

A heated object emits electromagnetic radiation. At temperatures below about 400°C this radiation can be felt as heat. As the temperature rises the object starts to emit visible radiation passing from red through yellow to white as the temperature rises. Intuitively we can use this radiation to qualitatively measure temperature as below:

Temperature	*Colour*
500°C	Barely visible dull red glow
800°C	Bright red glow
950°C	Orange
1000°C	Yellow
1200°C	White
1500°C	Dazzling white, eyes naturally avert

Pyrometers use the same effect to provide a non contact method of measuring temperature.

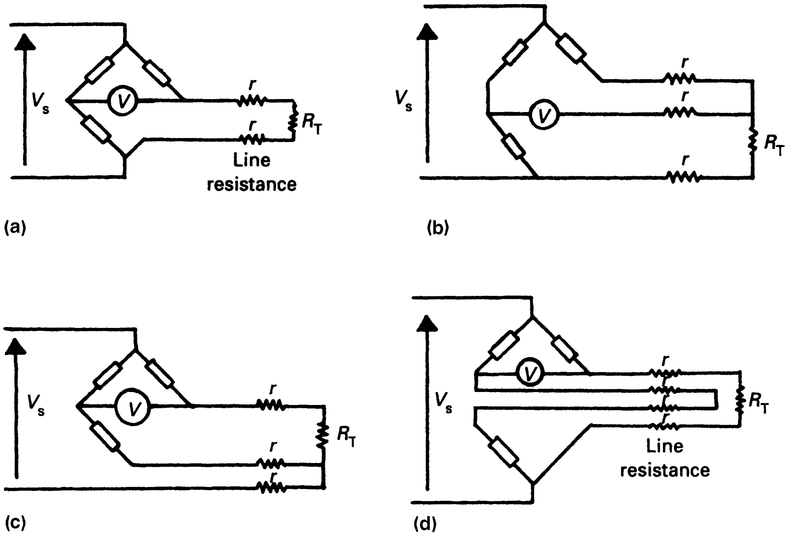


Figure 12.8 The effect of line resistance on RTDs. (a) Simple two wire circuit introduces an error of 2r; (b) Three wire circuit places line resistance into both legs of the measuring bridge; (c) Alternative three wire circuit; (d) Four wire circuit

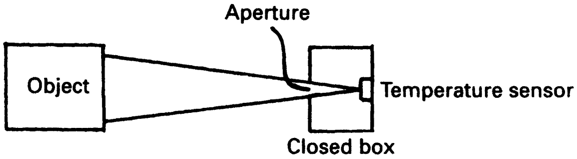


Figure 12.9 The principle of an optical pyrometer

A pyrometer is, in theory, a very simple device as shown on *Figure 12.9*. The object whose temperature is to be measured is viewed through a fixed aperture by a temperature measuring device. Part of the radiation emitted by the object falls on the temperature sensor causing its temperature to rise. The object's temperature is then inferred from the rise in temperature seen by the sensor. The sensor size must be very small, typically of the order of 1mm diameter. Often a circular ring of thermocouples connected in series (called a *thermopile*) is used. Alternatively a small resistance thermometer (called a *bolometer*) may be used. Some pyrometers measure the radiation directly using photo-electric detectors.

A major (and surprising) advantage of pyrometers is the temperature measurement is independent of distance from the object provided the field of view is full. As shown on *Figure 12.10* the source is radiating energy uniformly in all directions, so the energy received from a point is proportional to the solid angle subtended by the sensor. This will decrease as the square of the distance. There is, however, another effect. As the sensor moves away the scanned area also increases as the square of the distance. These two effects cancel giving a reading which is independent of distance.

Although all pyrometers operate on radiated energy there are various ways in which the temperature can be deduced from the received radiation. The simplest measures the total energy received from the object (which is proportional to T^4 where T is the temperature in kelvin). This method is susceptible to errors from lack of knowledge of the emissivity of the object's surface. This error can be reduced by using filters to restrict the measuring range to frequencies where the object's emissivity approaches unity.

An alternative method takes two measurements at two different frequencies (i.e. two different colours) and compares the relative intensities to give an indication of temperature. This method significantly reduces emissivity errors.

Pyrometers have a few major restrictions which should be appreciated. The main problem is they measure surface temperature and only surface temperature. Lack of knowledge of the emissivity is also a major source of error.

12.3 Flow

12.3.1 General

The term 'flow' can generally be applied in three distinct circumstances:

Volumetric flow is the commonest and is used to measure the volume of material passing a point in unit time (e.g. $m^3 \cdot s^{-1}$). It may be indicated at the local temperature and pressure or normalised to some standard conditions using the standard gas law relationship:

$$V_n = \frac{P_m V_m T_n}{P_n T_m}$$

where suffix 'm' denotes the measured conditions and suffix 'n' the normalised condition.

Mass flow is the mass of fluid passing a point in unit time (e.g. $kg \cdot s^{-1}$).

Velocity of flow is the velocity with which a fluid passes a given point. Care must be taken as the flow velocity may not be the same across a pipe, being lower at the walls. The effect is more marked at low flows.

12.3.2 Differential pressure flowmeters

If a constriction is placed in a pipe as *Figure 12.11* the flow must be higher through the restriction to maintain equal mass flow at all points. The energy in a unit mass of fluid has three components:

(1) Kinetic energy given by $mv^2/2$;
(2) Potential energy from the height of the fluid; and
(3) Energy caused by the fluid pressure, called, rather confusingly, *flow energy*. This is given by P/ρ where P is the pressure and ρ the density.

In *Figure 12.11* the pipe is horizontal, so the potential energy is the same at all points. As the flow velocity increases through the restriction, the kinetic energy will increase and, for conservation of energy, the flow energy (i.e. the pressure) must fall.

$$\frac{mv_1^2}{2} + \frac{P_1}{\rho_1} = \frac{mv_2^2}{2} + \frac{P_2}{\rho_2}$$

This equation is the basis of all differential flow meters.

Flow in a pipe can be smooth (called *streamline* or *laminar*) or *turbulent*. In the former case the flow velocity is not equal across the pipe being lower at the walls. With turbulent flow the flow velocity is equal at all points across a pipe.

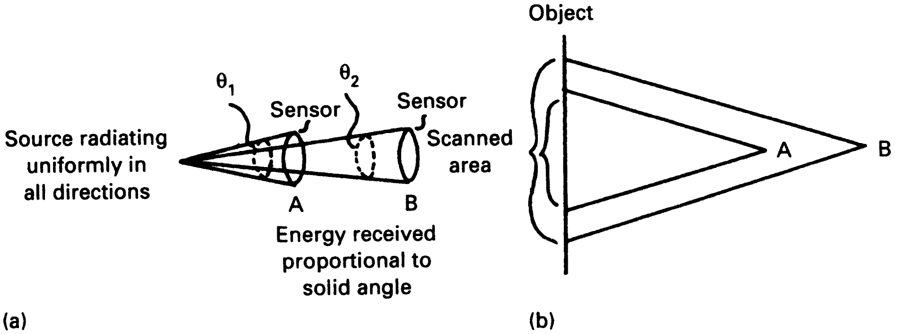

θ1
θ2
Sensor
Sensor
Source radiating uniformly in all directions
Scanned area
A
B
Energy received proportional to solid angle
(a)
Object
A
B
(b)

Figure 12.10 The effect of object distance on a pyrometer. The energy per unit area *decreases* with the square of the distance, however the scanned area *increases* as the square of the distance. Neglecting other influences (such as atmospheric absorption) these effects cancel as the pyrometer reading is independent of distance

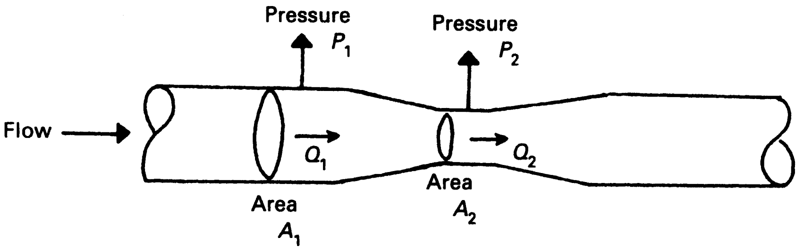


Figure 12.11 The basis of a differential flow meter. Because the mass flow must be equal at all points (i.e. $Q_1 = Q_2$) the flow velocity must increase in the region of A_2. As there is no net gain or loss of energy, the pressure must therefore decrease at A_2

For accurate differential measurement the flow must be turbulent.

The flow characteristic is determined by the *Reynolds number* defined as:

$$R_e = \frac{vD\rho}{\eta}$$

where v is the fluid velocity, D the pipe diameter, ρ the fluid density and η the fluid viscosity. Sometimes the *kinematic viscosity* ρ/η is used in the formula. The Reynolds number is a ratio and has no dimensions. If $R_e < 2000$ the flow is laminar. If $R_e > 10^5$ the flow is fully turbulent.

Calculation of the actual pressure drop is complex, especially for compressible gases, but is generally of the form:

$$Q = K\sqrt{\Delta P} \qquad (12.1)$$

where K is a constant for the restriction and ΔP the differential pressure. Methods of calculating K are given in British Standard BS 1042 and ISO 5167:1980. Computer programs can also be purchased.

The commonest differential pressure flow meter is the orifice plate shown on *Figure 12.12*. This is a plate inserted into the pipe with upstream tapping point at D and downstream tapping point at $D/2$ where D is the pipe diameter. The plate should be drilled with a small hole to release bubbles (liquid) or drain condensate (gases). An identity tag should be fitted showing the scaling and plant identification.

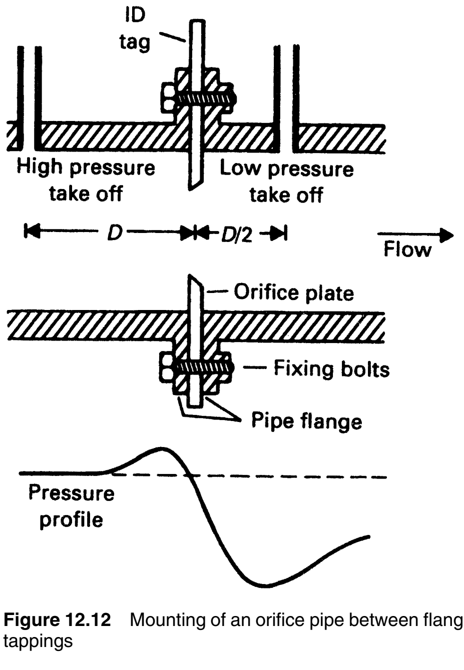


Figure 12.12 Mounting of an orifice pipe between flanges with D–$D/2$ tappings

The D-$D/2$ tapping is the commonest, but other tappings shown on *Figure 12.13* may be used where it is not feasible to drill the pipe.

Orifice plates suffer from a loss of pressure on the down stream side (called the *head loss*). This can be as high as 50%. The venturi tube and Dall tube of *Figure 12.14*(*b*) have lower losses of around 5% but are bulky and more expensive. Another low loss device is the pitot tube of *Figure 12.15*. Equation 12.1 applies to all these devices, the only difference being the value of the constant K.

Conversion of the pressure to an electrical signal requires a differential pressure transmitter and a linearising square root unit. This square root extraction is a major limit on the turndown as zeroing errors are magnified. A typical turndown is 4:1.

The transmitter should be mounted with a manifold block as shown on *Figure 12.16* to allow maintenance. Valves B and C are isolation valves. Valve A is an equalising valve and is used, along with B and C, to zero the transmitter. In normal operation A is closed and B and C are open. Valve A should always be opened before valves B and C are closed prior to removal of the transducer to avoid high pressure being locked into one leg. Similarly on replacement, valves B and C should both be opened before valve A is closed to prevent damage from the static pressure.

Gas measurements are prone to condensate in pipes, and liquid measurements are prone to gas bubbles. To avoid these effects, a gas differential transducer should be mounted above the pipe and a liquid transducer below the pipe with tap off points in the quadrants as shown later in Section 12.4.6.

Although the accuracy and turndown of differential flowmeters is poor (typically 4% and 4:1) their robustness, low cost and ease of installation still makes them the commonest type of flowmeter.

12.3.3 Turbine flowmeters

As its name suggests a turbine flowmeter consists of a small turbine placed in the flow as *Figure 12.17*. Within a specified flow range, (usually with about a 10:1 turndown for liquids, 20:1 for gases,) the rotational speed is directly proportional to flow velocity.

The turbine blades are constructed of ferromagnetic material and pass below a variable reluctance transducer producing an output approximating to a sine wave of the form:

$$E = A\omega \sin(N\omega t)$$

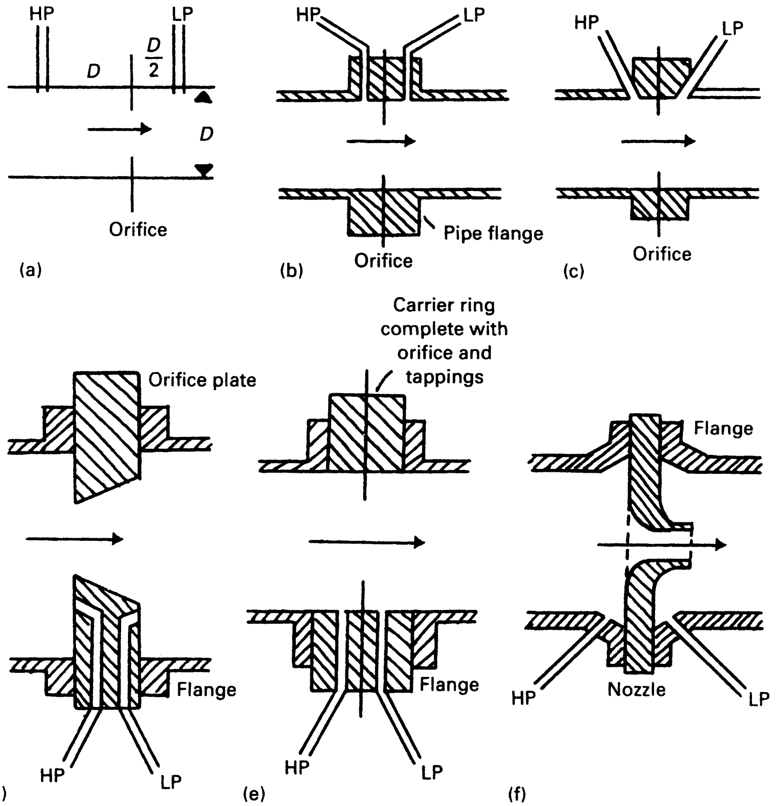

HP
LP
D
$\frac{D}{2}$
D
Orifice
(a)
HP
LP
Pipe flange
Orifice
(b)
HP
LP
Orifice
(c)
Orifice plate
Flange
HP
LP
(d)
Carrier ring complete with orifice and tappings
Flange
HP
LP
(e)
Flange
Nozzle
HP
LP
(f)

Figure 12.13 Common methods of mounting orifice plates. (a) *D*–*D*/2, probably the commonest; (b) Flange taps used on large pipes with substantial flanges; (c) Corner taps drilled through flange; (d) Plate taps, tappings built into the orifice plate; (e) Orifice carrier, can be factory made and needs no drilling on site; (f) Nozzle, gives smaller head loss

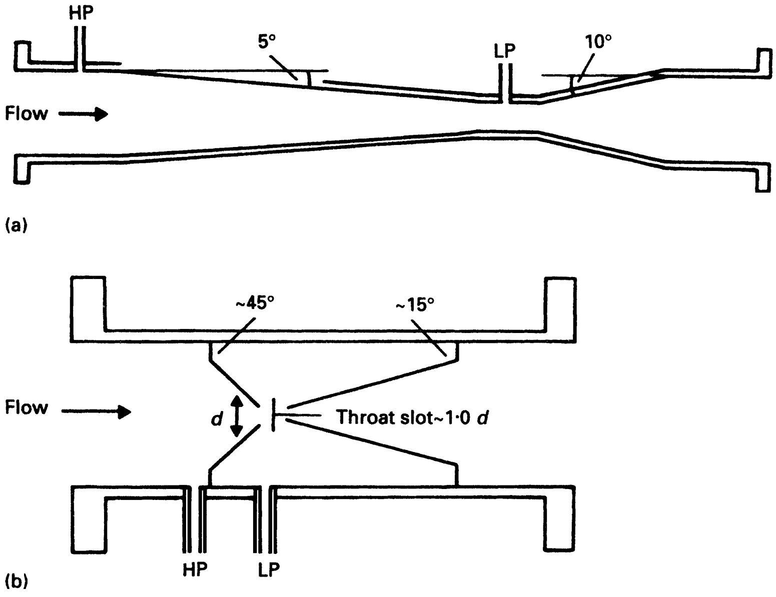

HP
5°
LP
10°
Flow
(a)
~45°
~15°
Flow
d
Throat slot~1·0 d
HP
LP
(b)

Figure 12.14 Low loss differential pressure primary sensors. Both give a much lower head loss than an orifice plate but at the expense of a great increase on pipe length. It is often impossible to provide the space for these devices. (a) Venturi tube; (b) Dall Tube

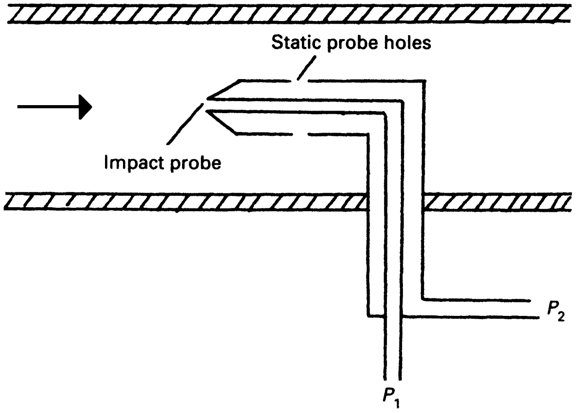

Static probe holes
Impact probe
P_2
P_1

Figure 12.15 An insertion pitot tube

where A is a constant, ω is the angular velocity (itself proportional to flow) and N is the number of blades. Both the output amplitude and the frequency are proportional to flow, although the frequency is normally used.

The turndown is determined by frictional effects and the flow at which the output signal becomes unacceptably low. Other non linearities occur from the magnetic and viscous drag on the blades. Errors can occur if the fluid itself is swirling and upstream straightening vanes are recommended.

Turbine flowmeters are relatively expensive and less robust than other flowmeters. They are particularly vulnerable to damage from suspended solids. Their main advantages are a linear output and a good turndown ratio. The pulse output can also be used directly for flow totalisation.

12.3.4 Vortex shedding flowmeters

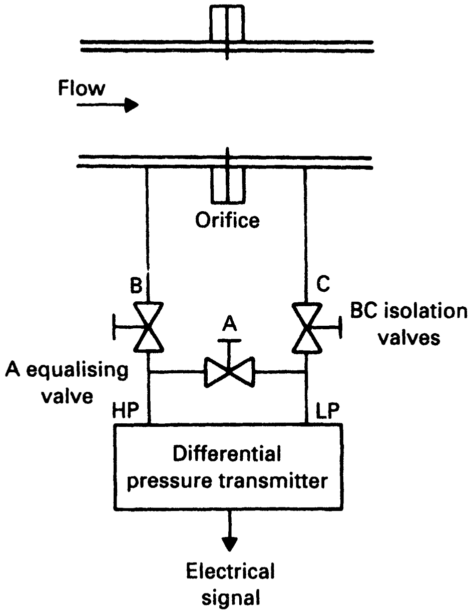

Flow
Orifice
B
C
A
BC isolation valves
A equalising valve
HP
LP
Differential pressure transmitter
Electrical signal

Figure 12.16 Connection of a differential pressure flow sensor such as an orifice plate to a differential pressure transmitter. Valves B and C are used for isolation and valve A for equalisation

If a bluff (non-streamlined) body is placed in a flow, vortices detach themselves at regular intervals from the downstream side as shown on *Figure 12.18*. The effect can be observed by moving a hand through water. In flow measurement the vortex shedding frequency is usually a few hundred Hz. Surprisingly at Reynolds numbers in excess of 10^3 the volumetric flow rate, Q, is directly proportional to the observed frequency of vortex shedding f, i.e.

$$Q = Kf$$

where K is a constant determined by the pipe and obstruction dimensions.

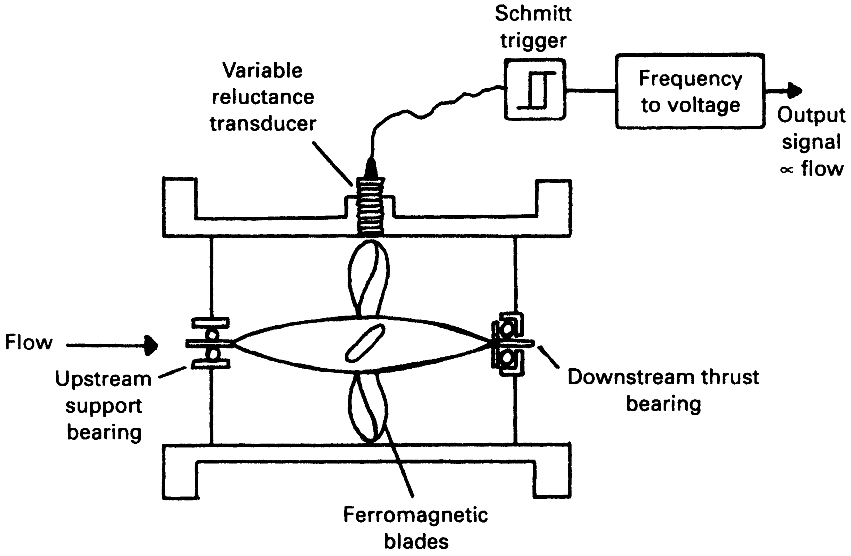

Schmitt trigger
Variable reluctance transducer
Frequency to voltage
Output signal ∝ flow
Flow
Upstream support bearing
Downstream thrust bearing
Ferromagnetic blades

Figure 12.17 A turbine flow meter. These are vulnerable to bearing failures if the fluid contains any solid particles

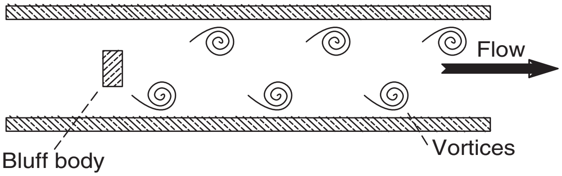


Figure 12.18 Vortex shedding flowmeter

The vortices manifest themselves as sinusoidal pressure changes which can be detected by a sensitive diaphragm on the bluff body or by a downstream modulated ultrasonic beam.

The vortex shedding flowmeter is an attractive device. It can work at low Reynolds numbers, has excellent turndown (typically 15:1), no moving parts and minimal head loss.

12.3.5 Electromagnetic flowmeters

In *Figure 12.19*(*a*) a conductor of length l is moving with velocity v perpendicular to a magnetic field of flux density B. By Faraday's law of electromagnetic induction a voltage E is induced where:

$$E = B \cdot l \cdot v \qquad (12.2)$$

This principle is used in the electromagnetic flowmeter. In *figure 12.19*(*b*) a conductive fluid is passing down a pipe with mean velocity v through an insulated pipe section. A magnetic field B is applied perpendicular to the flow. Two electrodes are placed into the pipe sides to form, via the fluid, a moving conductor of length D relative to the field where D is the pipe diameter. From Equation 12.2 a voltage will occur across the electrodes which is proportional to the mean flow velocity across the pipe.

Equation 12.2 and *Figure 12.19*(*b*) imply a steady d.c. field. In practice an a.c. field is used to minimise electrolysis and reduce errors from d.c. thermoelectric and electrochemical voltages which are of the same order of magnitude as the induced voltage.

Electromagnetic flowmeters are linear and have an excellent turndown of about 15:1. There is no practical size limit and no head loss. They do, though provide a few installation problems as an insulated pipe section is required, with earth bonding either side of the meter to avoid damage from any welding which may occur in normal service. They can only be used on fluids with a conductivity in excess of 1 mS m^{-1} which permits use with many, (but not all), common liquids but prohibits their use with gases. They are useful with slurries with a high solids content.

12.3.6 Ultrasonic flowmeters

The Doppler effect occurs when there is relative motion between a sound transmitter and receiver as shown on *Figure 12.20*(*a*). If the transmitted frequency is f_t Hz, V_s is the velocity of sound and V the relative velocity, the observed received frequency, f_r will be:

$$f_t \frac{(V + V_s)}{V}$$

A Doppler flowmeter injects an ultrasonic sound wave (typically a few hundred kHz) at an angle θ into a fluid moving in a pipe as shown on *Figure 12.20*(*b*). A small part of this beam will be reflected back off small bubbles, solid matter, vortices etc. and is picked up by a receiver mounted alongside the transmitter. The frequency is subject to two changes, one as it moves upstream against the flow, and one as it moves back with the flow. The received frequency is thus:

$$f_r = f_t \frac{(V_s + V\cos(\theta))}{(V_s - V\cos(\theta))}$$

which can be simplified to

$$\Delta f = \frac{2f_t}{V_s} V\cos(\theta)$$

The Doppler flowmeter measures mean flow velocity, is linear, and can be installed (or removed) without the need to break into the pipe. The turndown of about 100:1 is the best of all flowmeters. Assuming the measurement of mean flow velocity is acceptable it can be used at all Reynolds numbers. It is compatible with all fluids and is well suited for difficult applications with corrosive liquids or heavy slurries.

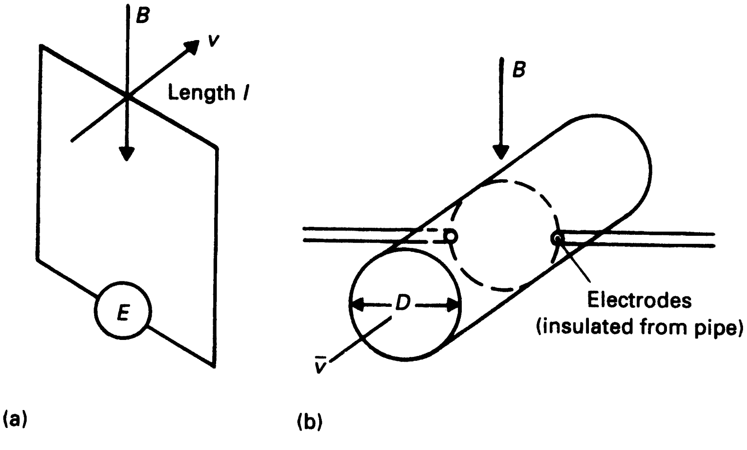


Figure 12.19 Electromagnetic flowmeter. (a) Electromagnetic induction in a wire moving in a magnetic field; (b) The principle applied with a moving conductive fluid

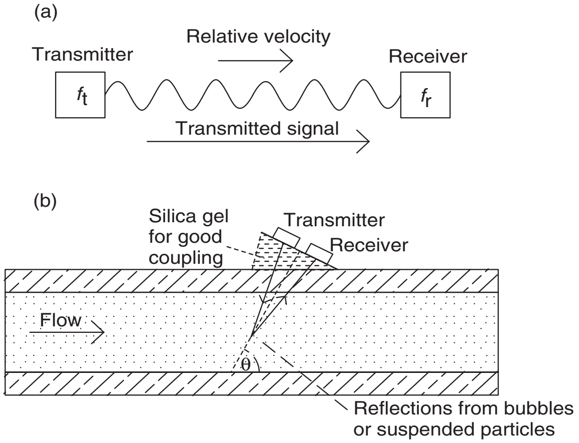

(a)
Relative velocity
Transmitter
Receiver
f_t
f_r
Transmitted signal
(b)
Silica gel for good coupling
Transmitter
Receiver
Flow
θ
Reflections from bubbles or suspended particles

Figure 12.20 An Ultrasonic Flowmeter. (a) Principle of operation; (b) Schematic of a clip on ultrasonic flowmeter

12.3.7 Hot wire anemometer

If fluid passes over a hot object, heat is removed. It can be shown that the power loss is:

$$P = \mathrm{A} + \mathrm{B}\sqrt{v} \qquad (12.3)\Leftarrow$$

where v is the flow velocity and A and B are constant. A is related to radiation and B to conduction.

Figure 12.21 shows a flowmeter based on Equation 12.3. A hot wire is inserted in the flow and maintained at a constant temperature by a self balancing bridge. Changes in the wire temperature result in a resistance change which unbalances the bridge. The bridge voltage is automatically adjusted to restore balance.

The current, I, though the resistor is monitored. With a constant wire temperature the heat dissipated is equal to the power loss from which:

$$v = (I^2R - \mathrm{A})/\mathrm{B}^2$$

Obviously the relationship is non linear, and correction will need to be made for the fluid temperature which will affect constants A and B.

12.3.8 Mass flowmeters

The volume and density of all materials are temperature dependent. Some applications will require true volumetric measurements, some, such as combustion fuels, will really require mass measurement. Previous sections have measured volumetric flow. This section discusses methods of measuring mass flow.

The relationship between volume and mass depends on both pressure and absolute temperature (measured in kelvin). For a gas:

$$\frac{P_1V_1}{T_1} = \Leftarrow\frac{P_2V_2}{T_2}$$

The relationship for a liquid is more complex, but if the relationship is known and the pressure and temperature are measured along with the delivered volume or volumetric flow the delivered mass or mass flow rate can be easily calculated. Such methods are known as *Inferential* flowmeters.

In *Figure 12.22* a fluid is passed over an in-line heater with the resultant temperature rise being measured by two temperature sensors. If the specific heat of the material is constant, the mass flow, F_m, is given by:

$$F_m = \Leftarrow\frac{E}{C_p\theta\varsigma}$$

where E is the heat input from the heater, C_p is the specific heat and $\theta\varsigma$the temperature rise. The method is only suitable for relatively small flow rates.

Many modern mass flowmeters are based on the Coriolis effect. In *Figure 12.23* an object of mass m is required to move with linear velocity v from point A to point B on a surface which is rotating with angular velocity. If the object

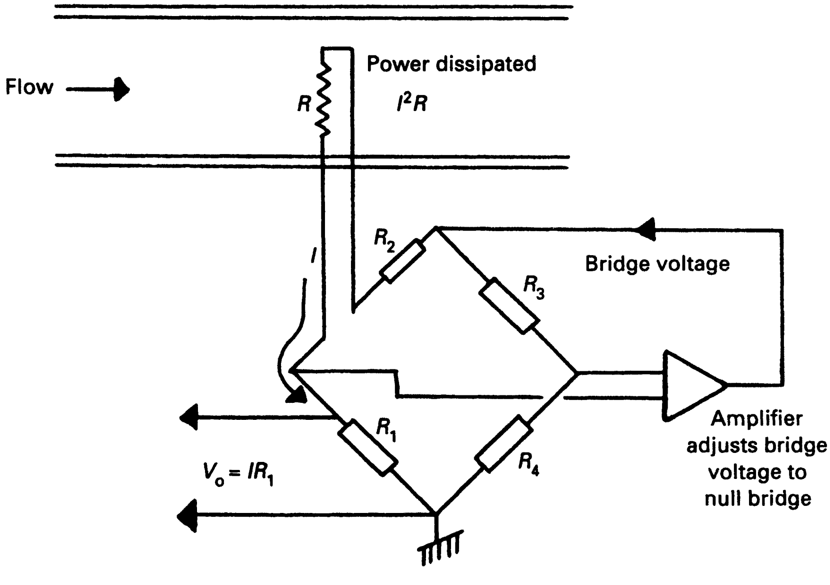

Flow
R
Power dissipated
I^2R
I
R_2
R_3
Bridge voltage
R_1
R_4
$V_o = IR_1$
Amplifier adjusts bridge voltage to null bridge

Figure 12.21 The hot wire anemometer

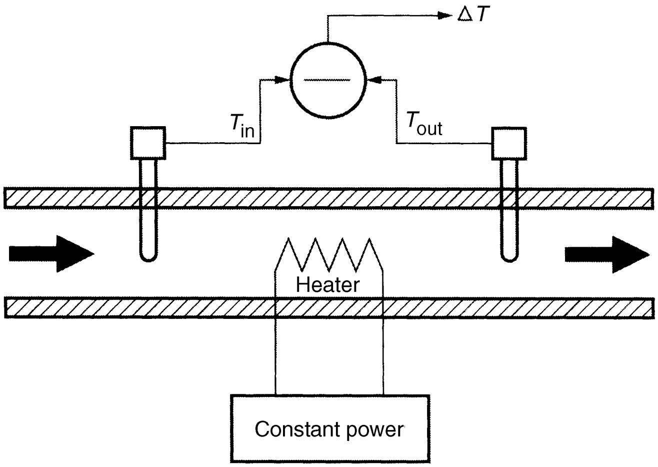

ΔT
T_{in}
T_{out}
Heater
Constant power

Figure 12.22 Mass flow measurement by noting the temperature rise caused by a constant input power

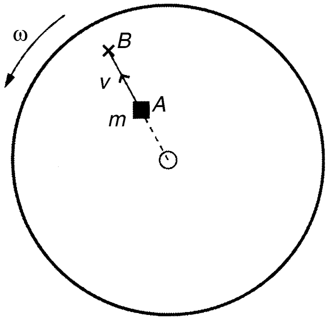


Figure 12.23 Definition of Coriolis force

moves in a straight line as viewed by a static observer, it will appear to veer to the right when viewed by an observer on the disc.

If the object is to move in a straight line as seen by an observer on the disc a force must be applied to the object as it moves out. This force is known as the *Coriolis force* and is given by

$$F = 2m\omega v$$

where m is the mass, ω is the angular velocity and v the linear velocity. The existence of this force can be easily experienced by trying to move along a radius of a rotating children's roundabout in a playground.

Coriolis force is not limited to pure angular rotation, but also occurs for sinusoidal motion. This effect is used as the basis of a Coriolis flowmeter shown on *Figure 12.24*. The flow passes through a 'C' shaped tube which is attached to a leaf spring and vibrated sinusoidally by a magnetic forcer. The Coriolis force arises not, as might be first thought, because of the semi-circle pipe section at the right-hand

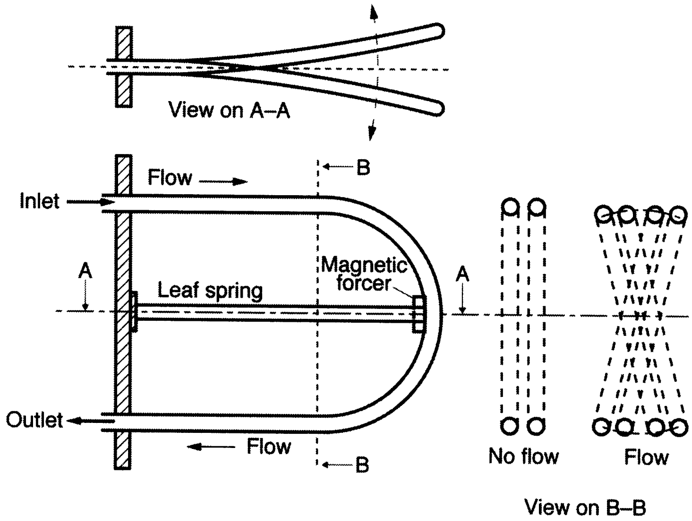

View on A–A
Flow
B
Inlet
A
Magnetic forcer
A
Leaf spring
Outlet
Flow
B
No flow
Flow
View on B–B

Figure 12.24 Simple Coriolis mass flowmeter. A multi-turn coil is often used in place of the 'C' segment

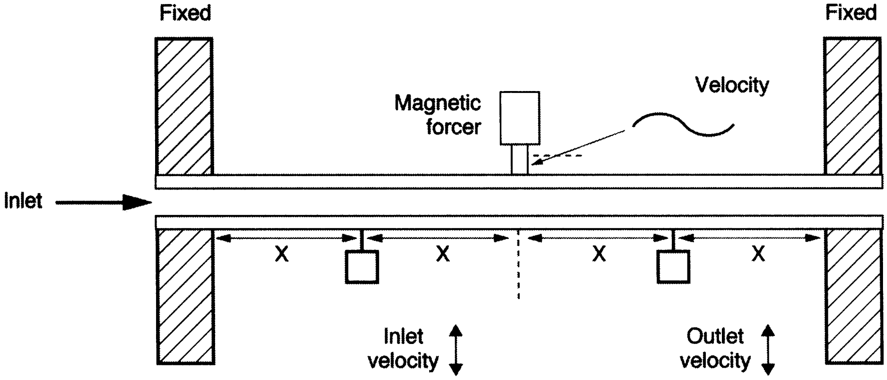


Figure 12.25 Vibrating straight pipe mass flowmeter

side, but from the angular motion induced into the two horizontal pipe sections with respect to the fixed base. If there is no flow, the two pipe sections will oscillate together. If there is flow, the flow in the top pipe is in the opposite direction to the flow in the bottom pipe and the Coriolis force causes a rolling motion to be induced as shown. The resultant angular deflection is proportional to the mass flow rate.

The original meters used optical sensors to measure the angular deflection. More modern meters use a coil rather than a 'C' section and sweep the frequency to determine the resonant frequency. The resonant frequency is then related to the fluid density by:

$$f_c = \Leftarrow \sqrt{\frac{K}{\text{density}}}$$

where K is a constant. The mass flow is determined either by the angular measurement or the phase shift in velocity. Using the resonant frequency maximises the displacement and improves measurement accuracy.

Coriolis measurement is also possible with a straight pipe. In *Figure 12.25* the centre of the pipe is being deflected with a sinusoidal displacement, and the velocity of the inlet and outlet pipe sections monitored. The Coriolis effect will cause a phase shift between inlet and outlet velocities. This phase shift is proportional to the mass flow.

12.4 Pressure

12.4.1 General

There are four types of pressure measurement.

Differential pressure is the difference between two pressures applied to the transducer. These are commonly used for flow measurement as described in Section 12.3.2.

Gauge pressure is made with respect to atmospheric pressure. It can be considered as a differential pressure measurement with the low pressure leg left open. Gauge pressure is usually denoted by the suffix 'g' (e.g. 37 psig). Most pressure transducers in hydraulic and pneumatic systems indicate gauge pressure.

Absolute pressure is made with respect to a vacuum.

$$\text{Absolute pressure} = \text{Gauge pressure} + \text{Atmospheric pressure}$$

Atmospheric pressure is approximately 1 bar, 100 kPa or 14.7 psi.

Head pressure is used in liquid level measurement and refers to pressure in terms of the height of a liquid (e.g. inches water gauge). It is effectively a gauge pressure measurement, but if the liquid is held in a vented vessel any changes in atmospheric pressure will affect both legs of the transducer equally giving a reading which is directly related to liquid height.

The head pressure is given by:

$$P = \rho g h \qquad (12.4)\Leftarrow$$

where P is the pressure in pascals, ρςis the density (kgm^{-2}), g is acceleration due to gravity (9.8 ms^{-2}) and h the column height in metres.

In Imperial units, pounds are a term of weight (i.e. force) not mass so Equation 12.4 becomes

$$P = \rho h$$

where P is the pressure in pounds per square unit (inch or foot), ρςthe density in pounds per cubic unit and h is the height in units.

12.4.2 Manometers

Although manometers are not widely used in industry they give a useful insight into the principle of pressure measurement. If a U-tube is part filled with liquid, and differing pressures applied to both legs as *Figure 12.26*, the liquid will fall on the high pressure side and rise on the low pressure side until the head pressure of liquid matches the pressure difference. If the two levels are separated by a height h then

$$h = (P_1 - P_2)/\rho g$$

where P_1 and P_2 are the pressures (in pascals), ρςis the density of the liquid and g is the acceleration due to gravity.

12.4.3 Elastic sensing elements

The Bourdon tube, dating from the mid nineteenth century, is still the commonest pressure indicating device. The tube is manufactured by flattening a circular cross-section tube to the section shown on *Figure 12.27* and bending it into a C shape. One end is fixed and connected to the pressure to be measured. The other end is closed and left free.

If pressure is applied to the tube it will try to straighten causing the free end to move up and to the right. This

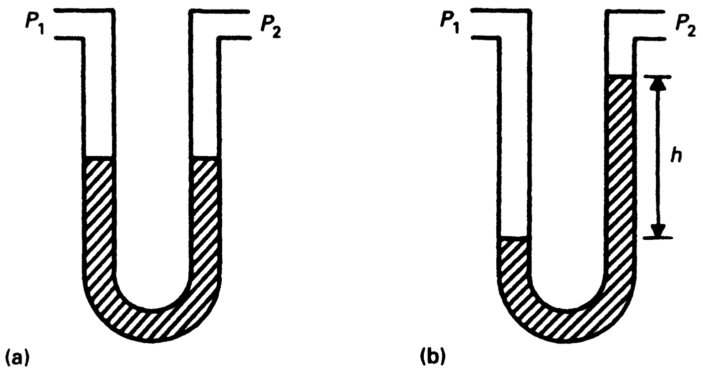


Figure 12.26 The U tube manometer. (a) Pressures P_1 and P_2 are equal; (b) Pressure P_1 is greater than P_2 and the distance h is proportional to $(P_1 - P_2)$

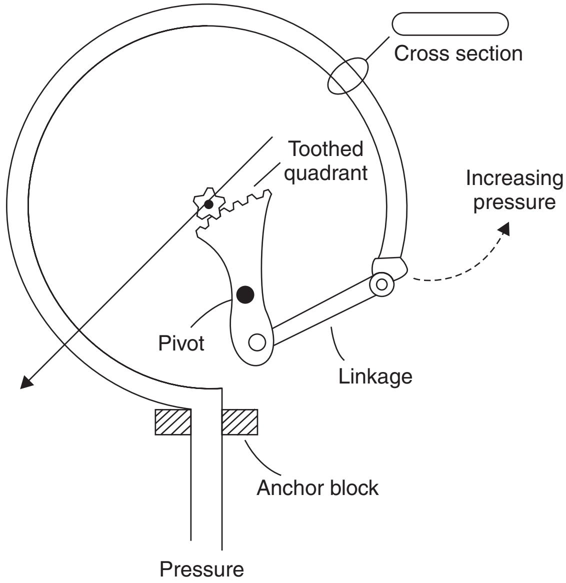


Figure 12.27 The Bourdon tube pressure gauge

motion is converted to a circular motion for a pointer with a quadrant and pinion linkage. A Bourdon tube inherently measures gauge pressure.

Bourdon tubes are usable up to about 50 MPa, (about 10 000 psi). Where an electrical output is required the tube can be coupled to a potentiometer or LVDT.

Diaphragms can also be used to convert a pressure differential to a mechanical displacement. Various arrangements are shown on *Figure 12.28*. The displacement can be measured by LVDTs (see Section 12.6.4), strain gauges (see Section 12.8.3). The diaphragm can also be placed between two perforated plates as *Figure 12.29* The diaphragm and the plates form two capacitors whose value is changed by the diaphragm deflection.

12.4.4 Piezo elements

The piezo electric effect occurs in quartz crystals. An electrical charge appears on the faces when a force is applied. This charge is directly proportional to the applied force. The force can be related to pressure with suitable diaphragms. The resulting charge is converted to a voltage by the circuit of *Figure 12.30* which is called a *charge amplifier*. The output voltage is given by:

$$V = -q/c$$

Because q is proportional to the force, V is proportional to the force applied to the crystal. In practice the charge leaks away, even with FET amplifiers and low leakage capacitors. Piezo-electric transducers are thus unsuitable for measuring static pressures, but they have a fast response and are ideal for measuring fast dynamic pressure changes.

A related effect is the piezo resistive effect which results in a change in resistance with applied force. Piezo resistive devices are connected into a Wheatstone bridge.

12.4.5 Force balance systems

Friction and non-linear spring constants can cause errors in elastic displacement transducers. The force balance system uses closed loop control to balance the force from the pressure with an opposing force which keeps the diaphragm in a fixed position as shown on *Figure 12.31*. As the force from the solenoid is proportional to the current, the current in the coil is directly proportional to the applied pressure. LVDTs are commonly used for the position feedback.

12.4.6 Vacuum measurement

Vacuum measurement is normally expressed in terms of the height of a column of mercury which is supported by the vacuum (e.g. mmHg). Atmospheric pressure (approximately 1 bar) thus corresponds to about 760 mmHg, and an absolute vacuum is 0 mmHg. The term '*torr*' is generally used for 1 mmHg. Atmospheric pressure is thus about 760 torr.

Conventional absolute pressure transducers are usable down to about 20 torr with special diaphragms. At lower pressures other techniques, described below, must be used. At low pressures the range should be considered as logarithmic, i.e. 1 to 0.1 torr is the same 'range' as 0.1 torr to 0.01 torr.

The first method, called the *Pirani gauge*, applies constant energy to heat a thin wire in the vacuum. Heat is lost from a hot body by conduction, convection and radiation. The first two losses are pressure dependent. The heat loss

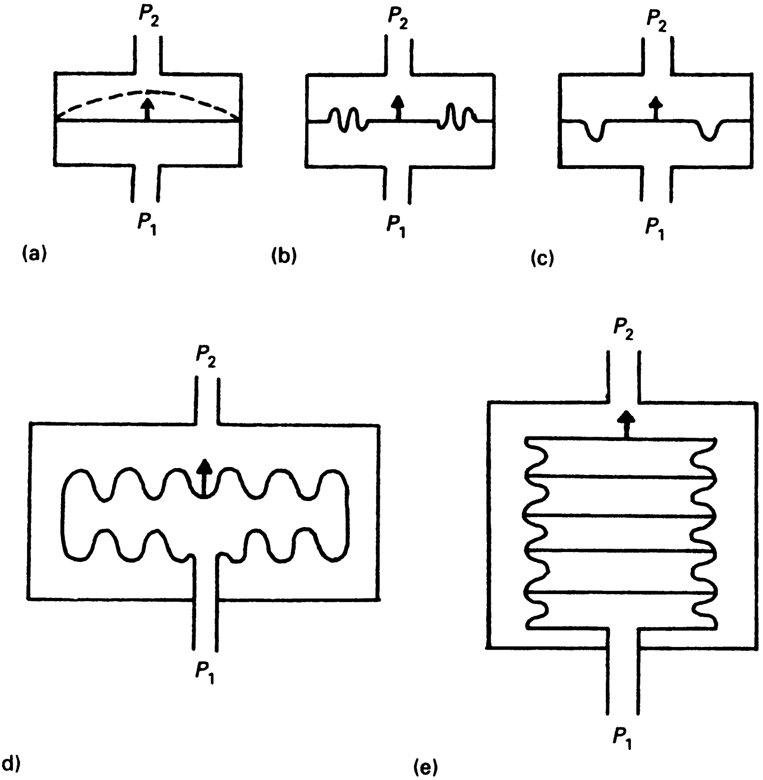


Figure 12.28 Various arrangements of elastic pressure sensing elements. (a) Diaphragm; (b) Corrugated diaphragm; (c) Catenary diaphragm; (d) Capsule; (e) Bellows

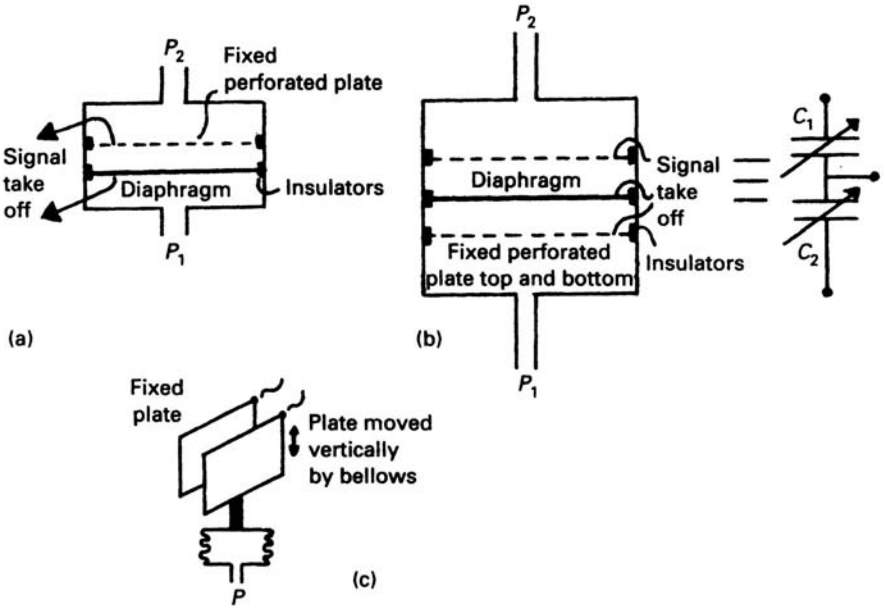

P_2
Fixed perforated plate
Signal take off
Diaphragm
Insulators
P_1
(a)
P_2
Diaphragm
Signal take off
Fixed perforated plate top and bottom
Insulators
C_1
C_2
P_1
(b)
Fixed plate
Plate moved vertically by bellows
P
(c)

Figure 12.29 Variable capacitance pressure sensing elements. (a) Single capacitor variable spacing; (b) Double capacitor variable spacing; (c) Variable area

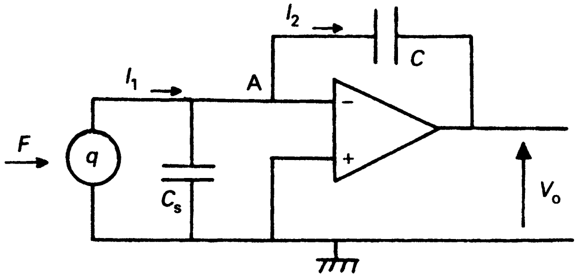


Figure 12.30 The charge amplifier

decreases as the pressure falls causing the temperature of the wire to rise. The temperature of the wire can be measured from its resistance (see Section 12.2.3) or by a separate thermocouple.

The range of a Pirani gauge is easily changed by altering the energy supplied to the wire. A typical instrument would cover the range 5 torr to 10^{-5} torr in three ranges. Care must be taken not to overheat the wire which will cause a change in the heat loss characteristics.

The second technique, called an *ionisation gauge*, is similar to a thermionic diode. Electrons emitted from the heated cathode cause a current flow which is proportional to the absolute pressure. The current is measured with high sensitivity microammeters. Ionisation gauges can be used over the range 10^{-3} to 10^{-12} torr.

12.4.7 Installation notes

To avoid a response with a very long time constant the pipe run between a plant and a pressure transducer must be kept as short as possible. This effect is particularly important with gas pressure transducers which can exhibit time constants of several seconds if the installation is poor.

Entrapped air can cause significant errors in liquid pressure measurement, and similarly condensed liquid will cause errors in gas pressure measurement. Ideally the piping should be taken from the top of the pipe and rise to the transducer for gas systems, and be taken from the lower half of the pipe and fall to the transducer for liquid pressure measurement. If this is not possible, vent/drain cocks must be fitted. Care must be taken to avoid sludge build up around the tapping with liquid pressure measurement. Typical installations are shown on *Figure 12.32*. Under no circumstances should the piping form traps where gas bubbles or liquid sumps can form.

Steam pressure transducers are a special case. Steam can damage the transducer diaphragm, so the piping leg to the transducer is normally arranged to naturally fill with water by taking the tapping of the top of the pipe and mounting the transducer (with unlagged pipes) below the pipe as *Figure 12.33*(*a*). If this is not possible, a steam trap is placed in the piping as *Figure 12.33*(*b*).

In both cases there will be an offset from the liquid head. A similar effect occurs when a liquid pressure transmitter is situated below the pipe. In both cases the offset pressure can usually be removed in the transducer setup.

Pressure transducers will inevitably need to be changed or maintained on line, so isolation valves should always be fitted. These should include an automatic vent on closure so pressure cannot be locked into the pipe between the transducer and the isolation valve.

Where a low pressure range differential transducer is used with a high static pressure, care must be taken to avoid damage by the static pressure getting 'locked in' to one side. *Figure 12.16* (Section 12.3.2) shows a typical manifold for a differential transducer. The bypass valve should always be opened first on removal and closed last on replacement.

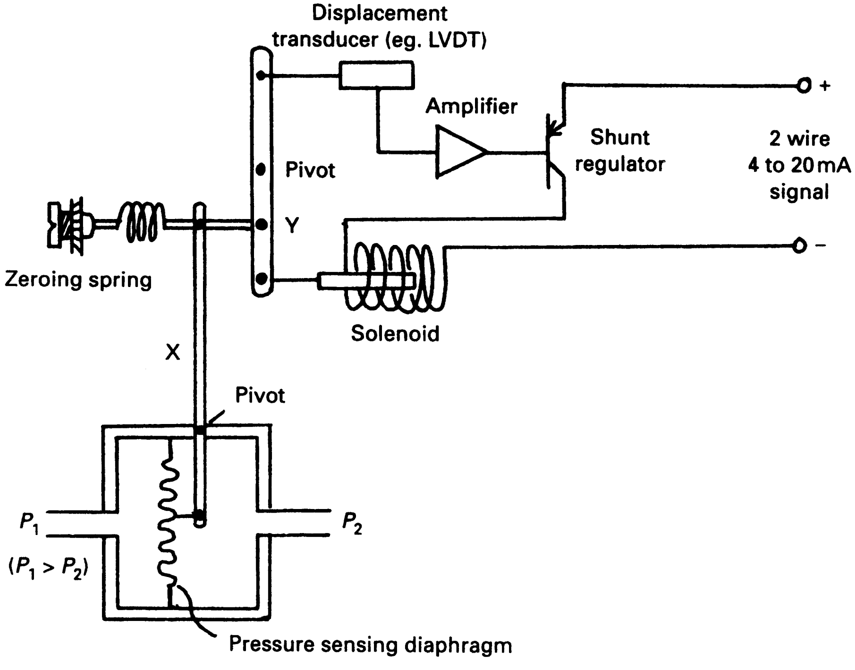


Figure 12.31 A force balance pressure transducer

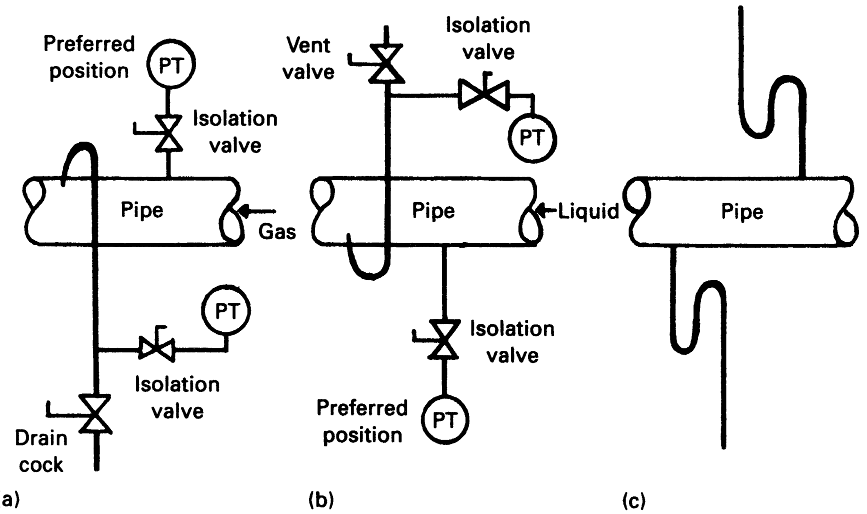

Preferred position
PT
Isolation valve
Pipe
Gas
PT
Isolation valve
Drain cock
(a)
Vent valve
Isolation valve
PT
Pipe
Liquid
Isolation valve
Preferred position
PT
(b)
Pipe
(c)

Figure 12.32 Installation of pressure transducers. Similar considerations apply to differential pressure transducers used with orifice plates and similar primary flow sensors. (a) Gas pressure transducer; (b) Liquid pressure transducer; (c) Faulty installations with potential traps

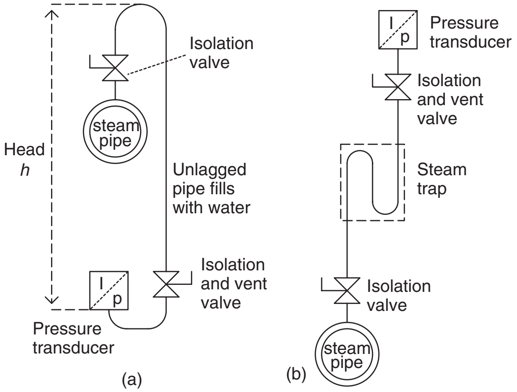

Isolation valve
steam pipe
Head h
Unlagged pipe fills with water
I p
Isolation and vent valve
Pressure transducer
(a)
I p
Pressure transducer
Isolation and vent valve
Steam trap
Isolation valve
steam pipe
(b)

Figure 12.33 Piping arrangements for steam pressure transducers. (a) Preferred arrangement; (b) Alternative arrangement

12.5 Level transducers

12.5.1 Float based systems

A float is the simplest method of measuring or controlling level. *Figure 12.34* shows a typical method of converting the float position to an electrical signal. The response is non linear, with the liquid level being given by:

$$h = H - l \cdot \sin(\theta) \Leftarrow$$

To minimise errors the float should have a large surface area.

12.5.2 Pressure based systems

The absolute pressure at the bottom of a tank has two components and is given by:

$$P = \rho g h + \text{atmospheric pressure}$$

where ρ is the density, g is gravitational acceleration and h the depth of the liquid. The term $\rho g h$ is called the head pressure. Most pressure transducers measure gauge pressure, (see Section 12.4.1), so the indicated pressure will be directly related to the liquid depth. It should be noted that 'h' is the height difference between the liquid surface and the pressure transducer itself. An offset may need to be added, or subtracted, to the reading as shown on *Figure 12.35(a)*.

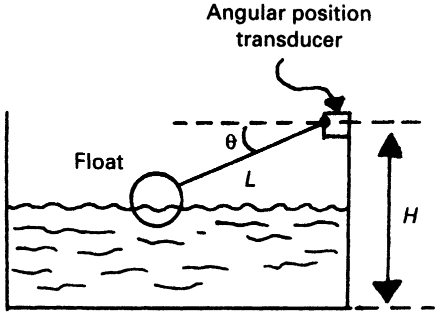

Angular position transducer
θ
Float
L
H

Figure 12.34 Lever arm float based level measurement system

If the tank is pressurised, a differential pressure transmitter must be used as *Figure 12.35(b)*. This arrangement may also require a correction if the pressure transducer height is not the same as the tank bottom. Problems can also occur if condensate can form in the low pressure leg as the condensate will have its own, unknown, head pressure.

In these circumstances the problems can be overcome by deliberately filling both legs with fluid as *Figure 12.35(c)*. Note that the high pressure leg is now connected to the tank top. The differential pressure is now:

$$\Delta P = (D - h) \cdot g \cdot (\rho_1 - \rho_2) \Leftarrow$$

where ρ_1 is the liquid density and ρ_2 the vapour density (often negligible). Note that the equation is independent of the transducer offset H_1. This arrangement is often used in boiler applications to measure the level of water in the drum. The filling of the pipes often occurs naturally if they are left unlagged.

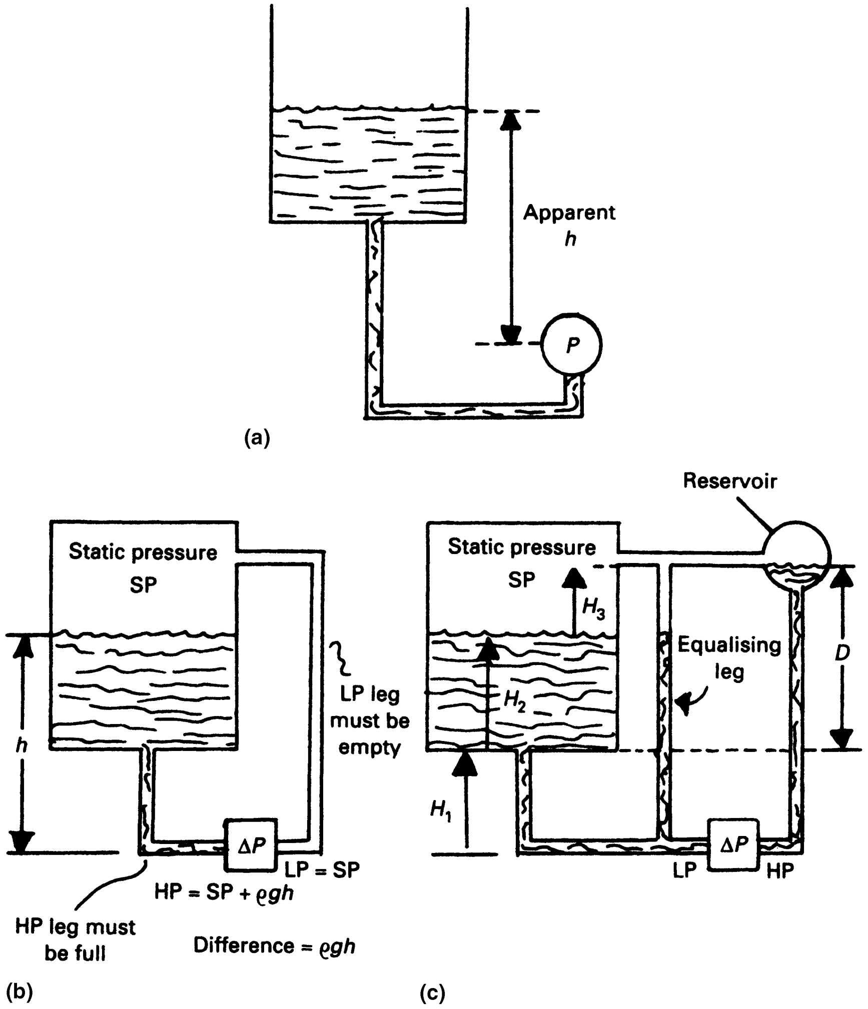

Apparent
h
P
(a)
Static pressure
SP
h
LP leg
must be
empty
ΔP
LP = SP
HP = SP + ϱgh
HP leg must
be full
Difference = ϱgh
(b)
Reservoir
Static pressure
SP
H_3
H_2
H_1
Equalising
leg
D
ΔP
LP
HP
(c)

Figure 12.35 Level measurement from hydrostatic pressure. (a) Head error arising from transducer position. This error can be removed by a simple offset; (b) Differential pressure measurement in a pressurised tank; (c) Level measurement with condensable liquid. This method is commonly used for drum level in steam boilers

One problem of all methods described so far is that the pressure transducer must come into contact with the fluid. If the fluid is corrosive or at a high temperature this may cause early failures. The food industries must also avoid stagnant liquids in the measuring legs. One solution is the gas bubbler of *Figure 12.36*. Here an inert gas (usually argon or nitrogen), is bubbled into the liquid, and the measured gas pressure will be the head pressure ρgh at the pipe exit.

12.5.3 Electrical probes

Capacitive probes can be used to measure the depth of liquids and solids. A rod, usually coated with PVC or PTFE, is inserted into the tank (*Figure 12.37*(*a*)) and the capacitance measured to the tank wall. This capacitance has two components, C_1 above the surface and C_2 below the surface. As the level rises C_1 will decrease and C_2 will increase. These two capacitors are in parallel, but as liquids and solids have a higher di-electric constant than vapour, the net result is that the capacitance rises for increasing level.

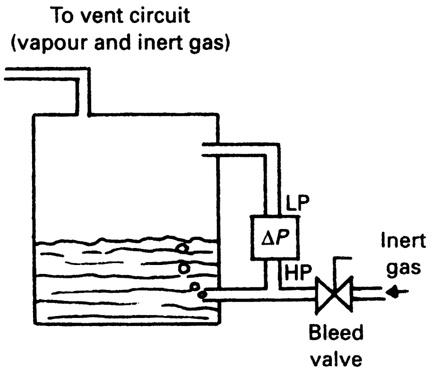

To vent circuit
(vapour and inert gas)
LP
ΔP
HP
Inert
gas
Bleed
valve

Figure 12.36 Level measurement using a gas bubbler

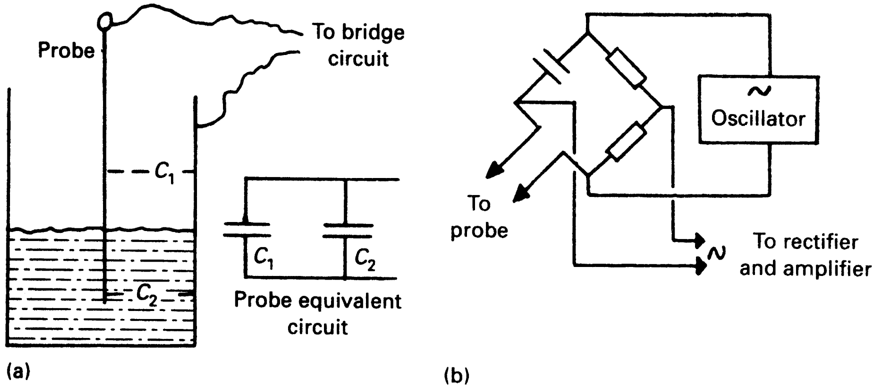


Figure 12.37 Level measurement using a capacitive probe. (a) Physical arrangement; (b) A.c. bridge circuit

The effect is small, however, and the change is best measured with an a.c. bridge circuit driven at about 100 kHz (*Figure 12.37(b)*). The small capacitance also means that the electronics must be close to the tank to prevent errors from cable capacitance.

The response is directly dependent on the dielectric constant of the material. If this changes, caused, say, by bubbling or frothing, errors will be introduced.

If the resistance of a liquid is reasonably constant, the level can be inferred by reading the resistance between two submerged metal probes. Stainless steel is often used to reduce corrosion. A bridge measuring circuit is again used with an a.c. supply to avoid electrolysis and plating effects. Corrosion can be a problem with many liquids, but the technique works well with water. Cheap probes using this principle are available for stop/start level control applications.

12.5.4 Ultrasonic transducers

Ultrasonic methods use high frequency sound produced by the application of a suitable a.c. voltage to a piezo electric crystal. Frequencies in the range 50 kHz to 1 MHz can be used, although the lower end of the range is more common in industry. The principle is shown on *Figure 12.38*. An ultrasonic pulse is emitted by a transmitter. It reflects off the surface and is detected by a receiver. The time of flight is given by:

$$t = \frac{2d}{v}$$

where v is the velocity of sound in the medium above the surface. The velocity of sound in air is about 3000 ms^{-1}, so for a tank whose depth can vary from 1 to 10 m, the delay will vary from about 7 ms (full) to 70 ms (empty).

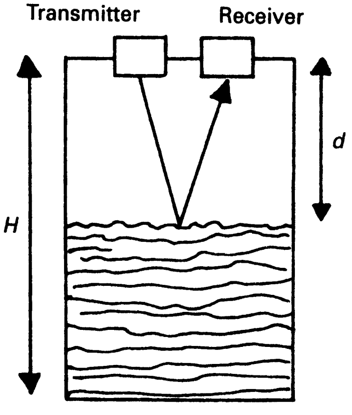


Figure 12.38 Basic arrangement for ultrasonic level measurement

There are two methods used to measure the delay. The simplest, assumed so far and most commonly used in industry, is a narrow pulse. The receiver will see several pulses, one almost immediately through the air, the required surface reflection and spurious reflections from sides, the bottom and rogue objects above the surface (e.g. steps and platforms). The measuring electronics normally provides adjustable filters and upper and lower limits to reject unwanted readings.

Pulse driven systems lose accuracy when the time of flight is small. For a distance below a few millimetres a swept frequency is used where a peak in the response will be observed when the path difference is a multiple of the wavelength, i.e.

$$d = \frac{v}{2f}$$

where v is the velocity of propagation and f the frequency at which the peak occurs. Note that this is ambiguous as peaks will also be observed at integer multiples of the wavelength.

Both methods require accurate knowledge of the velocity of propagation. The velocity of sound is 1440 ms^{-1} in water, 3000 ms^{-1} in air and 5000 ms^{-1} in steel. It is also temperature dependent varying in air by 1% for a 30°C temperature change. Pressure also has an effect. If these changes are likely to be significant they can be measured and correction factors applied.

12.5.5 Nucleonic methods

Radioactive isotopes (such as cobalt 60) spontaneously emit gamma or beta radiation. As this radiation passes through the material it is attenuated according to the relationship:

$$I = I_0 \exp(-\mathrm{m}\rho d)$$

where I_0 is the initial intensity, m is a constant for the material, ρ is the density and d the thickness of the material. This equation allows a level measurement system to be constructed in one of the ways shown on *Figure 12.39*. In each case the intensity of the received radiation is dependent on the level, being a maximum when the level is low.

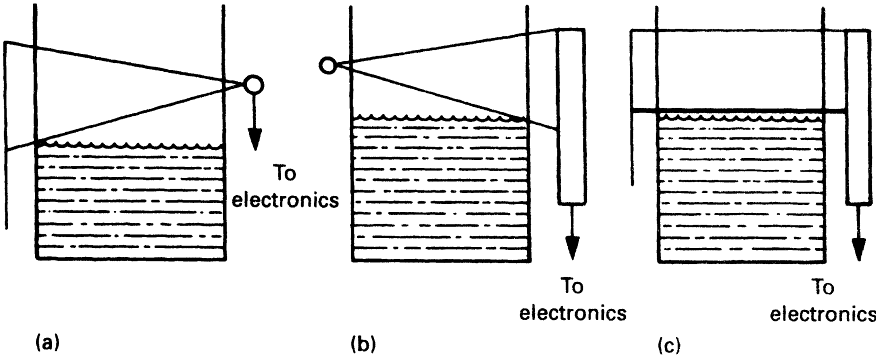


Figure 12.39 Various methods of level measurement using a radio-active source. (a) Line source, point detector; (b) Point source, line detector; (c) Collimated line source, line detector

The methods are particularly attractive for aggressive materials or extreme temperatures and pressures as all the measuring equipment can be placed outside the vessel.

Sources used in industry are invariably sealed, i.e. contained in a lead or similar container with a shuttered window allowing a narrow beam (typically 40° × 2°) to be emitted. The window can usually be closed when the system is not in use and to allow safe transportation. Source strength is determined by the number of disintegrations per second (dps) at the source. There are two units in common use; the curie, (Ci 3.7×10^{10} dps) and the SI derived unit the bequerel (Bq 1 dps). The typical industrial source would be 500 mCi (18.5 GBq) caesium 137 although there are variations by a factor of at least ten dependent on the application.

The strength of a source decays exponentially with time and this decay is described by the *half life* which is the time taken for the strength to decrease by 50%. Cobalt 60, a typical industrial source, for example has a half life of 3.5 years and will hence change by about 1% per month. Sources thus have a built in drift and a system will require regular calibration.

The biological effects of radiation are complex and there is no real safe level; all exposure must be considered harmful and legislation is based around the concept of ALARP; '*as low as reasonably practical*'. Exposures should be kept as low as possible, and figures quoted later should not be considered as design targets.

The *adsorbed dose* (AD) is a measure of the energy density adsorbed. Two units are again used; the rad and the SI derived unit the gray. These are more commonly encountered as the millirad (mrad) and the milligray (mGy):

$$1\,\mathrm{mrad} = 6.25 \times 10^{4}\,\mathrm{MeVg}^{-1}$$

$$1\,\mathrm{mGy} = 6.25 \times 10^{6}\,\mathrm{MeVg}^{-1}$$

The adsorbed dose is not however directly related to biological damage as it ignores the differing effect of α, β, and γ, radiation. A quality factor Q is defined which allows a *dose equivalent* (DE) to be calculated from:

$$\mathrm{DE} = Q \times \mathrm{AD}$$

There are, yet again, two units in common use; the rem and SI derived unit the sievert, again these are more commonly encountered as the millirem and millisievert (mSv);

$$\mathrm{mrem} = Q \times \mathrm{mrad}$$

$$\mathrm{mSv} = Q \times \mathrm{mGy}$$

The so called *dose rate* is expressed as DE/time (e.g. mrem/hr). Dose rates fall off as the inverse square of the distance, i.e. doubling the distance from the source gives one quarter of the dose rate.

British legislation recognises three groups of people:

Classified workers are trained individuals wearing radiation monitoring devices (usually film badges). They must not be exposed to dose rates of more than 2.5 mrem/hr and their maximum annual exposure (DE) is 5 rem. Medical records and dose history must be kept.

Supervised workers operate under the supervision of classified workers. Their maximum dose rate is 0.75 mrem/hr and maximum annual exposure is 1.5 rem.

Unclassified workers refers to others and the general public. Here the dose rate must not exceed 0.25 mrem/hr and the annual exposure must be below 0.5 rem.

It must be emphasised that the principle of ALARP applies and these are not design criteria.

Nucleonic level detection also requires detectors. Two types are commonly used, the *Geiger-Muller* (GM) tube and the *scintillation counter*. Both produce a semi-random pulse stream, the number of pulses received in a given time being dependent on the strength of the radiation. This pulse chain must be converted to a d.c. voltage by suitable filtering to give a signal which is dependent on the level of material between the source and the detector.

Conceptually nucleonic level measurement is simple and reliable. Public mistrust and the complex legislation can, however, make it a minefield for the unwary. Professional advice should always be taken before implementing a system.

12.5.6 Level switches

Many applications do not require analogue measurement but just a simple material present/absent signal. For liquids the simplest is a float which hangs down when out of a liquid and inverts when floating. The float orientation is sensed by two internal probes and a small mercury pool. Solids can be detected by rotating paddles or vibrating reeds which seize solid when submerged.

Two non-moving detectors are shown on *Figure 12.40*. The first is a heated tube whose heat loss will be higher (and hence the sensed temperature lower) when submerged. The second is a light reflective probe which will experience total internal reflection when submerged.

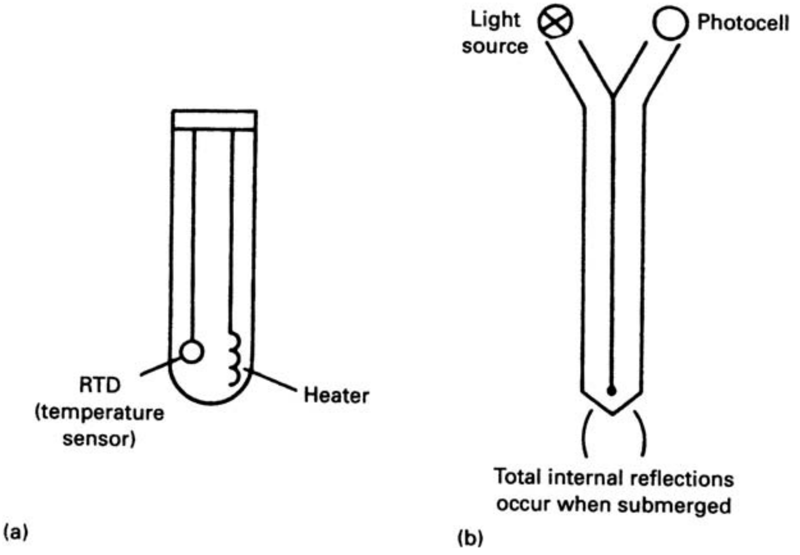


Figure 12.40 Non contact level switches. (a) Using heat loss from submerged sensor; (b) Optical level switch based on total internal reflection

12.6 Position transducers

12.6.1 Introduction

The measurement of position or distance is of fundamental importance in many applications. It also appears as an intermediate variable in other measurement transducers where the measured variable (such as pressure or force) is converted to a displacement.

12.6.2 The potentiometer

The simplest position transducer is the humble potentiometer. They can directly measure angular or linear displacements. *Figure 12.41* shows a potentiometer with track resistance R, connected to a supply V_i and a load R_L. If the span is D and the slider at position d, we can define a fractional displacement $x = d/D$. The output voltage is then:

$$V_0 = \frac{xV_i}{1 + x(1 - x)(R/R_L)}$$

The error is zero at both ends of the travel and maximum at $x = 2/3$. If R_L is significantly larger than R, the error is approximately $25 \cdot R/R_L\%$ of value or $15 \cdot R/R_L\%$ of full scale. A 10 K potentiometer connected to a 100 K load will thus have an error of about 1.5% FSD. If $R_L \gg R$, then $V_0 = x \cdot V_i$.

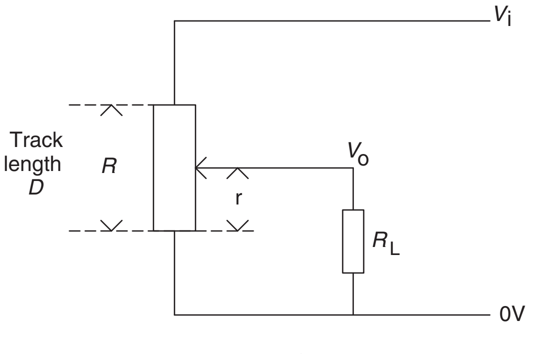


Figure 12.41 Potentiometer used for position measurement

Loading errors are further augmented by linearity and resolution errors. Potentiometers have finite resolution from the wire size or grain size of the track. They are also mechanical devices which require a force to move and hence can suffer from stiction, backlash and hysteresis.

The failure mode of a potentiometer also needs consideration. A track break can cause the output signal to be fully high above the failure point and fully low below it. In closed loop position control systems this manifests itself as a high speed dither around the break point.

Potentiometers can be manufactured with track resistance to give a specified non-linear response (e.g. trigonometric, logarithmic, square root etc.).

12.6.3 Synchros and resolvers

Figure 12.42(a) shows a transformer whose secondary can be rotated with respect to the primary. At angle θ the output voltage will be given by

$$V_0 = K \cdot V_i \cdot \cos\theta$$

where K is a constant. This relationship is shown on *Figure 12.42(b)*. The output amplitude is dependent on the angle, and the signal can be in phase (from $\theta = 270°$ to 90°) or anti-phase (from $\theta = 90°$ to 270°). This principle is the basis of synchros and resolvers.

A synchro link in its simplest form consists of a transmitter and receiver connected as *Figure 12.43*. Although this looks superficially like a three-phase circuit it is fed from a single-phase supply, often at 400 Hz. The voltage applied to the transmitter induces in-phase/anti-phase voltages in the windings as described above and causes currents to flow through the stator windings at the receiver. These currents produce a magnetic field at the receiver which aligns with the angle of the transmitter rotor.

The receiver rotor also produces a magnetic field. If the two receiver fields do not align, torque will be produced on

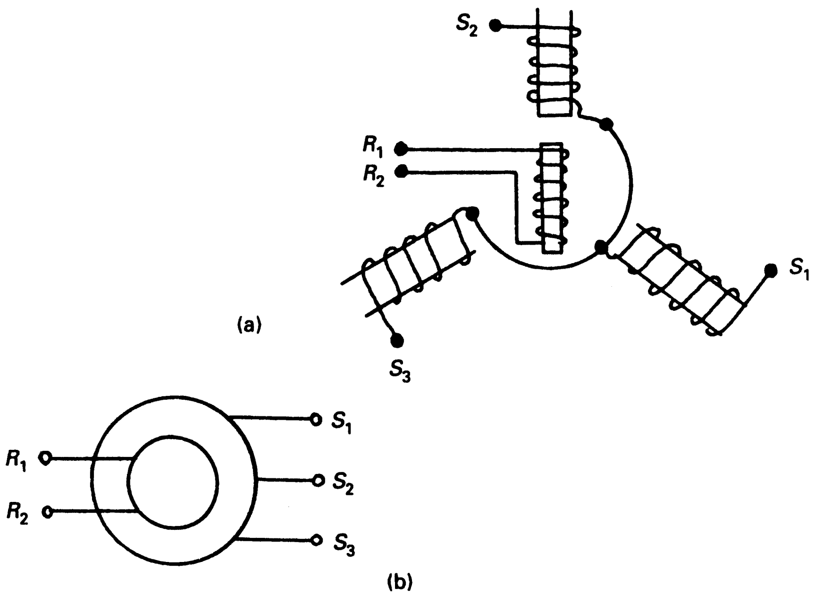


Figure 12.42 The synchro torque transmitter. (a) Construction; (b) Schematic representation

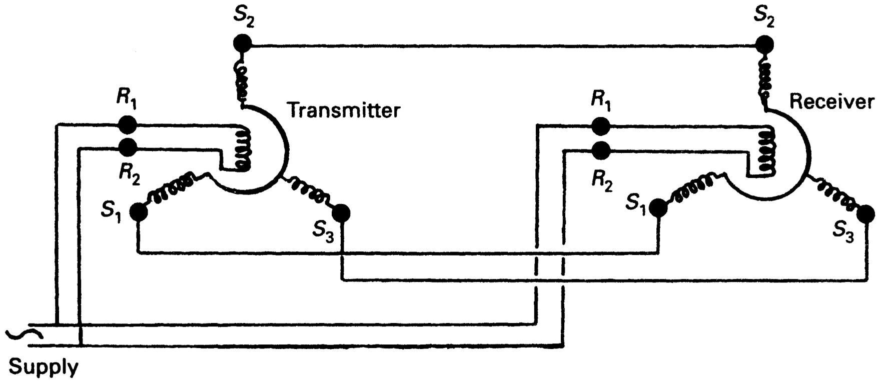


Figure 12.43 A synchro transmitter receiver link

the receiver shaft which causes the rotor to rotate until the rotor and stator fields align, i.e. the receiver rotor is at the same position as the transmitter rotor.

Such a link is called a *torque link* and can be used to remotely position an indicator. A more useful circuit, shown on *Figure 12.44* can be used to give an electrical signal which represents the difference in angle of the two shafts. The receiving device is called a *control transformer*.

The transmitter operates as described above and induces a magnetic field in the control transformer. An a.c. voltage will be induced in the control transformer's rotor if it is not perpendicular to the field. The magnitude of this voltage is related to the angle error and the phase to the direction.

The a.c. output signal must be converted to d.c. by a phase sensitive rectifier. The *Cowan rectifier* (*Figure 12.45*) is commonly used. Another common circuit is the positive/negative amplifier (*Figure 12.46*) with the electronic polarity switch driven by the reference supply.

Resolvers have two stator coils at right angles and a rotor coil as shown on *Figure 12.47*. The voltages induced in the two stator coils are simply

$$V_1 = K \cdot V_i \cos\theta$$
$$V_2 = K \cdot V_i \sin\theta$$

Resolvers are used for co-ordinate conversion and conversion from rectangular to polar co-ordinates. They are also widely used with solid state digital converters which give a binary output signal (typically 12 bit). One useful feature is that $(V_1^2 + V_2^2)$ is a constant which makes an open circuit winding easy to detect.

Resolvers can also be used with a.c. applied to the stator windings and the output signal taken from the rotor (*Figure 12.48*). One stator signal is shifted by 90°. Both signals are usually obtained from a pre-manufactured quadrature oscillator. The output signal is then

$$V_0 = K \cdot V_i(\sin\omega t \cdot \cos\theta + \cos\omega t \cdot \sin\theta)$$

which simplifies to

$$V_0 = K \cdot V_i(\sin\omega t + \theta)$$

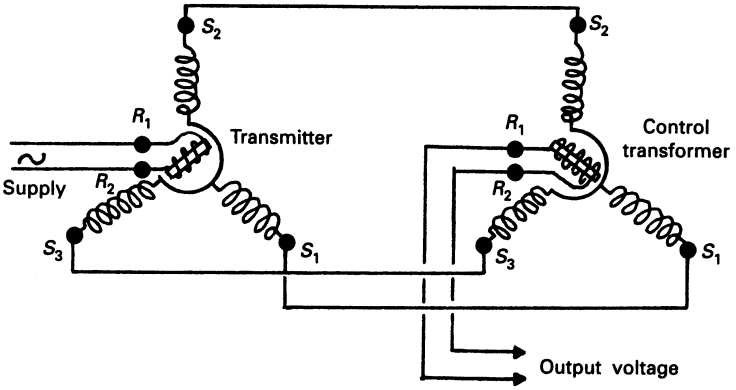

S_2
S_2
R_1
Transmitter
R_1
Control transformer
Supply
R_2
R_2
S_3
S_1
S_3
S_1
Output voltage

Figure 12.44 Connection of a position control system using a control transformer. The output voltage is zero when the rotor on the control transformer is at 90°⇐to the rotor on the transmitter. Note there are two zero positions 180°⇐apart

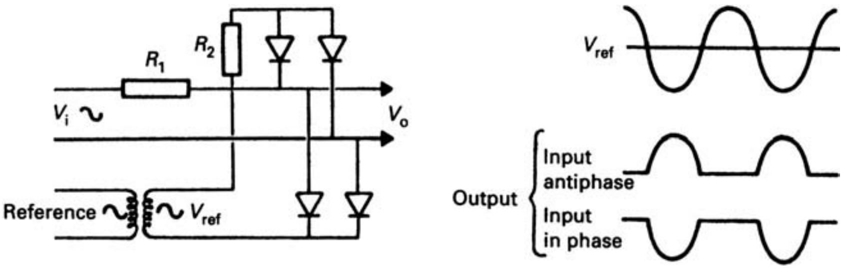

R_1
R_2
V_i ∼
V_o
Reference ∼
∼ V_{ref}
V_{ref}
Output
Input antiphase
Input in phase

Figure 12.45 The Cowan half wave phase sensitive rectifier

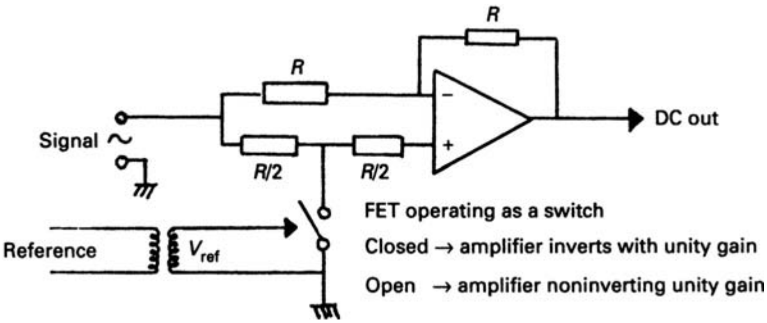

R
R
Signal ∼
R/2
R/2
−
+
DC out
Reference
V_{ref}
FET operating as a switch
Closed → amplifier inverts with unity gain
Open → amplifier noninverting unity gain

Figure 12.46 Full wave phase sensitive rectifier based on an operational amplifier. The switch represents a FET driven by a reference signal

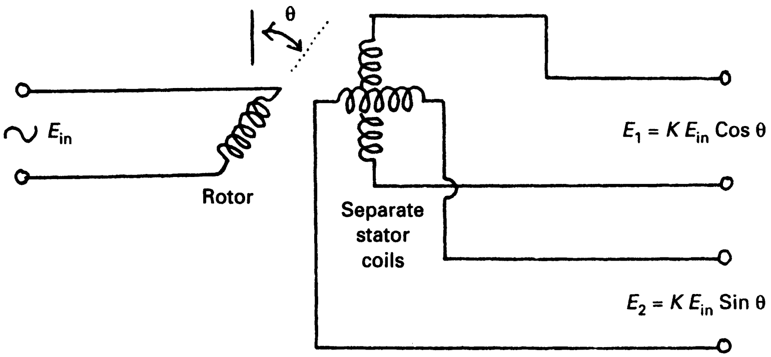

θ
∼ E_{in}
Rotor
Separate stator coils
$E_1 = K E_{in} \cos\theta$
$E_2 = K E_{in} \sin\theta$

Figure 12.47 The resolver; the reference voltage is applied to the rotor and the output signals read from the two stator windings

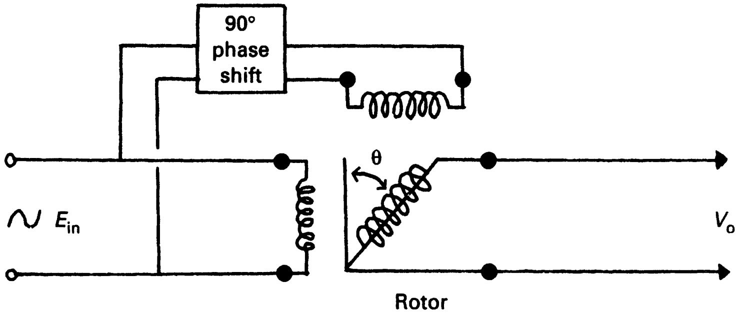


Figure 12.48 Alternative connection for a resolver. Two reference voltages shifted by 90° are applied to the stator windings and the output read from the rotor. The output signal is a constant amplitude sine wave with phase shift determined by the rotor angle

This a.c. signal can be converted to d.c. by a phase sensitive rectifier.

Complete rotary transducers comprising quadrature oscillator, resolver and phase sensitive rectifier can be obtained which give a d.c. output signal proportional to shaft angle. Because there is little friction these are low torque devices.

12.6.4 Linear variable differential transformer (LVDT)

An LVDT is used to measure small displacements, typically less than a few mm. It consists of a transformer with two secondary windings and a movable core as shown on *Figure 12.49*(*a*). At the centre position the voltages induced in the secondary windings are equal but of opposite phase giving zero output signal. As the core moves away from the centre one induced voltage will be larger, giving a signal whose amplitude is proportional to the displacement and whose phase shows the direction as *Figure 12.49*(*b*). Again a d.c. output signal can be obtained with a phase sensitive rectifier. *Figure 12.49*(*c*) shows the same idea used to produce a small displacement angular transducer.

12.6.5 Shaft encoders

Shaft encoders give a digital representation of an angular position. They exist in two forms.

An *absolute* encoder gives a parallel output signal, typically 12 bits (one part in 4095) per revolution. Binary coded decimal (BCD) outputs are also available. A simplified four bit encoder would therefore operate as in *Figure 12.50*(*a*). Most absolute encoders use a coded wheel similar to *Figure 12.50*(*b*) moving in front of a set of photocells.

A simple binary coded shaft encoder can give anomalous readings as the outputs change state. Suppose a four bit encoder is going from 0111 to 1000. It is unlikely that all photocells will change together, so the output states could go 0111 > 0000 > 1000 or 0111 > 1111 > 1000 or any other sequence of bits.

There are two solutions to this problem. The first uses an additional track, called an *anti-ambiguity track*, which is used by the encoder's internal logic to inhibit changes around transition points.

The second solution uses a *unit distance code*, such as the *Gray code* described in Section 14.6.4 which has no ambiguity. The conversion logic from Gray to binary is usually contained within the encoder itself.

Incremental encoders give a pulse output with each pulse representing a fixed distance. These pulses are counted by an external counter to give indication of position. Simple encoders can be made by reading the output from a proximity detector in front of a toothed wheel.

A single pulse output train carries no information as to direction. Commercial incremental encoders usually provide two outputs shifted 90° in phase as *Figure 12.51*(*a*). For clockwise rotation, A will lead B as *Figure 12.51*(*b*) and the positive edge of A will occur when B is low. For anti-clockwise rotation B will lead A as *Figure 12.51*(*c*) and the positive edge of A will occur when B is high. The logic (A and not B) can be used to count up, and (A and B) to count down. The two output incremental encoder can thus be used to follow reversals without cumulative error, but can still lose its datum position after a power failure.

Most PLC systems have dedicated high speed counter cards which can read directly from a two channel incremental encoder.

12.6.6 Variable capacitance transducers

The small deflections obtained in weigh systems or pressure transducers are often converted into an electrical signal by varying the capacitance between two plates. Very small displacements can be detected by these methods.

The capacitance C of a parallel plate capacitor is given by:

$$C = \frac{\varepsilon A}{d}$$

where ε is the permitivity of the material between the plates, A the area and d the separation.

Variation in capacitance can be obtained from varying e (sliding in a dielectric, or moving the plates horizontally (change A) or apart (change d)). Although the change is linear for e and A, the effect is small. A common arrangement uses a movable plate between two fixed plates as *Figure 12.52*(*a*). Here two capacitors are formed which can be directly connected into an a.c. bridge as *Figure 12.52*(*b*). If the fixed plates are distance D apart, d is the null position separation and the centre plate is displaced by x, the output voltage change can be shown to be:

$$V_0 = V_i x/2d$$

12.6.7 Laser distance measurement

Lasers can give very accurate measurement of distance and are often used in surveying. There are two broad classes of laser distance measurement.

In the first, shown on *Figure 12.53* and called a *triangulation laser*, a laser beam is used to produce a very bright

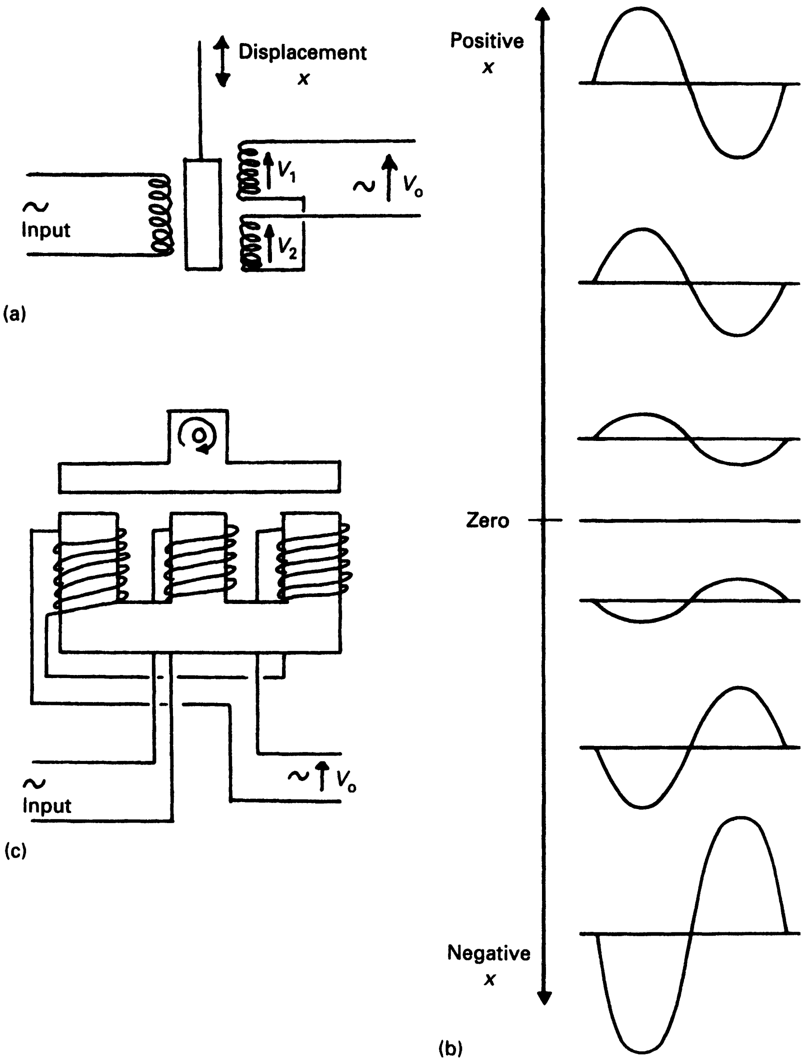


Figure 12.49 The linear variable displacement transformer (LVDT). (a) Construction; (b) Output signal for various positions; (c) Small displacement angular device

spot on the target object. This is viewed by an imaging device which can be considered as a linear array of tiny photocells. The position of the spot on the image will vary according to the distance as shown, and the distance found by simple triangulation trigonometry.

The second type of distance measurement is used at longer distances and times how long it takes a laser pulse to reach the target and be reflected back. For very long distances, (in surveying for example), a pulse is simply timed. At shorter distances a continuous amplitude modulated laser beam is sent and the phase shift between transmitted and received signals used to calculate the distance. This latter method can give ambiguous results, so often the two techniques are combined to give a coarse measurement from time of flight which is fine tuned by the phase shift.

Lasers can blind, and are classified according to their strength. Class 1 lasers are inherently safe. Class 2 lasers can cause retinal damage with continuous viewing but the natural reaction to blink or look away gives some protection. Class 3 lasers can cause damage before the natural reaction occurs. Risk assessments, safe working practices and suitable guarding must be provided for Class 2 and 3 lasers. Optical devices such as binoculars must never be used whilst working around lasers.

12.6.8 Proximity switches and photocells

Mechanical limit switches are often used to indicate the positional state of mechanical systems. These devices are bulky, expensive and prone to failure where the environment is hostile (e.g. dust, moisture, vibration, heat).

Proximity switches can be considered to be solid state limit switches. The commonest type is constructed around a coil whose inductance changes in the presence of a metal surface. Simple two wire a.c. devices act just like a switch, having low impedance and capable of passing several hundred mA when covered by metal, and a high impedance

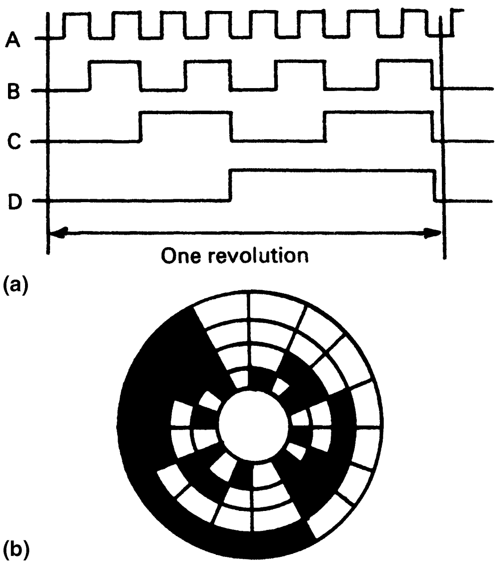


Figure 12.50 Absolute position shaft encoder. (a) Output from a four bit encoder. Real devices use 12 bits and employ unit distance coding such as the Gray code; (b) The wheel on a four bit encoder.

(typical leakage 1 mA) when uncovered. Sensing distances up to 20 mm are feasible, but 5–10 mm is more common.

The leakage in the off state (typically 1 mA) can cause problems with circuits such as high impedance PLC input cards. In these circumstances a dummy load must be provided in parallel with the input.

D.c. powered switches use three wires, two for the supply and one for the output. D.c. sensors use an internal high frequency oscillator to detect the inductance change which gives a much faster response. Operation at several kHz is easily achieved.

A d.c. proximity detector can have an output in either of the two forms of *Figure 12.54*. A PNP output switches a positive supply to a load with a return connected to the negative supply. A PNP output is sometimes called a *current sourcing* output. With an NPN output the load is connected to the positive supply and the proximity detector connects the load to the negative supply. NPN outputs are therefore sometimes called a *current sinking* output. It is obviously important to match the detector to the load connection.

Inductive detectors only work with metal targets. Capacitive proximity detectors work on the change in capacitance caused by the target and accordingly work with wood, plastic, paper and other common materials. Their one disadvantage is they need adjustable sensitivity to cope with different applications.

Ultrasonic sensors can also be used, operating on the same principle as the level sensor described in Section 12.5.4. Ultrasonic proximity detectors often have adjustable near and far limits so a sensing window can be set.

Photocells (or PECs) are another possible solution. The presence of an object can be detected either by the breaking of a light beam between a light source and a PEC or by bouncing a light beam off the object which is detected by the PEC. In the first approach the PEC sees no light for an object present. In the second approach (called a *retro-reflective* PEC) light is seen for an object present. It is usual to use a modulated light source to avoid erratic operation with changes in ambient light and prevent interaction between adjacent PECs.

12.7 Velocity and acceleration

12.7.1 Introduction

Velocity is the rate of change of position, and acceleration is the rate of change of velocity. Conversion between them can therefore be easily achieved by the use of integration and differentiation circuits based on d.c. amplifiers. Integration circuits can, though, be prone to long term drift.

12.7.2 Velocity

Many applications require the measurement of angular velocity. The commonest method is the tachogenerator which is a simple d.c. generator whose output voltage is proportional to speed. A common standard is 10 volts per 1000 rpm. Speeds up to 10 000 rpm can be measured, the upper limit being centrifugal force on the tacho commutator.

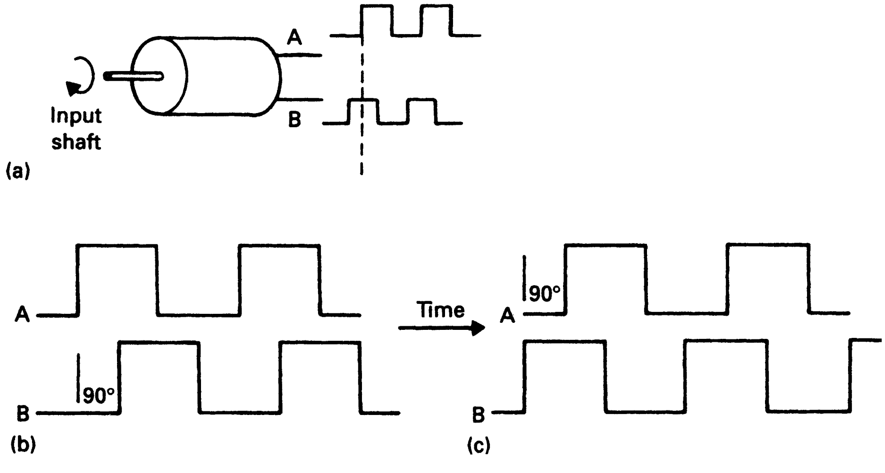


Figure 12.51 The incremental encoder. (a) Physical arrangements—the two phase shifted outputs give directional information; (b) Clockwise rotation; (c) Anti-clockwise rotation

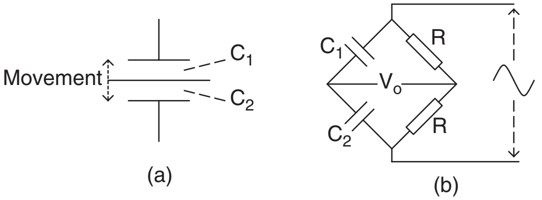


Figure 12.52 Capacitive based displacement transducer. (a) Physical arrangement; (b) A.c. bridge

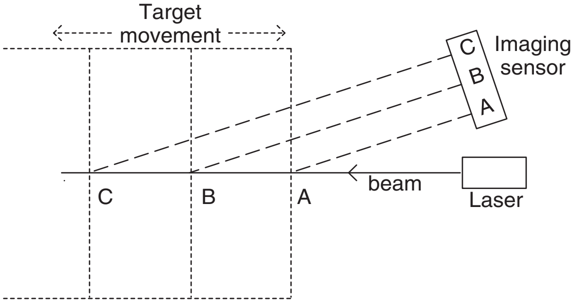


Figure 12.53 Triangulation laser position measurement

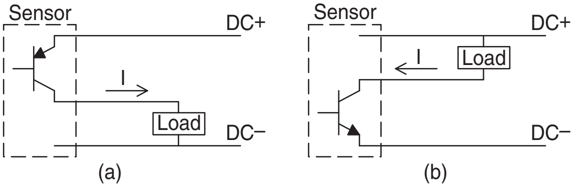


Figure 12.54 Current sourcing and sinking proximity detectors. (a) PNP current sourcing; (b) NPN current sinking

Pulse tachometers are also becoming common. These are essentially identical to incremental encoders described in Section 12.6.5 and produce a constant amplitude pulse train whose frequency is directly related to rotational speed. This pulse train can be converted into a voltage in three ways.

The first, basically analog, method used fires a fixed width monostable as *Figure 12.55* which gives a mark/space ratio which is speed dependent. The monostable output is then passed through a low pass filter to give an output proportional to speed. The maximum achievable speed is determined by the monostable pulse width.

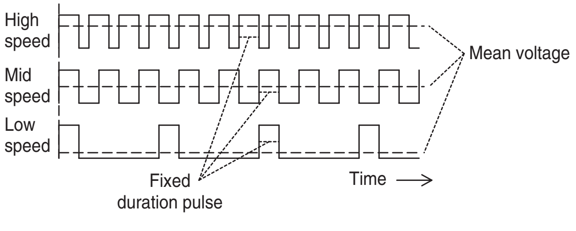


Figure 12.55 Speed measurement using incremental pulse encoder. The pulses from the encoder fire a fixed duration monostable. The resulting pulse train is filtered to give a mean voltage proportional to speed

The second, digital, method counts the number of pulses in a given time. This effectively averages the speed over the count period which is chosen to give a reasonable balance between resolution and speed of response.

Counting over a fixed time is not suitable for slow speeds as adequate resolution can only be obtained with a long duration sample time. At slow speeds, therefore, the period of the pulses is often directly timed giving an average speed per pulse. The time is obviously inversely proportional to speed.

Digital speed control systems often use both of the last two methods and switch between them according to the speed.

Linear velocity can be measured using *Doppler shift*. This occurs when there is relative motion between a source of sound (or electromagnetic radiation) and an object. It is commonly experienced as a change of pitch of a car horn as the vehicle passes.

Suppose an observer is moving with velocity v towards a source emitting sound with a frequency f. The observer will see each wavefront arrive early and the perceived frequency will be

$$f_{\rm v} = (c + v)/\lambda$$

where c is the velocity of propagation. As the original frequency is c/l, the frequency shift is

$$\delta f = (f_{\rm v} - f) = v/\lambda \approx fv/c \qquad (12.5)$$

i.e. proportional to speed.

The Doppler shift thus allows remote measurement of velocity. The principle is shown on *Figure 12.56*. A transmitter, (ultrasonic, radar or light,) emits a constant frequency signal. This is bounced off the target. Two Doppler shifts occur as the object is acting as both moving observer and moving (reflected) source. The frequency shift is thus twice that for the simple case of Equation 12.5.

The transmitted and received frequencies are mixed to give a beat frequency equal to the frequency shift and hence proportional to the target object's speed. The beat frequency can be measured by counting the number of cycles in a given time.

12.7.3 Accelerometers and vibration transducers

Both accelerometers and vibration transducers consist of a seismic mass linked to a spring as *Figure 12.57*(*a*). The movement will be damped by the dashpot giving a second order response defined by the natural frequency and the damping factor, typically chosen to be 0.7.

If the frame of the transducer is moved sinsusoidally, the relationship between the mass displacement and the frequency will be as *Figure 12.57*(*b*).

In region A, the device acts as an accelerometer and the mass displacement is proportional to the acceleration. In region C, the mass does not move, and the displacement with respect to the frame is solely dependent on the frame movement (but is shifted by 180°). In this region the device acts as a vibration transducer. Typically region A extends to about one third of the natural frequency and region C starts at about three times the natural frequency.

In both cases the displacement is converted to an electrical signal by LVDTs or strain gauges.

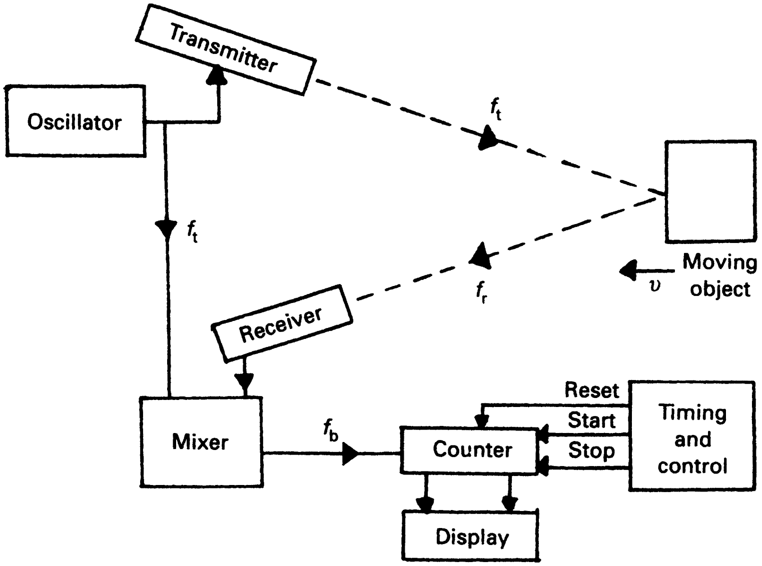

Transmitter
Oscillator
f_t
f_t
Moving object
v
f_r
Receiver
Mixer
f_b
Counter
Reset
Start
Stop
Timing and control
Display

Figure 12.56 Doppler shift velocity measurement

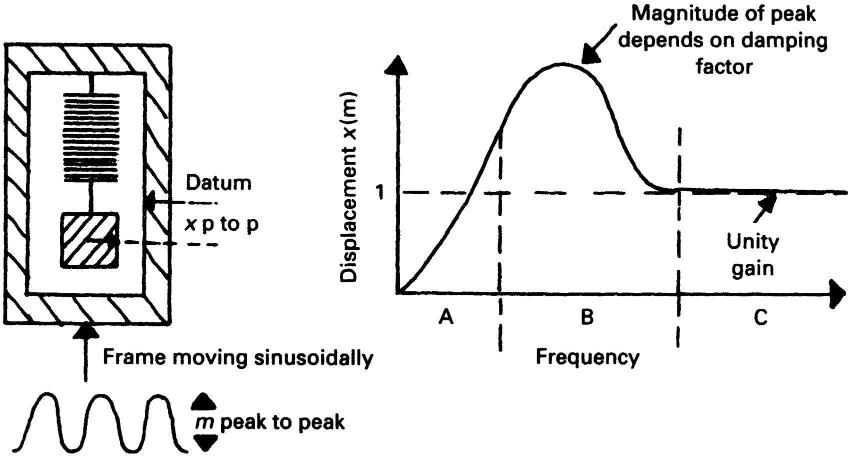

Datum
x p to p
Frame moving sinusoidally
m peak to peak
Displacement x(m)
1
Magnitude of peak depends on damping factor
Unity gain
A
B
C
Frequency

Figure 12.57 Principle of operation of accelerometers and vibration transducers. (a) Schematic diagram; (b) Frequency response. In region A the device acts as an accelerometer. In region C the device acts as a vibration transducer

12.8 Strain gauges, loadcells and weighing

12.8.1 Introduction

Accurate measurement of weight is required in many industrial processes. There are two basic techniques in use. In the first, shown in *Figure 12.58*(*a*) and called a *force balance system*, the weight of the object is opposed by some known force which will equal the weight. Kitchen scales where weights are added to one pan to balance the object in the other pan use the force balance system. Industrial systems use hydraulic or pneumatic pressure to balance the load, the pressure required being directly proportional to the weight.

The second, and commoner, method is called *strain weighing*, and uses the gravitational force from the load to cause a change in the structure which can be measured. The simplest form is the spring balance of *Figure 12.58*(*b*) where the deflection is proportional to the load.

12.8.2 Stress and strain

The application of a force to an object will result in deformation of the object. In *Figure 12.59* a tensile force F has been applied to a rod of cross-sectional area A and length L. This results in an increase in length ΔL. The effect of the force will depend on the force and the area over which it is applied. This is called the *stress*, and is the force per unit area:

Stress $= F/A$

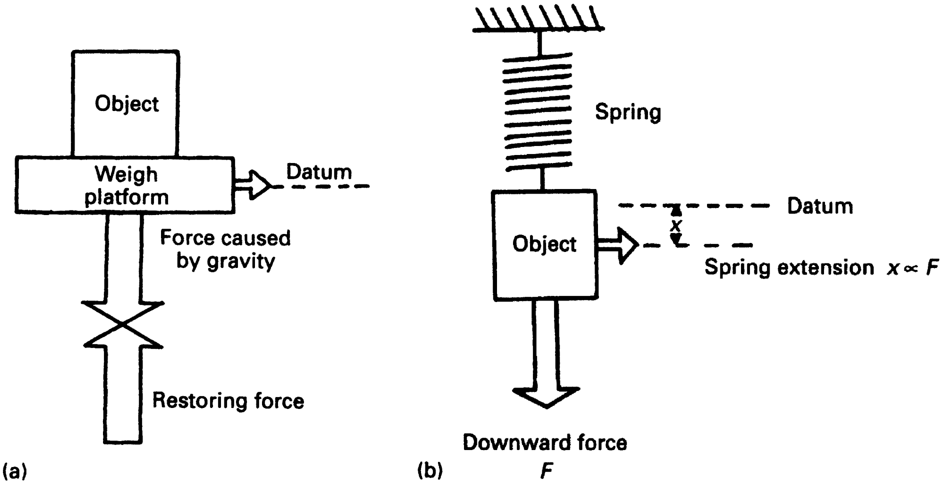


Figure 12.58 The two different weighing principles. (a) Force balance; (b) Strain weighing

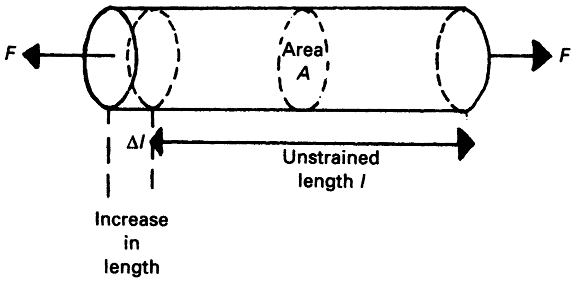


Figure 12.59 Tensile strain

Stress has the units of N/m^2 (i.e. the same as pressure, so pascals are sometimes used).

The resulting deformation is called the *strain*, and is defined as the fractional change in length.

Strain = $\Delta L/L$

Strain is dimensionless. Because the change in length is small, microstrain (µstrain defined as strain*10^6) is often used. A 10 m rod exhibiting a change in length of 0.0045 mm because of an applied force is exhibiting 45 µstrain.

The strain will increase as the stress increases as shown on *Figure 12.60*. Over region AB the object behaves as a spring; the relationship is linear and there is no hysteresis (i.e. the object returns to its original dimension when the force is removed. Beyond Point B the object suffers deformation and the change is not reversible. The relationship is now non linear, and with increasing stress the object fractures at point C. The region AB, called the *elastic region*, is used for strain measurement. Point B is called the *elastic limit*. Typically AB will cover a range of 10 000 µstrain.

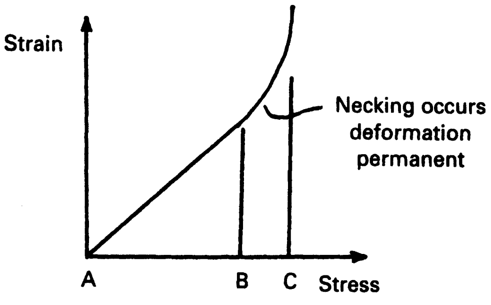


Figure 12.60 The relationship between stress and strain

The inverse slope of the line AB is sometimes called the *elastic modulus* (or *modulus of elasticity*). It is more commonly known as *Young's modulus* defined as

Young's modulus = Stress/Strain = $(F/A)/(\Delta L/L)\Leftarrow$

Young's modulus has dimensions of Nm^{-2}, i.e. the same as pressure. It is commonly given in pascals. Typical values are:

Steel	210 GPa
Copper	120 GPa
Aluminium	70 GPa
Plastics	30 GPa

When an object experiences strain, it displays not only a change in length but also a change in cross-sectional area. This is defined by *Poisson's ratio*, denoted by the Greek letter ν. If an object has a length L and width W in its unstrained state and experiences changes ΔL and ΔW when strained, Poisson's ratio is defined as:

νς= $(\Delta W/W)/(\Delta L/L)\Leftarrow$

Typically νςis between 0.2 and 0.4. Poisson's ratio can be used to calculate the change in cross-sectional area.

12.8.3 Strain gauges

The electrical resistance of a conductor is proportional to the length and inversely proportional to the cross-section. i.e.

$R = \rho L/A$

where ρ is a constant called the *resistivity* of the material.

When a conductor suffers stress, its length and area will both change resulting in a change in resistance. For tensile stress, the length will increase and the cross-sectional area

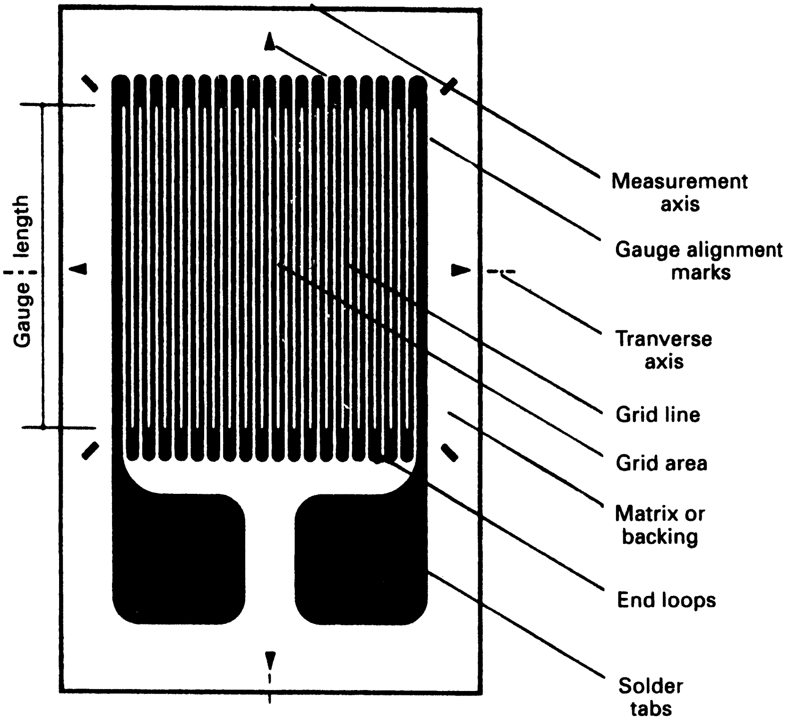


Figure 12.61 A typical strain gauge (courtesy of Welwyn Strain Measurements)

decrease both resulting in increased resistance. Similarly the resistance will decrease for compressive stress.

Ignoring second order effects the change in resistance is given by:

$$\Delta R/R = G \cdot \Delta L/L \qquad (12.6)$$

where G is a constant called the *gauge factor*. $\Delta L/L$ is the strain, E, so Equation 12.6 can be re-written:

$$\Delta R = G \cdot R \cdot E \qquad (12.7)$$

where E is the strain experienced by the conductor.

Devices based on the change of resistance with load are called *strain gauges*, and Equation 12.7 is the fundamental strain gauge equation.

Practical strain gauges are not a slab of material as implied so far but consist of a thin small, (typically a few mm), foil similar to *Figure 12.61* with a pattern to increase the conductor length and hence the gauge factor. The gauge is attached to some sturdy stressed member with epoxy resin and experiences the same strain as the member. Early gauges were constructed from thin wire but modern gauges are photo-etched from metallised film deposited onto a polyester or plastic backing. Normal gauges can experience up to 10 000 μstrain without damage. Typically the design will aim for 2000 μstrain under maximum load.

A typical strain gauge will have a gauge factor of 2, a resistance of 120 Ω and experience 1000 μstrain. From Equation 12.7 this will result in a resistance change of 0.24 Ω.

Strain gauges must ignore strains in unwanted directions. A gauge has two axis, an active axis along which the strain is applied and a passive axis (usually at 90°) along which the gauge is least sensitive. The relationship between these is defined by the *cross-sensitivity*:

$$\text{cross-sensitivity} = \frac{\text{sensitivity along passive axis}}{\text{sensitivity along active axis}}$$

Cross-sensitivity is typically about 0.002.

12.8.4 Bridge circuits

The small change in resistance is superimposed on the large unstrained resistance. Typically the change will be 1 part in 5000. In *Figure 12.62* the strain gauge R_g is connected into a classical Wheatstone bridge. In the normal laboratory method, R_a and R_b are made equal and calibrated resistance box R_c adjusted until V_0, (measured by a sensitive millivoltmeter,) is zero. Resistance R_c then equals R_g.

With a strain gauge, however, we are not interested in the actual resistance, but the change caused by the applied load. Suppose R_b and R_c are made equal, and R_a is made equal to the unloaded resistance of the strain gauge. Voltages V_1 and V_2 will both be half the supply voltage and V_0 will be zero. If a load is applied to the strain gauge such that its resistance changes by fractional change x (i.e. $x = \Delta R/R$) it can be shown that

$$V_0 = V_s \cdot x/2(2 + x) \qquad (12.8)$$

In a normal circuit x will be very small compared with 2. For the earlier example x has the value $0.24/120 = 0.002$, so Equation 12.8 can be simplified to:

$$V_0 = V_s \cdot \Delta R/4 \cdot R$$

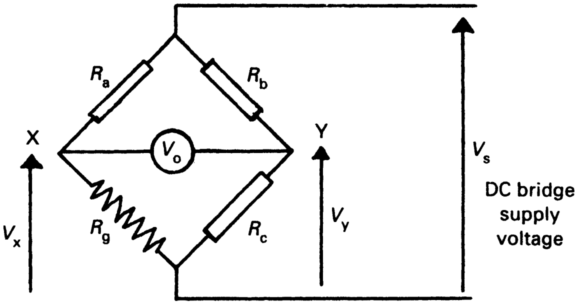


Figure 12.62 Simple measurement of strain with a Wheatstone bridge

But $\Delta R = E \cdot G \cdot R$ where E is the strain and G the gauge factor giving

$$V_0 = E \cdot G \cdot V_s/4 \qquad (12.9)\Leftarrow$$

For small values of x, the output voltage is thus linearly related to the strain.

It is instructive to put typical values into Equation 12.9. For a 24 V supply, 1000 µstrain and gauge factor 2 we get 12 mV. This output voltage must be amplified before it can be used. Care must be taken to avoid common mode noise so the differential amplifier circuit of *Figure 12.63* is commonly used.

The effect of temperature on resistance was described in Section 12.2.3. and resistance changes from temperature variation are of a similar magnitude to resistance changes from strain. The simplest way of preventing this is to use two gauges arranged as *Figure 12.64*(*a*). One gauge has its active axis and the other gauge the passive axis aligned with the load. If these are connected into a bridge circuit as shown on *Figure 12.64*(*b*) both gauges will exhibit the same resistance change from temperature and these will cancel leaving the output voltage purely dependent on the strain.

Temperature errors can also occur from dimensional changes in the member to which the strain gauges are attached. Gauges are often temperature compensated by having coefficients of linear expansion identical to the material to which they are attached.

Many applications use gauges in all four arms of the bridge as shown on *Figure 12.64*(*c*) with two active gauges and two passive gauges. This provides temperature compensation and doubles the output voltage giving

$$V_0 = E \cdot G \cdot V_s/2$$

In the arrangement of *Figure 12.64*(*d*) four gauges are used with two gauges experiencing compressive strain and two experiencing tensile strain. Here all four gauges are active, giving

$$V_0 = E \cdot G \cdot V_s \qquad (12.10)\Leftarrow$$

Again temperature compensation occurs because all gauges are at the same temperature.

Gauges are manufactured to a tolerance of about 0.5%, i.e. about +/−0.6 Ω for a typical 120 Ω gauge. As this is larger than the resistance change from strain some form of zeroing will be required. Three methods of achieving this are shown on *Figure 12.65*. Arrangement b and c are preferred if the zeroing is remote from the bridge.

Equation 12.10 shows that the output voltage is directly related to the supply voltage. This implies that a large voltage should be used to increase the sensitivity. Unfortunately a high voltage supply cannot be used or I^2R heating in the gauges will cause errors and ultimately failure. Bridge voltages of 15–30 V and currents of 20–100 mA are typically used.

Equation 12.10 also implies that the supply voltage must be stable. If the bridge is remote from the supply and electronics, as is usually the case, voltage drops down the cabling can introduce significant errors. *Figure 12.66* shows a typical cabling scheme used to give remote zeroing and compensation for cabling resistance. The supply is provided on cores 2 and 6 and monitored on cores 1 and 7 to keep constant voltage at the bridge itself.

12.8.5 Load cells

A loadcell converts a force (usually the gravitational force from an object being weighed) to a strain which can then be converted to an electrical signal by strain gauges. A load cell will typically have the local circuit of *Figure 12.67* with four gauges, two in compression and two in tension, and six external connections. The span and zero adjust on test (AOT) are factory set to ensure the bridge is within limits that can be further adjusted on site. The temperature compensation resistors compensate for changes in Young's

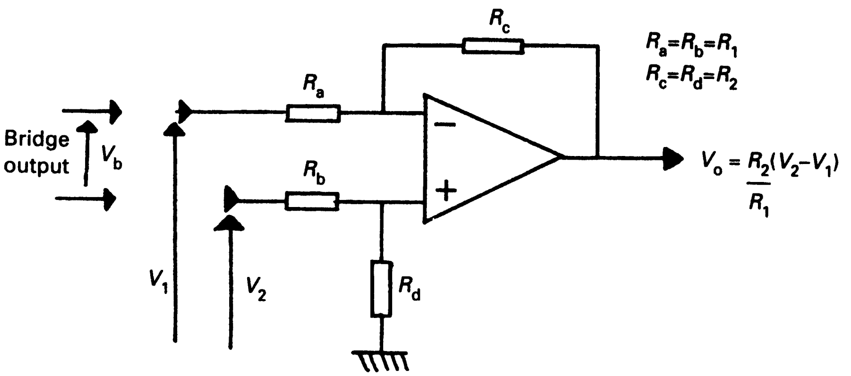


Figure 12.63 Differential amplifier. The output voltage is determined by the difference between the two input voltages

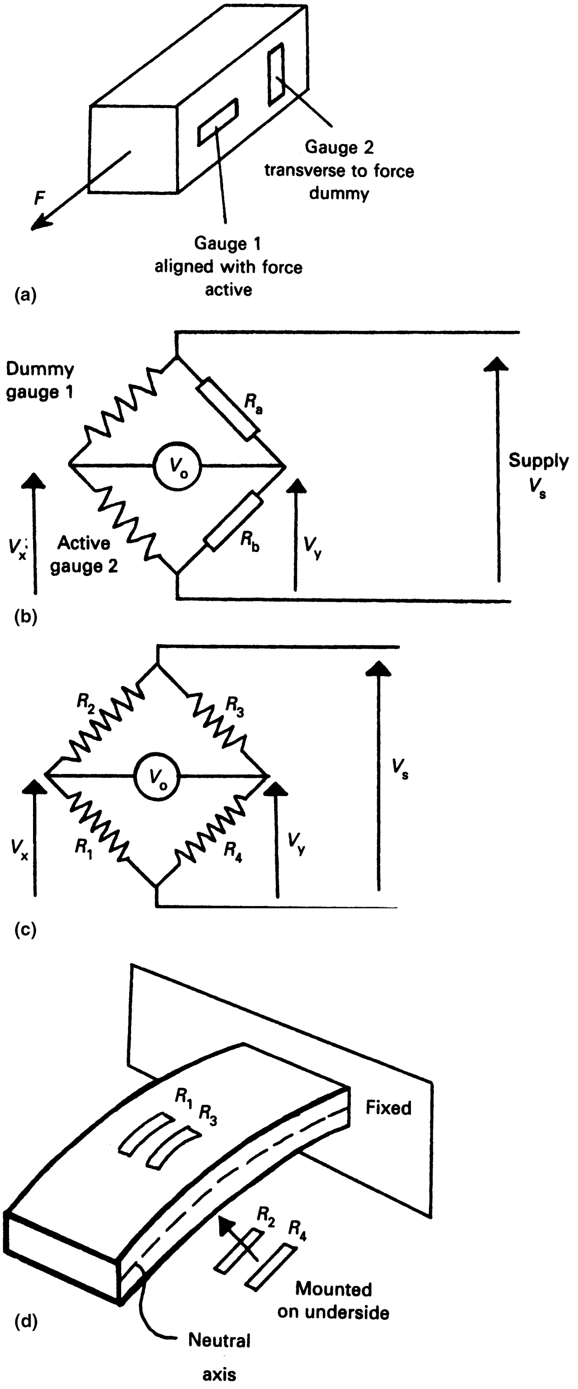


Figure 12.64 Use of multiple strain gauges. (a) Two gauges used to give temperature compensation; (b) Two gauges connected into bridge. Temperature effects will affect both gauges equally and have no effect; (c) Four gauge bridge gives increased sensitivity and temperature compensation; (d) Arrangement for four gauges. Two will be in compression and two in tension for load changes. Temperature affects all gauges equally

modulus with temperature, not changes of the gauges themselves which the bridge inherently ignores.

Coupling of the load requires care, a typical arrangement being shown on *Figure 12.68*. A pressure plate applies the load to the proof-ring via a knuckle and avoids error from slight misalignment. A flexible diaphragm seals against dust and weather. A small gap ensures that shock overloads will make the loadcell bottom out without damage to the proof-ring or gauges. Maintenance must ensure that dust does not close this gap and cause the proof-ring to carry only part of the load. Bridging and binding are common causes of loadcells reading below the load weight.

Multi-cell weighing systems can be used, with the readings from each cell being summed electronically. Three cell systems inherently spread the load across all cells. With four cell systems the support structure must ensure that all cells are in contact with the load at all times.

Usually the load cells are the only route to ground from the weigh platform. It is advisable to provide a flexible earth strap, not only for electrical safety but also to provide a route for any welding current which might arise from repairs or later modification.

12.8.6 Weighing systems

A weigh system is usually more than a collection of loadcells and a display. *Figure 12.69* is a taring weigh system used when a 'recipe' of several materials is to be collected in a hopper. The gross weight is the weight from the load cells, i.e. the materials and the hopper itself. Each time the tare command is given, the gross weight is stored and this stored value subtracted from the gross weight to give the net weight display. In this way the weight of each new material can be clearly seen. This type of weigh system can obviously be linked into a supervisory PLC or computer control network.

Figure 12.70 is a batch feeder system where material is fed from a vibrating feeder into a weigh hopper. The system is first tared, then a two speed system used with a changeover from fast to dribble feed a pre-set weight before the target weight. The feeder turns off just before the correct weight to allow for the material in flight from the feeder. This is known as '*in flight compensation*', '*anticipation*' or '*preact*'. To reduce the weigh time the fast feed should be set as fast as possible without causing avalanching in the feed hopper and the changeover to slow feed made as late as possible. The accuracy of the system is mainly dependent on the repeatability of the preact weight, so the slow speed should be set as low as possible.

Material is often carried on conveyor belts and *Figure 12.71* shows a method of providing feed rate in weight per unit time (e.g. kg per min).

The conveyor passes over a load cell of known length and the linear speed is obtained from the drive motor, probably from a tacho-generator. If the weigh platform has length L m, and is indicating weight W kg at speed V ms^{-1}, the feed rate is simply $W \cdot V/L$ kgs^{-1}.

12.9 Fieldbus systems

A Fieldbus, discussed in more detail in Chapter 16, Section 16.5 is a way of interconnecting several devices (e.g. transducers, controllers, actuators etc.) via a simple and cheap serial cable. Devices on the network are identified by addresses which allows messages to be passed between them.

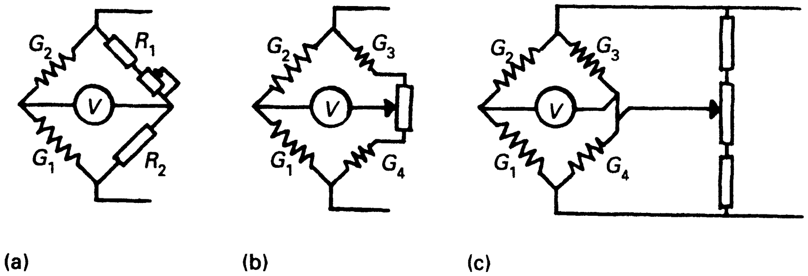

G₂
R₁
V
G₁
R₂
G₂
G₃
V
G₁
G₄
G₂
G₃
V
G₁
G₄
(a)
(b)
(c)

Figure 12.65 Common methods of bridge balancing. (a) Single leg balancing. R_1 is less than R_2; (b) Apex balancing; (c) Parallel balancing

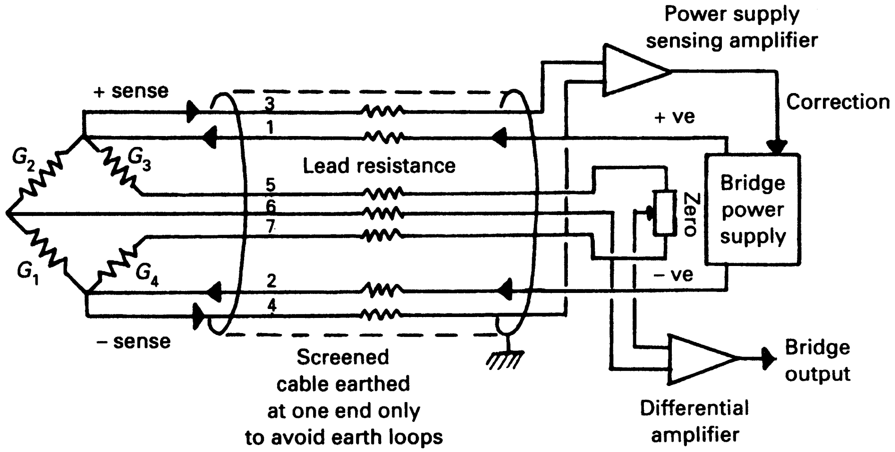

Power supply sensing amplifier
Correction
+ sense
3
1
+ ve
G₂
G₃
Lead resistance
5
6
7
Zero
Bridge power supply
G₁
G₄
2
4
– ve
– sense
Screened cable earthed at one end only to avoid earth loops
Bridge output
Differential amplifier

Figure 12.66 Typical cabling scheme for a remote bridge

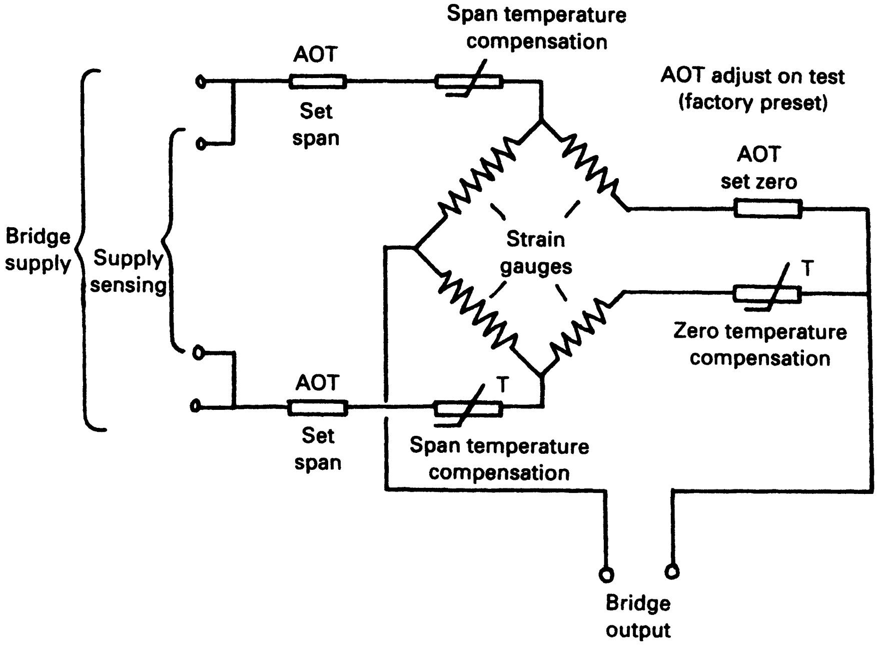

Span temperature compensation
AOT
Set span
AOT adjust on test (factory preset)
AOT set zero
Bridge supply
Supply sensing
Strain gauges
T
Zero temperature compensation
AOT
T
Set span
Span temperature compensation
Bridge output

Figure 12.67 Connection diagram for a typical commercial load cell. Note that a six core cable is required

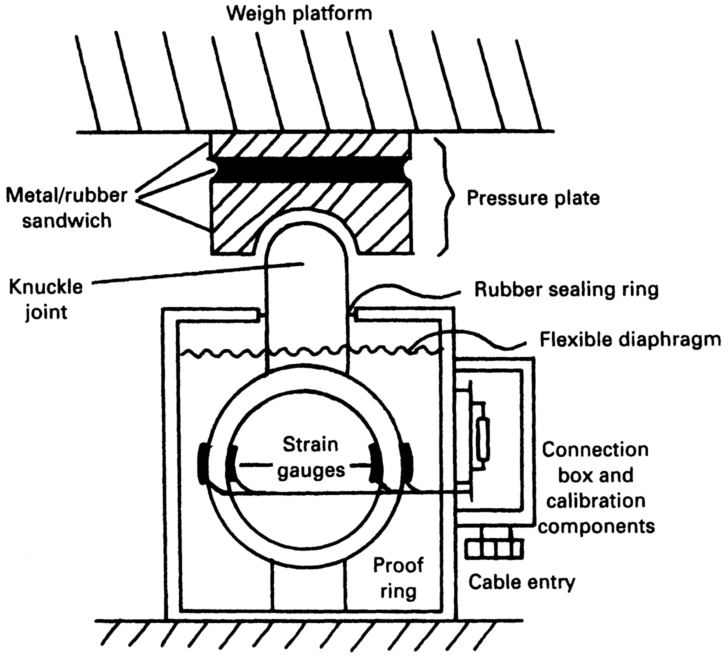

Weigh platform
Metal/rubber sandwich
Pressure plate
Knuckle joint
Rubber sealing ring
Flexible diaphragm
Strain gauges
Connection box and calibration components
Proof ring
Cable entry

Figure 12.68 Construction of a typical commercial loadcell

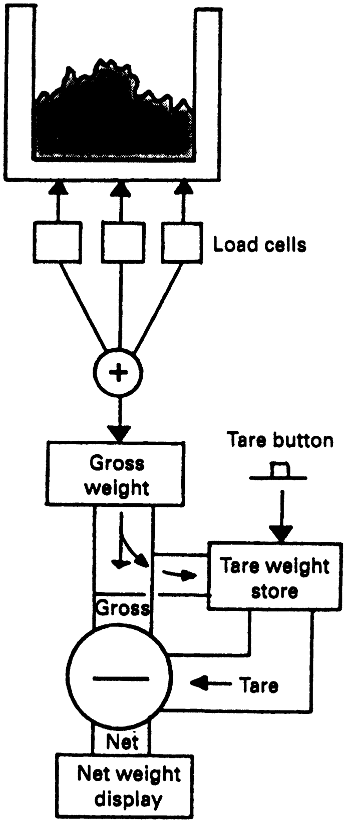

Load cells
+
Gross weight
Tare button
Tare weight store
Gross
—
Tare
Net
Net weight display

Figure 12.69 Taring weighing system

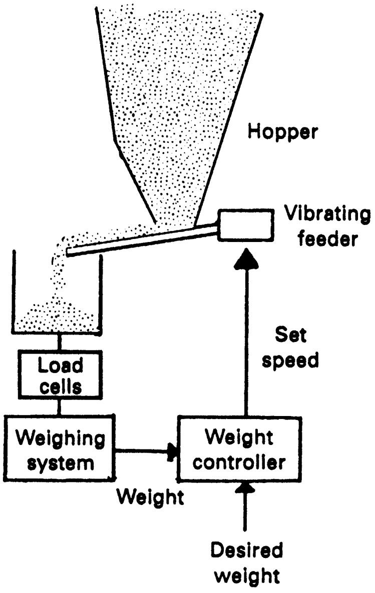

Hopper
Vibrating feeder
Set speed
Load cells
Weighing system
Weight controller
Weight
Desired weight

Figure 12.70 A batch weighing system

Of particular interest in instrumentation is the HART protocol. HART (for Highway Addressable Remote Transducer) was originally introduced by Rosemount as a closed standard for their own devices. Generously, Rosemount have passed the rights to the Hart Communication Foundation and HART is now an open protocol which can be used by anyone. It is a very simple master/slave system; devices only speak when requested and the

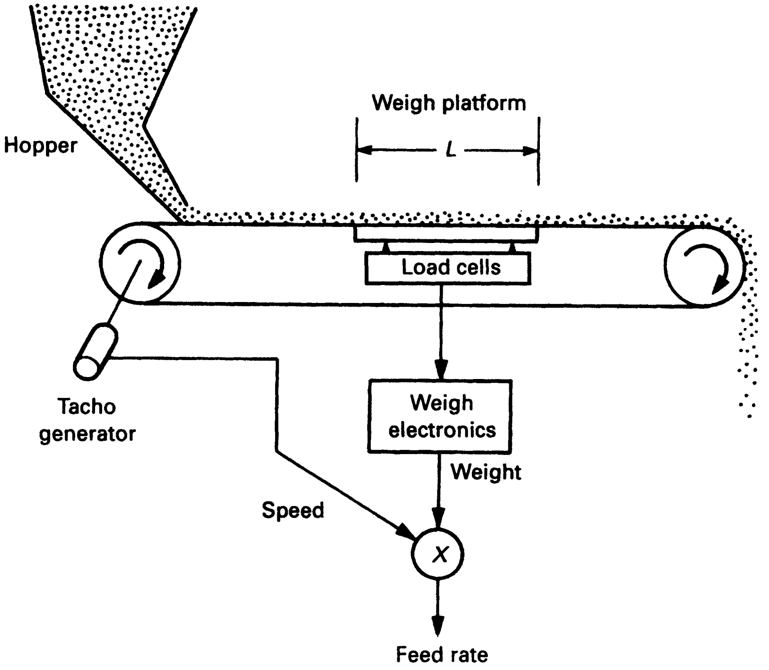

Weigh platform
L
Hopper
Load cells
Tacho generator
Weigh electronics
Weight
Speed
X
Feed rate

Figure 12.71 Continuous belt weighing system

operation is always master request, slave replies. Up to fifteen slaves can be connected to each master.

HART can work in two ways. In its simplest, and probably commonest, form of point to point it superimposes the serial communication data on to a standard 4–20 mA loop signal as shown in *Figure 12.72*(*a*). Frequency shift keying (FSK) is used with frequencies of 1200 Hz for a '1' and 2200 Hz for a '0'. These frequencies are far too high to affect the analogue instrumentation so the analog signal is still used in the normal manner. The system operates with one master (usually a computer, PLC or hand held programming terminal) and one slave (a transducer or actuator). The attraction of this approach is that it allows HART to be retro-fitted onto existing cabling and instrumentation schemes. The disadvantage is that the full cabling benefits of a fieldbus system are not realised.

The serial data, though, allows much more information to be conveyed in addition to the basic analog signal. HART

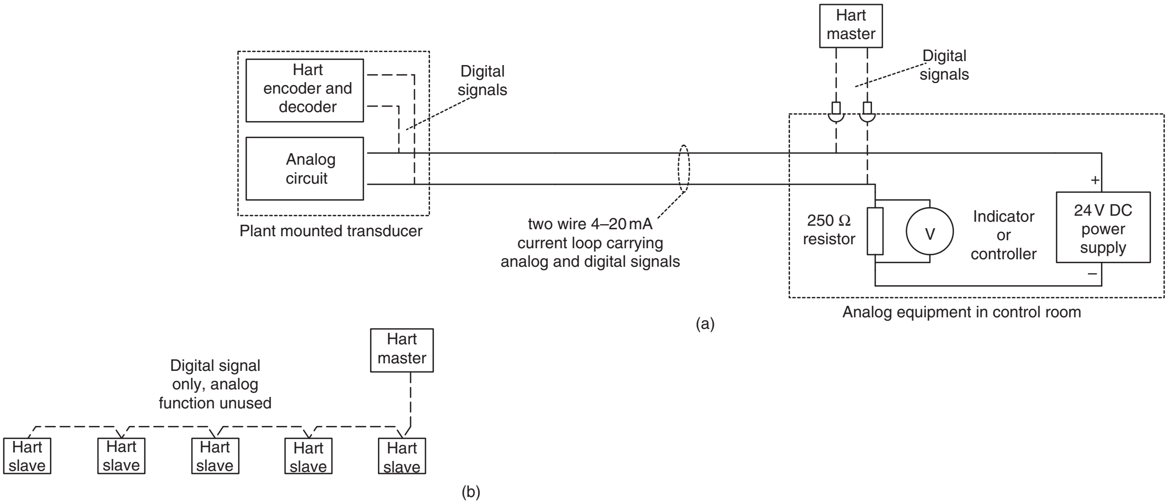

Hart encoder and decoder
Analog circuit
Digital signals
Plant mounted transducer
two wire 4–20 mA current loop carrying analog and digital signals
Hart master
Digital signals
250 Ω resistor
V
Indicator or controller
24 V DC power supply
+
−
Analog equipment in control room
(a)
Digital signal only, analog function unused
Hart master
Hart slave
Hart slave
Hart slave
Hart slave
Hart slave
(b)

Figure 12.72 The Hart communication system. (a) Hart digital data superimposed onto a two wire 4–20 mA analog loop to give additional data. A Hart master (e.g. a PLC or hand held terminal) is used to read the digital data. The transducer can also be configured by the master; (b) A Hart based fieldbus system. All data transfer is done via digital communications. Hart is not as efficient in this mode of operation as normal Fieldbus systems

devices can all be remotely configured and monitored allowing very simple diagnostics and quick replacement after a failure. In addition much more plant data can be passed from the transducer. A flow transducer giving flow as the 4–20 mA analogue signal can also give temperature and pressure via the HART link along with diagnostic information about the transducer status. With a HART programming terminal connected to the line, the transducer can also be made to send fixed currents to aid loop checks and fault diagnosis.

The second method of using HART is in a normal Fieldbus multidrop system as *Figure 12.72(b)*. Here each device has an identifying address and sends data to the master on request. Usually devices (and their relevant parameters) are polled on a regular cyclic scan. The multidrop brings all the cost savings from simple cabling and makes the system easy to modify and expand. By comparison with other fieldbus systems, however, HART is rather slow, and this slow response time can cause difficulties in some applications. The speed is generally adequate for simple monitoring systems where the process variables cannot change quickly.

12.10 Installation notes

Analog systems are generally based on low voltages and are consequently vulnerable to electrical noise. In most plants, a PLC may be controlling 415 V high power motors at 100 A, and reading thermocouple signals of a few mV. Great care must be taken to avoid interference from the high voltage signals.

The first precaution is to adopt a sensible earthing layout. A badly laid out system, as *Figure 12.73*, will have common return paths, and currents from the high powered load returning through the common impedance Z_e will induce error voltages into the low level analog circuit. It should be realised that there are at least three distinct 'earths' in a system.

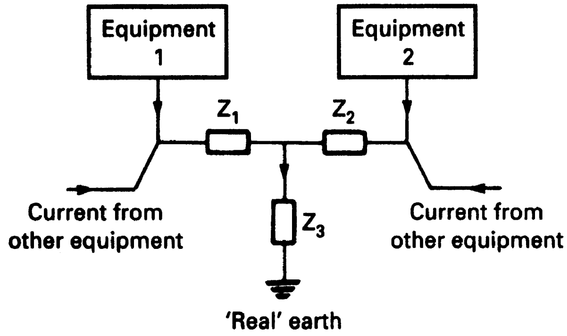


Figure 12.73 A poorly installed earthing system which leads to interaction between signals

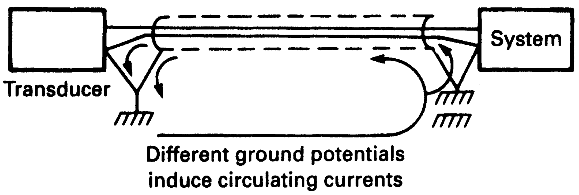


Figure 12.74 An earth loop formed by connecting both ends of a screened cable to the local earth. Screens should be earthed at one point only

- A safety earth (used for doors, frames etc.)
- A dirty earth (used for high voltage/high current signals)
- A clean earth (for low voltage analog signals)

These should meet at one point and one point only (which implies all analog signals should return, and hence be referenced to, the same point).

Screened cable is needed for all analog signals, with foil screening to be used in preference to braided screen. The screen should NOT be earthed at both ends as any difference in earth potential between the two points will cause current to flow in the screen as *Figure.12.74*, and induce noise onto the signal lines. A screen must be earthed at one point only, ideally the receiving end. When a screened cable goes through intermediate junction boxes, screen continuity must be maintained, AND the screen must be sleeved to prevent it touching the frame of the junction boxes. Earthing faults in screened cables can cause very elusive faults.

High voltage and low voltage cables should be well separated, most manufacturers suggest at least one metre between 415 V and low voltage cables but this can be difficult to achieve in practice. In any case, separation can only be achieved until some other person, not knowing the system well, straps a 415 V cable to the same cable tray as a multicore thermocouple cable. It is therefore good practice to use trunking or conduit for low voltage signals as a way of identifying low voltage cables for future installers. The same result can also be achieved by using cables with different coloured PVC sleeves. Inevitably high voltage and low voltage cables will have to cross at some points. If spacing can not be achieved the crossings should always be at 90°.

In an ideal world, separate cubicles should be provided for 110 V/high current devices, and low voltage signals, but this is often not cost effective. Where both types of signals have to share a cubicle the cables should take separate, well separated routes, and high and low voltage devices separated as far as possible.

13

Control Systems

J O Flower Bsc(Eng), PhD, DSc(Eng), CEng, FIEE, FIMarE, MSNAME
University of Warwick

E A Parr MSc, CEng, MIEE, MInstMC
CoSteel Sheerness

Contents

13.1 Introduction 13/3

13.2 Laplace transforms and the transfer function 13/3
  13.2.1 Laplace transformation 13/3
  13.2.2 The transfer function 13/3
  13.2.3 Certain theorems 13/5

13.3 Block diagrams 13/6

13.4 Feedback 13/6

13.5 Generally desirable and acceptable behaviour 13/7

13.6 Stability 13/8

13.7 Classification of system and static accuracy 13/9
  13.7.1 Classification 13/9
  13.7.2 Static accuracy 13/9
  13.7.3 Steady-state errors due to disturbances 13/10

13.8 Transient behaviour 13/10
  13.8.1 First-order system 13/10
  13.8.2 Second-order system 13/11
  13.8.3 Velocity feedback 13/12
  13.8.4 Incorporation of integral control 13/12

13.9 Root-locus method 13/13
  13.9.1 Rules for construction of the root locus 13/14

13.10 Frequency-response methods 13/15
  13.10.1 Nyquist plot 13/16
  13.10.2 Bode diagram 13/17
  13.10.3 Nichols chart 13/20

13.11 State-space description 13/20

13.12 Sampled-data systems 13/24

13.13 Some necessary mathematical preliminaries 13/25
  13.13.1 The *z* transformation 13/25

13.14 Sampler and zero-order hold 13/25

13.15 Block diagrams 13/26

13.16 Closed-loop systems 13/27

13.17 Stability 13/28

13.18 Example 13/28

13.19 Dead-beat response 13/30

13.20 Simulation 13/30
  13.20.1 System models 13/30
  13.20.2 Integration schemes 13/32
  13.20.3 Organisation of problem input 13/32
  13.20.4 Illustrative example 13/32

13.21 Multivariable control 13/33

13.22 Dealing with non-linear elements 13/35
  13.22.1 Introduction 13/35
  13.22.2 The describing function 13/35
  13.22.3 State space and the phase plane 13/38

13.23 Disturbances 13/42
  13.23.1 Introduction 13/42
  13.23.2 Cascade control 13/43
  13.23.3 Feedforward 13/44

13.24 Ratio control 13/45
  13.24.1 Introduction 13/45
  13.24.2 Slave follow master 13/46
  13.24.3 Lead lag control 13/47

13.25 Transit delays 13/47
  13.25.1 Introduction 13/47
  13.25.2 The Smith predictor 13/48

13.26 Stability 13/48
  13.26.1 Introduction 13/48
  13.26.2 Definitions and performance criteria 13/48
  13.26.3 Methods of stability analysis 13/51

13.27 Industrial controllers 13/52
  13.27.1 Introduction 13/52
  13.27.2 A commercial controller 13/52
  13.27.3 Bumpless transfer 13/54
  13.27.4 Integral windup and desaturation 13/54
  13.27.5 Selectable derivative action 13/55
  13.27.6 Variations on the PID algorithm 13/55
  13.27.7 Incremental controllers 13/56
  13.27.8 Scheduling controllers 13/56
  13.27.9 Variable gain controllers 13/56
  13.27.10 Inverse plant model 13/56

13.28 Digital control algorithms 13/57
  13.28.1 Introduction 13/57
  13.28.2 Shannon's sampling theorem 13/58
  13.28.3 Control algorithms 13/59

13.29 Auto-tuners 13/59

13.30 Practical tuning methods 13/60
  13.30.1 Introduction 13/60
  13.30.2 Ultimate cycle methods 13/61
  13.30.3 Bang/bang oscillation test 13/61
  13.30.4 Reaction curve test 13/61
  13.30.5 A model building tuning method 13/62
  13.30.6 General comments 13/63

13.1 Introduction

Examples of the conscious application of feedback control ideas have appeared in technology since very early times: certainly the float-regulator schemes of ancient Greece were notable examples of such ideas. Much later came the automatic direction-setting of windmills, the Watt governor, its derivatives, and so forth. The first third of the 1900s witnessed applications in areas such as automatic ship steering and process control in the chemical industry. Some of these later applications attracted considerable analytical effort aimed at attempting to account for the seemingly capricous dynamic behaviour that was sometimes found in practice.

However, it was not until during, and immediately after, World War II that the fundamentals of the above somewhat disjointed control studies were subsumed into a coherent body of knowledge which became recognised as a new engineering discipline. The great thrust in achieving this had its main antecedents in work done in the engineering electronics industry in the 1930s. Great theoretical strides were made and the concept of feedback was, for the first time, recognised as being all pervasive. The practical and theoretical developments emanating from this activity, constitute the classical approach to control which are explored in some detail in this chapter.

Since the late 1940s, tremendous efforts have been made to expand the boundaries of control engineering theory. For example, ideas from classical mechanics and the calculus of variations have been adapted and extended from a control-theoretic viewpoint. This work is based largely on the state-space description of systems (this description is briefly described in Section 13.11). However, it must be admitted that the practical uses and advantages of many of these developments have yet to be demonstrated. Most control system design work is still based on the classical work mentioned previously. Moreover, nowadays these applications rely, very heavily, on the use of computer techniques; indeed, computers are commonly used as elements in control loops.

Techniques from the 'classical' period of control engineering development is easily understood, wide-ranging in application and, perhaps most importantly, capable of coping with deficiencies in detailed knowledge about the system to be controlled.

These techniques are easily adapted for use in the computer-aided design of control systems, and have proved themselves capable of extension into the difficult area of multi-variable system control; however, this latter topic is beyond the scope of this chapter. So with the above comments in mind, a conventional basic approach to control theory is presented, with a short discussion of the state-space approach and a more extensive forage into sampled-data systems. These latter systems have become important owing to the incorporation of digital computers, particularly microcomputers, into the control loop. Fortunately, an elementary theory for sampled data can be established which nicely parallels the development of basic continuous control theory.

The topics covered in this introduction, and extensions of them, have stood practitioners in good stead for several decades now, and can be confidently expected to go on delivering good service for some decades to come.

13.2 Laplace transforms and the transfer function

In most engineering analysis it is usual to produce mathematical models (of varying precision) to predict the behaviour of physical systems. Often such models are manifested by a differential equation description. This appears to fit in with the causal behaviour of idealised components, e.g. Newton's law relating the second derivative of displacement to the applied force. It is possible to model such behaviour in other ways (for example, using integral equations), although these are much less familiar to most engineers. All real systems are non-linear; however, it is fortuitous that most systems behave approximately like linear ones, with the implication that superposition holds true to some extent. We further restrict the coverage here in that we shall be concerned particularly with systems whose component values are not functions of time—at least over the time-scale of interest to us.

In mathematical terms this latter point implies that the resulting differential equations are not only linear, but also have constant coefficients, e.g. many systems behave approximately according to the equation

$$\frac{d^2x}{dt^2} + 2\zeta\omega_n \frac{dx}{dt} + \omega_n^2 x = \omega_n^2 f(t) \tag{13.1}$$

where x is the dependent variable (displacement, voltage, etc.), $f(t)$ is a forcing function (force, voltage source, etc.), and ω_n^2 and ζ are constants the values of which depend on the size and interconnections of the individual physical components making up the system (spring-stiffness constant, inductance values, etc.).

Equations having the form of Equation (13.1) are called 'linear constant coefficient ordinary differential equations' (LCCDE) and may, of course, be of any order. There are several techniques available for solving such equations but the one of particular interest here is the method based on the Laplace transformation. This is treated in detail elsewhere, but it is useful to outline the specific properties of particular interest here.

13.2.1 Laplace transformation

Given a function $f(t)$, then its Laplace transformation $F(s)$ is defined as

$$L[f(t)] = F(s) = \int_0^{\infty} f(t)\exp(-st)\,dt$$

where, in general, s is a complex variable and of such a magnitude that the above integral converges to a definite functional value.

A list of Laplace transformation pairs is given in *Table 13.1*.

The essential usefulness of the Laplace transformation technique in control engineering studies is that it transforms LCCDE and integral equations into algebraic ones and, hence, makes for easier and standard manipulation.

13.2.2 The transfer function

This is a central notion in control work and is, by definition, the Laplace transformation of the output of a system divided by the Laplace transformation of the input, with the tacit assumption that all initial conditions are at zero.

Thus, in *Figure 13.1*, where $y(t)$ is the output of the system and $u(t)$ is the input, then the transfer function $G(s)$ is

$$L[y(t)]/L[u(t)] = Y(s)/U(s) = G(s)$$

Supposing that $y(t)$ and $u(t)$ are related by the general LCCDE

$$a_n \frac{d^n y}{dt^n} + a_{n-1}\frac{d^{n-1}y}{dt^{n-1}} + \cdots + a_0 y$$

Table 13.1 Laplace transforms and *z* transforms

$f(t)$	$F(s)$	$F(z)$
0	0	0
$f(t-nT)$	$\exp(-nsT)F(s)$	$z^{-n}F(z)$
$\delta(t)$	1	1
$\delta(t-nT)$	$\exp(-nsT)$	z^{-n}
$\sum_{n=0}^{\infty}\delta(t-nT)$	$[1-\exp(-st)]^{-1}$	$z(z-1)^{-1}$
$h(t)$	s^{-1}	$z(z-1)^{-1}$
$u_T(t)$	$[1-\exp(-sT)]s^{-1}$	—
A	As^{-1}	$Az(z-1)^{-1}$
t	s^{-2}	$Tz(z-1)^{-2}$
$f(t)t$	$-\mathrm{d}F(s)/\mathrm{d}s$	—
$(t-nT)h(t-nT)$	$\exp(-nsT)s^{-2}$	$Tz^{-(n-1)}(z-1)^{-2}$
t^2	$2s^{-3}$	$T^2z(z+1)(z-1)^{-3}$
t_n	$n!s^{-(n+1)}$	—
$\exp(\alpha t)$	$(s-\alpha)^{-1}$	$z(z-\exp(\alpha T))^{-1}$
$f(t)\exp(\alpha t)$	$F(s-\alpha)$	$F[z\exp(-\alpha T)]$
$\delta(t)+\alpha\exp(\alpha t)$	$s(s-\alpha)^{-1}$	—
$t\ \exp(\alpha t)$	$(s-\alpha)^{-2}$	$TZ\ \exp(\alpha T)[z-\exp(\alpha T)]^{-2}$
$t^n\ \exp(\alpha t)$	$n!(s-\alpha)^{-(n+1)}$	—
$\sin\ \omega t$	$\dfrac{\omega}{s^2+\omega^2}$	$\dfrac{z\sin\omega T}{z^2-2z\cos\omega T+1}$
$\cos\ \omega t$	$\dfrac{s}{s^2+\omega^2}$	$\dfrac{z(z-\cos\omega T)}{z^2-2z\cos\omega T+1}$
$\dfrac{t}{2\omega}\sin\omega t$	$\dfrac{s}{(s^2+\omega^2)}$	—
$\dfrac{1}{2\omega}(\sin\omega t-\omega t\cos\omega t)$	$\dfrac{\omega^2}{(s^2+\omega^2)^2}$	—
$\dfrac{1}{\cos\delta}\sin(\omega t+\delta)$	$\dfrac{A}{s^2+\omega^2}\left(s+\dfrac{\omega}{A}\right)$ where $\tan\delta=A$	—
$\dfrac{1}{\cos\delta}\cos(\omega t+\delta)$	$\dfrac{1}{s^2+\omega^2}(s-A\omega)$	—
$\exp(\alpha t)\ \sin\ \omega t$	$\dfrac{\omega}{(s-\alpha)^2+\omega^2}=\dfrac{\omega}{(s-\alpha+j\omega)(s-\alpha-j\omega)}$	$\dfrac{z\exp(\alpha T)\sin\omega T}{z^2-2z\exp(\alpha T)\cos\omega T+\exp(2\alpha T)}$
$\exp(\alpha t)\ \cos\ \omega t$	$\dfrac{s-\alpha}{(s-\alpha)^2+\omega^2}$	$\dfrac{z[z-\exp(\alpha T)\cos\omega T]}{z^2-2z\exp(\alpha T)\cos\omega T+\exp(2\alpha T)}$
$\dfrac{t}{2\omega}\exp(\alpha t)\sin\omega t$	$\dfrac{s-\alpha}{[(s-\alpha)^2+\omega^2]^2}$	—
$\dfrac{1}{2\omega}\exp(\alpha t)(\sin\omega t-\cos\omega t)$	$\dfrac{\omega^2}{[(s-\alpha)^2+\omega^2]^2}$	—
$\dfrac{1}{\cos\delta}\exp(\alpha t)\sin(\omega t+\delta)$	$\dfrac{A}{(s-\alpha)^2+\omega^2}\left(s-\alpha+\dfrac{\omega}{A}\right)$ where $\tan\delta=A$	—
$\dfrac{1}{\cos\delta}\exp(\alpha t)\cos(\omega t+\delta)$	$\dfrac{1}{(s-\alpha)^2+\omega^2}(s-\alpha-A\omega)$	—
$\sinh\ \omega t$	$\omega(s^2-\omega^2)^{-1}$	—
$\cosh\ \omega t$	$s(s^2-\omega^2)^{-1}$	—

cont'd

Table 13.1 (*continued*)

$f'(t)$	$sF(s)-f(0-)$	—
$f''(t)$	$s^2F(s)-sf(0-)-f'(0-)$	—
$f^n(t)$	$s^nF(s)-s^{n-1}f(0-)-s^{n-2}f'(0-)\ \ldots\ -f^{n-1}(0-)$	—
$f^{-1}(t)$	$\dfrac{F(s)}{s}+\Leftarrow\dfrac{f^{-1}(0-)}{s}$	—
$f(t)$ $t\to 0$	$sF(s)$ $s\to\infty\Leftarrow$	$F(z)$ $z\to\infty\Leftarrow$
$f(t)$ $t\to\infty\Leftarrow$	$sF(s)$ $s\to 0$	$(z-1)z^{-1}F(z)$ $z\to 1$

$\delta(t)$, The unit impulse function.
$h(t)$, The unit step function.
$u_T(t)$, The unit step function followed by a unit negative step at $t=T$, where T is the sampling period.

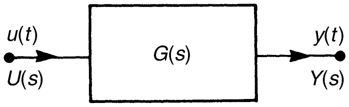


Figure 13.1 Input–output representation

$$= b_m\frac{\mathrm{d}^m u}{\mathrm{d}t^m}+b_{m-1}\frac{d^{m-1}u}{\mathrm{d}t^{m-1}}+\cdots+b_0u \qquad (13.2)$$

then, on Laplace transforming and ignoring initial conditions, we have (see later for properties of Laplace transformation)

$$(a_ns^n+a_{n-1}s^{n-1}+\cdots+a_0)Y(s)\Leftarrow$$
$$=(b_ms^m+b_{m-1}s^{m-1}+\cdots+b_0)U(s)\Leftarrow$$

whence

$$\frac{Y(s)\Leftarrow}{U(s)\Leftarrow}\equiv G(s)=\sum_{i=0}^{m}\Bigg(b_is^i\Big/\sum_{i=0}^{n}a_is^i$$

There are a number of features to note about $G(s)$.

(1) Invariably $n >$ɋn for physical systems.
(2) It is a ratio of two polynomials which may be written

$$G(s)=\Leftarrow\frac{b_m(s-z_1)\ldots(s-z_m)\Leftarrow}{a_n(s-p_1)\ldots(s-p_n)\Leftarrow}$$

$z_1,\ldots,z_m$ are called the *zeros* and $p_1,\ldots,p_n$ are called the *poles* of the transfer function.
(3) It is not an explicit function of input or output, but depends entirely upon the nature of the system.
(4) The block diagram representation shown in *Figure 13.1* may be extended so that the interaction of composite systems can be studied (provided that they do not load each other); see below.
(5) If $u(t)$ is a delta function $\delta(t)$, then $U(s)=1$, whence $Y(s)=G(s)$ and $y(t)=g(t)$, where $g(t)$ is the *impulse response* (or weighting function) of the system.
(6) Although a particular system produces a particular transfer function, a particular transfer function does not imply a particular system, i.e. the transfer function specifies merely the input–output relationship between two variables and, in general, this relationship may be realised in an infinite number of ways.
(7) Although we might expect that all transfer functions will be ratios of finite polynomials, an important and common element which is an exception to this is the pure-delay element. An example of this is a loss-free transmission line in which any disturbance to the input of the line will appear at the output of the line without distortion, a finite time (say τ) later. Thus, if $u(t)$ is the input, then the output $y(t)=u(t-\tau)$ and the transfer function $Y(s)/U(s)=\exp(-s\tau)$. Hence, the occurrence of this term within a transfer function expression implies the presence of a pure delay; such terms are common in chemical plant and other fluid-flow processes.

Having performed any manipulations in the Laplace transformation domain, it is necessary for us to transform back to the time domain if the time behaviour is required. Since we are dealing normally with the ratio of polynomials, then by partial fraction techniques we can arrange $Y(s)$ to be written in the following sequences:

$$Y(s)=\Leftarrow\frac{K(s-z_1)(s-z_2)\ldots(s-z_m)\Leftarrow}{(s-p_1)(s-p_2)\ldots(s-p_n)\Leftarrow}$$

$$Y(s)=K\left[\Bigg(\frac{A_1}{s-p_1}+\Leftarrow\frac{A_2}{s-p_2}+\cdots+\Leftarrow\frac{A_n}{s-p_n}\right]\Bigg($$

and by so arranging $Y(s)$ in this form the conversion to $y(t)$ can be made by looking up these elemental forms in *Table 13.1.*

Example Suppose that

$$Y(s)=\Leftarrow\frac{5(s^2+4s+3)\Leftarrow}{s^3+6s^2+8s}=\Leftarrow\frac{5(s^2+4s+3)\Leftarrow}{s(s+2)(s+4)\Leftarrow}$$
$$=5\left[\frac{3}{8s}+\frac{1}{4(s+2)\Leftarrow}+\Leftarrow\frac{3}{8(s+4)}\Leftarrow\right]\Bigg($$

Then

$$y(t)=\Leftarrow\frac{5}{4}\left[\frac{3}{2}\{1+\exp(-4t)\}+\exp(-2t)\Leftarrow\right]\Bigg($$

13.2.3 Certain theorems

A number of useful transform theorems are quoted below, without proof.

(1) *Differentiation*
If $F(s)$ is the Laplace transformation of $f(t)$, then

$$L[d^nf(t)/\mathrm{d}t^n]=s^nF(s)-s^{n-1}f(0)=s^{n-2}f'(0)-\cdots-f^{n-1}(0)\Leftarrow$$

For example, if $f(t) = \exp(-bt)$, then

$$L\left[\frac{d^3}{dt^3}\exp(-bt)\right] = \frac{s^3}{s+b} - s^2 + bs - b^2$$

(2) *Integration*
If $L[f(t)] = F(s)$, then

$$L\left[\int_0^1 f(t)\,dt\right] = \frac{F(s)}{s} + f(0)$$

Repeated integration follows in a similar fashion.

(3) *Final-value theorem*
If $f(t)$ and $f'(t)$ are Laplace transformable and if $L[f(t)] = F(s)$, then if the limit of $f(t)$ exists as t goes towards infinity, then

$$\lim_{s\to 0} sF(s) = \lim_{t\to\infty} f(t)$$

For example,

$$F(s) = \frac{b-a}{s(s+a)(s+b)}$$

then

$$\lim_{s\to 0}\frac{s(b-a)}{s(s+a)(s+b)} = \frac{b-a}{ab} = \lim_{t\to\infty} f(t)$$

(4) *Initial-value theorem*
If $f(t)$ and $f'(t)$ are Laplace transformable and if $L[f(t)] = F(s)$,
then

$$\lim_{s\to\infty} sF(s) = \lim_{t\to 0} f(t)$$

(5) *Convolution*
If $L[f_1(t)] = F_1(s)$ and $L[f_2(t)] = F_2(s)$, then

$$F_1(s)\cdot F_2(s) = L\left[\int_0^\infty f_1(t-\tau)\cdot f_2(\tau)\,d\tau\right]$$

13.3 Block diagrams

It is conventional to represent individual transfer functions by boxes with an input and output (see note (4) in Section 13.2.2). Provided that the components represented by the transfer function do not load those represented by the transfer function in a connecting box, then simple manipulation of the transfer functions can be carried out. For example, suppose that there are two transfer functions in cascade (see *Figure 13.2*): then we may write $X(s)/U(s) = G_1(s)$ and $Y(s)/X(s) = G_2(s)$. Eliminating $X(s)$ by multiplication, we have

$$Y(s)/U(s) = G_1(s)G_2(s)$$

which may be represented by a single block. This can obviously be generalised to any number of blocks in cascade.

Another important example of block representation is the prototype feedback arrangement shown in *Figure 13.3*. We see that $Y(s) = G(s)E(s)$ and $E(s) = U(s) - H(s)Y(s)$. Eliminating $E(s)$ from these two equations results in

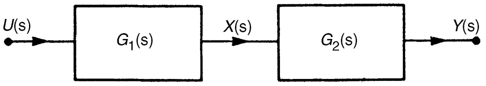


Figure 13.2 Systems in cascade

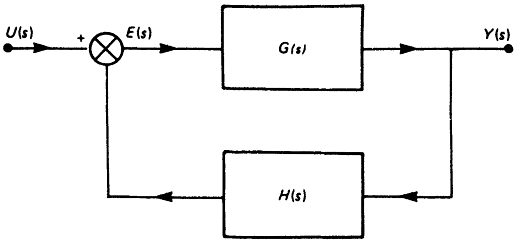


Figure 13.3 Block diagram of a prototype feedback system

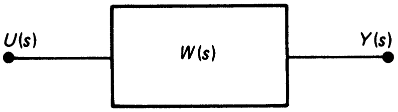


Figure 13.4 Reduction of the diagram shown in *Figure 13.3* to a single block

$$\frac{Y(s)}{U(s)} = \frac{G(s)}{1 + H(s)G(s)} = W(s)$$

In block diagram form we have *Figure 13.4*. If we eliminate $Y(s)$ from the above equations, we obtain

$$\frac{E(s)}{U(s)} = \frac{1}{1 + (H(s)G(s))}$$

13.4 Feedback

The last example is the basic feedback conceptual arrangement, and it is pertinent to investigate it further, as much effort in dealing with control systems is devoted to designing such feedback loops. The term ‘feedback’ is used to describe situations in which a portion of the output (and/or processed parts of it) are fed back to the input of the system. The appropriate application may be used, for example, to improve bandwidth, improve stability, improve accuracy, reduce effects of unwanted disturbances, compensate for uncertainty and reduce the sensitivity of the system to component value variation.

As a concrete example consider the system shown in *Figure 13.5*, which displays the arrangements for an angular position control system in which a desired position θ_r is indicated by tapping a voltage on a potentiometer. The actual position of the load being driven by the motor (usually via a gearbox) is monitored by θ_o, indicated, again electrically, by a potentiometer tapping. If we assume identical potentiometers energised from the same voltage supply, then the misalignment between the desired output and the actual output is indicated by the difference between the respective potentiometer voltages. This difference (proportional to error) is fed to an amplifier whose output, in turn,

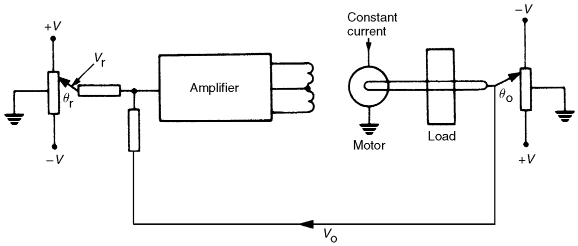


Figure 13.5 Schematic diagram of a simple position and control system

drives the motor. Thus, the arrangement seeks to drive the system until the output θ_o and input θ_r are coincident (i.e. the error is zero).

In the more general block diagram form, the above schematic will be transformed to that shown in *Figure 13.6*, where $\theta_r(s)$, $\theta_o(s)$ are the Laplace transforms of the input, output position: $K_1(s)$ and $K_2(s)$ are the potentiometer transfer functions (normally taken as straight gains); $Vr(s)$ is the Laplace transform of the reference voltage; $Vo(s)$ is the Laplace transform of the output voltage; $G_m(s)$ is the motor transfer function; $G_l(s)$ is the load transfer function; and $A(s)$ is the amplifier transfer function.

Let us refer now to *Figure 13.3* in which $U(s)$ is identified as the transformed input (reference or demand) signal, $Y(s)$ is the output signal and $E(s)$ is the error (or actuating) signal. $G(s)$ represents the *forward transfer function* and is the product of all the transfer functions in the forward loop, i.e. $G(s) = A(s)G_m(s)G_l(s)$ in the above example.

$H(s)$ represents the *feedback transfer function* and is the product of all transfer functions in the feedback part of the loop.

We saw in Section 13.3 that we may write

$$\frac{Y(s)}{U(s)} = \frac{G(s)}{1 + H(s)G(s)}$$

and

$$\frac{E(s)}{U(s)} = \frac{1}{1 + H(s)G(s)}$$

i.e. we have related output to input and the error to the input.

The product $H(s)G(s)$ is called the *open-loop transfer function* and $G(s)/[1 + H(s)G(s)]$ the *closed-loop transfer function.* The open-loop transfer function is most useful in studying the behaviour of the system, since it relates the error to the demand. Obviously it would seem desirable for this error to be zero at all times, but since we are normally considering systems containing energy storage components, total elimination of error at all times is impossible.

13.5 Generally desirable and acceptable behaviour

Although specific requirements will normally be drawn up for a particular control system, there are important general requirements applicable to the majority of systems. Usually an engineering system will be assembled from readily available components to perform some function, and the choice of these components will be restricted. An example of this would be a diesel engine–alternator set for delivering electrical power, in which normally the most convenient diesel engine–alternator combination will be chosen from those already manufactured.

Even if such a system were assembled from customer-designed components, it would be fortuitous if it performed in a satisfactory self-regulatory way without further consideration of its control dynamics. Hence, it is the control engineer's task to take such a system and devise economical ways of making the overall system behave in a satisfactory manner under the expected operational conditions.

For example, a system may oscillate, i.e. it is unstable; or, although stable, it might tend to settle after a change in input demand to a value unacceptably far from this new demand, i.e. it lacks static accuracy. Again, it might settle to a satisfactory new steady state, but only after an unsatisfactory transient response. Alternatively, normal operational load disturbances on the system may cause unacceptably wide variation of the output variable, e.g. voltage and frequency of the engine–alternator system.

All these factors will normally be quantified in an actual design specification, and fortunately a range of techniques is available for improving the behaviour. But the application of a particular technique to improve the performance of one aspect of behaviour often has a deleterious effect on another, e.g. improved stability with improved static accuracy tends to be incompatible. Thus, a compromise is sought which gives the 'best' acceptable all-round performance. We now discuss

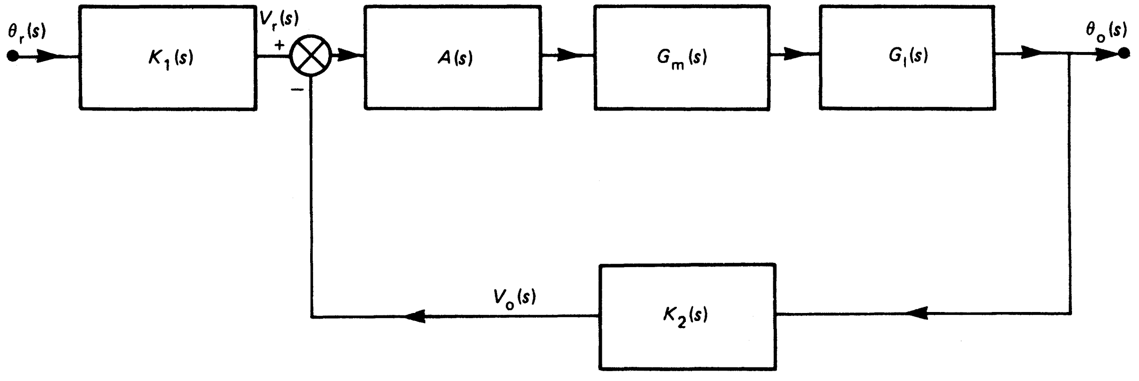


Figure 13.6 Block diagram of the system shown in *Figure 13.5*

some of these concepts and introduce certain techniques useful in examining and designing systems.

13.6 Stability

This is a fairly easy concept to appreciate for the types of system under consideration here. Equation (13.2) with the right-hand side made equal to zero governs the free (or unforced, or characteristic) behaviour of the system, and because of the nature of the governing LCCDE it is well known that the solution will be a linear combination of exponential terms, *viz*.

$$y(t) = \sum_{i=1}^{n} A_i \exp(\alpha_i t) \Leftarrow$$

where the α_i values are the roots of the so-called 'characteristic equation'.

It will be noted that should any α_i have a positive real part (in general, the roots will be complex), then any disturbance will grow in time. Thus, for stability, no roots must lie in the right-hand half of the complex plane or s plane. In a transfer function context this obviously translates to 'the roots of the denominator must not lie in the right-hand half of the complex plane'.

For example, if $W(s) = G(s)/[1 + H(s)G(s)]$, then the roots referred to are those of the equation

$$1 + H(s)G(s) = 0$$

In general, the determination of these roots is a non-trivial task and, as at this stage we are interested only in whether the system is stable or not, we can use certain results from the theory of polynomials to achieve this without the necessity for evaluating the roots.

A preliminary examination of the location of the roots may be made using the *Descartes rule of signs*, which states: if $f(x)$ is a polynomial, the number of positive roots of the equation $f(x) = 0$ cannot exceed the number of changes of sign of the numerical coefficients of $f(x)$, and the number of negative roots cannot exceed the number of changes of sign of the numerical coefficients of $f(-x)$. 'A change of sign' occurs when a term with a positive coefficient is immediately followed by one with a negative coefficient, and vice versa.

Example Suppose that $f(x) = x^3 + 3x - 2 = 0$; then there can be at most one positive root. Since $f(-x) = -x^3 - 3x - 2$, the equation has no negative roots. Further, the equation is cubic and must have at least one real root (complex roots occur in conjugate pairs); therefore the equation has one positive-real root.

Although Descartes' result is easily applied, it is often indefinite in establishing whether or not there is stability, and a more discriminating test is that due to Routh, which we give without proof.

Suppose that we have the polynomial

$$a_0 s^n + a_1 s^{n-1} \dots a_{n-1} s + a_n = 0$$

where all coefficients are positive, which is a necessary (but not sufficient) condition for the system to be stable, and we construct the following so-called 'Routh array':

$$\begin{array}{lllllll}
s^n & : & a_0 & a_2 & a_4 & a_6 & \dots \\
s^{n-1} & : & a_1 & a_3 & a_5 & a_7 & \dots \\
s^{n-2} & : & b_1 & b_2 & b_3 & \dots & \\
s^{n-3} & : & c_1 & c_2 & c_3 & \dots & \\
s^{n-4} & : & d_1 & d_2 & \dots & &
\end{array}$$

where

$$b_1 = \Leftarrow \frac{a_1 a_2 - a_0 a_3}{a_1}, \varsigma b_2 = \Leftarrow \frac{a_1 a_4 - a_0 a_5}{a_1}, \varsigma b_3 = \Leftarrow \frac{a_1 a_6 - a_0 a_7}{a_1}, \dots$$

$$c_1 = \Leftarrow \frac{b_1 a_3 - a_1 b_2}{b_1}, \varsigma c_2 = \Leftarrow \frac{b_1 a_5 - a_1 b_3}{b_1}, \dots \%$$

$$d_1 = \Leftarrow \frac{c_1 b_2 - b_1 c_2}{c_1}, \dots$$

This array will have $n + 1$ rows.

If the array is complete and *none* of the elements in the first column vanishes, then a sufficient condition for the system to be stable (i.e. the characteristic equation has all its roots with negative-real parts) is for all these elements to be positive. Further, if these elements are not all positive, then the number of changes of sign in this first column indicates the number of roots with positive-real parts.

Example Determine whether the polynomial $s^4 + 2s^3 + 6s^2 + 7s + 4 = 0$ has any roots with positive-real parts. Construct the Routh array:

$$\begin{array}{llll}
s^4 : & 1 & 6 & 4 \\
s^3 : & 2 & 7 & \\
s^2 : & \dfrac{(2)(6) - (1)(7)}{2} = 2.5 & \dfrac{(2)(4) - (1)(0)}{2} = 4 & \\
s :! & \dfrac{(2.5)(7) - (2)(4)}{2.5} = 3.8 & & \\
s^0 : & 4 & &
\end{array}$$

There are five rows with the first-column elements all positive, and so a system with this polynomial as its characteristic would be stable.

There are cases that arise which need a more delicate treatment.

(1) Zeros occur in the first column, while other elements in the row containing a zero in the first column are non-zero.

In this case the zero is replaced by a small positive number, ε, which is allowed to approach zero once the array is complete.

For example, consider the polynomial equation

$$s^5 + 2s^4 + 2s^3 + 4s^2 + 11s + 8 = 0:$$

$$\begin{array}{llll}
s^5 : & 1 & 2 & 11 \\
s^4 : & 2 & 4 & 8 \\
s^3 : & \varepsilon\varsigma & 5 & 0 \\
s^2 : & \alpha_1 & 8 & \\
s^1 : & \alpha_2 & 0 & \\
s^0 : & 8 & &
\end{array}$$

where

$$\alpha_1 = \Leftarrow \frac{4\varepsilon\varsigma - 10}{\varepsilon\varsigma} \simeq - \Leftarrow \frac{10}{\varepsilon} \text{ and } \alpha_2 = \frac{5\alpha_1 - 8\varepsilon\varsigma}{\alpha_1} \simeq \Leftarrow 5$$

Thus, α_1 is a large negative number and we see that there are effectively two changes of sign and, hence, the equation has two roots which lie in the right-hand half of this plane.

(2) Zeros occur in the first column and other elements of the row containing the zero are also zero.

This situation occurs when the polynomial has roots that are symmetrically located about the origin of the s plane, i.e. it contains terms such as $(s + j\omega)(s - j\omega)$ or $(s + v)(s - v)$.

This difficulty is overcome by making use of the auxiliary equation which occurs in the row immediately before the zero entry in the array. Instead of the all-zero row the equation formed from the preceding row is differentiated and the resulting coefficients are used in place of the all-zero row.

For example, consider the polynomial $s^3 + 3s^2 + 2s + 6 = 0$.

s^3 : 1 2
s^2 : 3 6 (auxiliary equation $3s^2 + 6 = 0$)
s^1 : 0 0

Differentiate the auxiliary equation giving $6s = 0$, and compile a new array using the coefficients from this last equation, *viz*.

s^3 : 1 2
s^2 : 3 6
s^1 : 6 0
s^0 : 1

Since there are no changes of sign, the system will not have roots in the right-hand half of the s plane.

Although the Routh method allows a straightforward algorithmic approach to determining the stability, it gives very little clue as to what might be done if stability conditions are unsatisfactory. This consideration is taken up later.

13.7 Classification of system and static accuracy

13.7.1 Classification

The discussion in this section is restricted to unity-feedback systems (i.e. $H(s) = 1$) without seriously affecting generalities. We know that the open-loop system has a transfer function $KG(s)$, where K is a constant and we may write

$$KG(s) = \frac{K(s - z_1)(s - z_2)\ldots(s - z_m)}{s^l(s + p_1)(s + p_2)\ldots(s - p_3)} = \frac{K\sum_{k=0}^{m} b_k s^k}{s^l \sum_{k=0}^{n-l} a_k s^k}$$

and for physical systems $n \geq m + 1$.

The *order* of the system is defined as the degree of the polynomial in s appearing in the denominator, i.e. n.

The *rank* of the system is defined as the difference in the degree of the denominator polynomial and the degree of the numerator polynomial, i.e. $n - m \geq 1$.

The *class* (or *type*) is the degree of the s term appearing in the denominator (i.e. l), and is equal to the number of integrators in the system.

Example

(1) $G(s) = \dfrac{s + 1}{s^4 + 6s^3 + 9s^2 + 3s}$

implies order 4, rank 3 and type 1.

(2) $G(s) = \dfrac{s^2 + 4s + 1}{(s + 1)(s^2 + 2s + 4)}$

implies order 3, rank 1 and type 0.

13.7.2 Static accuracy

When a demand has been made on the system, then it is generally desirable that after the transient conditions have decayed the output should be equal to the input. Whether or not this is so will depend both on the characteristics of the system and on the input demand. Any difference between the input and output will be indicated by the error term $e(t)$ and we know that for the system under consideration

$$E(s) = \frac{U(s)}{1 + KG(s)}$$

Let $e_{ss} = \lim_{t\to\infty} e(t)$ (if it exists), and so e_{ss} will be the steady-state error. Now from the final-value theorem we have

$$e_{ss} = \lim_{t\to\infty} e(t) = \lim_{s\to 0}[sE(s)]$$

Thus,

$$e_{ss} = \lim_{s\to 0}\left[\frac{sU(s)}{1 + KG(s)}\right]$$

13.7.2.1 *Position-error coefficient K_p*

Suppose that the input is a unit step, i.e. $R(s) = 1/s$; then

$$e_{ss} = \lim_{s\to 0}\left[\frac{1}{1 + KG(s)}\right] = \frac{1}{1 + \lim_{s\to 0}[KG(s)]} = \frac{1}{1 + K_p}$$

where $K_p = \lim_{s\to 0}[KG(s)]$ and this is called the *position-error coefficient*.

Example For a type-0 system

$$KG(s) = \left[K\sum_{k=0}^{m} b_k s^k\right] \Big/ \left[\sum_{k=0}^{n} a_k s^k\right]$$

Therefore $K_p = K(b_0/a_0)$ and $e_{ss} = 1/(1 + K_p)$.

It will be noted that, after the application of a step, there will always be a finite steady-state error between the input and the output, but this will decrease as the gain K of the system is increased.

Example For a type-1 system

$$KG(s) = \left[K\sum_{k=0}^{m} b_k s^k\right] \Big/ \left[s\left(\sum_{k=0}^{n-1} a_k s^k\right)\right]$$

and

$$K_p = \lim_{s\to 0}\left[K\sum_{k=0}^{m} b_k s^k\right] \Big/ \left[s\left(\sum_{k=0}^{n-1} a_k s^k\right)\right] \to \infty$$

Thus,

$$e_{ss} = \frac{1}{1 + \infty} \to 0$$

i.e. there is no steady-state error in this case and we see that this is due to the presence of the integrator term $1/s$. This is an important practical result, since it implies that steady-state errors can be eliminated by use of integral terms.

13.7.2.2 *Velocity-error coefficient, K_v*

Let us suppose that the input demand is a unit ramp, i.e. $u(t) = t$, so $U(s) = 1/s^2$. Then

$$e_{ss} = \lim_{s\to 0}[sE(s)] = \lim_{s\to 0}\left[\frac{1}{s + sKG(s)}\right] = \frac{1}{\lim_{s\to 0}[sKG(s)]}$$
$$= \frac{1}{K_v}$$

where $K_v = \lim_{s\to 0}[sKG(s)]$ is called the *velocity-error coefficient*.

Examples For a type-0 system $K_v = 0$, whence $e_{ss} \to \infty$.

For a type-1 system $K_v = K(b_0/a_0)$ and so this system can follow but with a finite error.

For a type-2 system

$$K_v = \lim_{s\to 0}\left[\frac{K}{s}\frac{b_0}{a_0}\right]\Bigg(\to \infty \Leftarrow$$

whence $e_{ss} \to 0$ and so the system can follow in the steady state without error.

13.7.2.3 Acceleration-error coefficient K_a

In this case we assume that $u(t) = t^2/2$, so $U(s) = 1/s^3$ and so

$$e_{ss} = \lim_{s\to 0}[sE(s)] = \lim_{s\to 0}\left[\frac{1}{s^2 + s^2KG(s)}\Leftarrow\right]\Bigg($$
$$= \frac{1}{\lim_{s\to 0}[s^2KG(s)]\Leftarrow} = \frac{\Leftarrow 1}{K_a}$$

where $K_a = \lim_{s\to 0}[s^2\ KG(s)]$ is called the *acceleration-error coefficient* and similar analyses to the above may be performed.

These error-coefficient terms are often used in design specifications of equipment and indicate the minimum order of the system that one must aim to design.

13.7.3 Steady-state errors due to disturbances

The prototype unity-feedback closed-loop system is shown in *Figure 13.7* modified by the intrusion of a disturbance $D(s)$ being allowed to affect the loop. For example, the loop might represent a speed-control system and $D(s)$ might represent the effect of changing the load. Now, since linear systems are under discussion, in order to evaluate the effects of this disturbance on $Y(s)$ (denoted by $Y_D(s)$), we may tacitly assume $U(s) = 0$ (i.e. invoke the superposition principle)

$$Y_D(s) = D(s) - KG(s)\,Y_D(s)\Leftarrow$$
$$Y_D(s) = D(s)/[1 + KG(s)]\Leftarrow$$

Now $E_D(s) = -Y_D(s) = -D(s)/[1 + KG(s)]$, and so the steady-state error, e_{ssD} due to the application of the disturbance, may be evaluated by use of the final-value theorem as

$$e_{ssD} = -\lim_{s\to 0}\left[\Bigg(\frac{sD(s)\Leftarrow}{1 + KG(s)}\right]\Bigg($$

Obviously the disturbance may enter the loop at other places but its effect may be established by similar analysis.

13.8 Transient behaviour

Having developed a means of assessing stability and steady-state behaviour, we turn our attention to the transient behaviour of the system.

13.8.1 First-order system

It is instructive to examine first the behaviour of a first-order system (a first-order lag with a time constant T) to a unit-step input (*Figure 13.8*).

Now

$$\frac{Y(s)\Leftarrow}{U(s)\Leftarrow} = G(s) = \frac{\Leftarrow 1}{1 + sT}$$

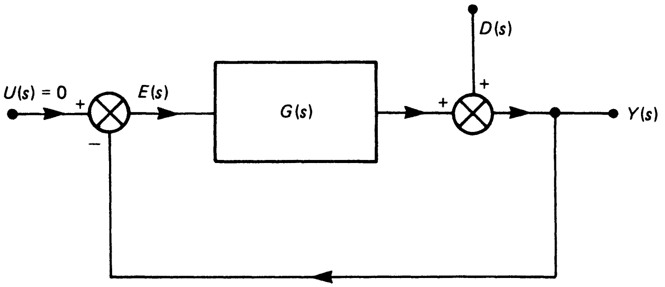


Figure 13.7 Schematic diagram of a disturbance entering the loop

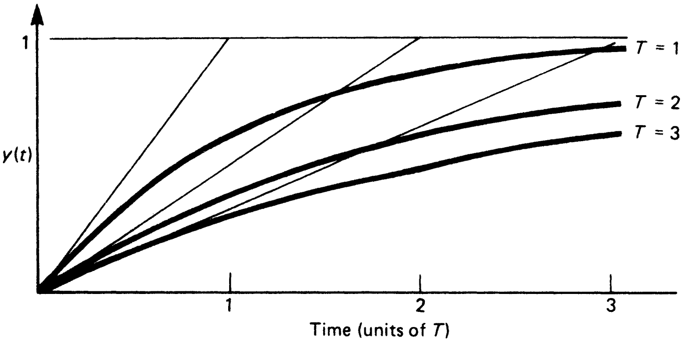


Figure 13.8 First-order lag response to a unit step (time constant = 1, 2, 3, units)

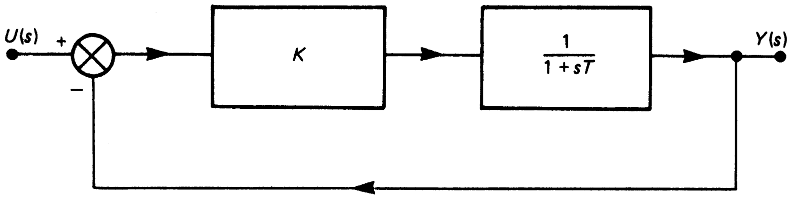


Figure 13.9 First-order lag incorporated in a feedback loop

where $U(s) = 1/s$

$$Y(s) = \frac{1}{s(1+sT)} = \frac{1}{Ts[s+(1/T)]} = \frac{1}{s} - \frac{1}{[s+(1/T)]}$$

or $y(t) = 1 - \exp(-t/T)$;
note also that $\mathrm{d}y/\mathrm{d}t = (1/T)\exp(-t/T)$.

Figure 13.8 shows this time response for different values of T where it will be noted that the corresponding trajectories have slopes of $1/T$ at time $t = 0$ and reach approximately 63% of their final values after T.

Suppose now that such a system is included in a unity-feedback arrangement together with an amplifier of gain K (*Figure 13.9*); therefore

$$\frac{Y(s)}{U(s)} = \frac{K/(1+sT)}{1+K/(1+sT)} = \frac{K}{(1+K)\left(1+s\dfrac{T}{1+K}\right)}$$

For a unit-step input the time response will be

$$y(t) = \frac{K}{1+K}[1 - \exp\{-(1+K)(t/T)\}]$$

This expression has the same form as that obtained for the open loop but the effective time constant is modified by the gain and so is the steady-state condition (*Figure 13.10*). Such an arrangement provides the ability to control the effective time constant by altering the gain of an amplifier, the original physical system being left unchanged.

13.8.2 Second-order system

The behaviour characteristics of second-order systems are probably the most important of all, since many systems of seemingly greater complexity may often be approximated by a second-order system because certain poles of their transfer function dominate the observed behaviour. This has led to system specifications often being expressed in terms of second-order system behavioural characteristics.

In Section 13.2 the importance of the second-order behaviour of a generator was mentioned, and this subject is now taken further by considering the system shown in *Figure 13.11*.

The closed-loop transfer function for this system is given by

$$W(s) = \frac{KG(s)}{1+KG(s)} = \frac{K}{s^2+as+K}$$

and this may be rewritten in general second-order terms in the form

$$W(s) = \frac{\omega_n^2}{s^2+2\zeta\omega_n s+\omega_n^2}$$

where $K = \omega_n^2$ and $\zeta = a/(2\sqrt{K})$. The unit-step response is given by

$$y(t) = 1 - \exp(-\zeta\omega_n t)[\cos(\gamma\omega_n t) - (\zeta/\gamma)\sin(\gamma\omega_n t)]$$

where $\gamma = \sqrt{(1-\zeta^2)}$. This assumes, of course, that $\zeta < 1$, so giving an oscillating response decaying with time.

The *rise time* t_r will be defined as the time to reach the first overshoot (note that other definitions are used and it is important to establish which particular definition is being used in a particular specification):

$$t_r = \pi/(\gamma\omega_n) = \pi/\sqrt{[K-(a/2)^2]}$$

i.e. the rise time decreases as the gain K is increased.

The *percentage overshoot* is defined as:

$$\text{Percentage overshoot} = \frac{100(\text{Max. value of } y(t) - \text{Steady-state value})}{\text{Steady-state value}}$$

$$= 100\exp(-\zeta\pi/\gamma) = 100\exp[-a\pi/\sqrt{(4K-a^2)}]$$

i.e. the percentage overshoot increases as the gain K increases.

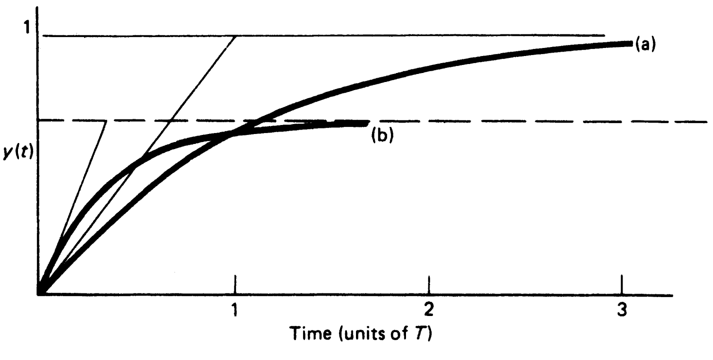


Figure 13.10 Response of first-order lag: (a) open-loop condition ($T = 1$); (b) closed-loop condition ($T = 1$, $K = 2$)

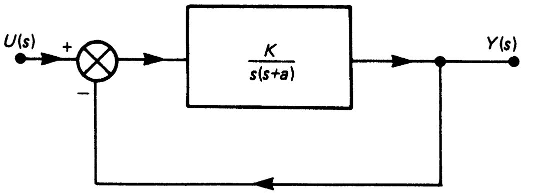


Figure 13.11 Second-order system

The *frequency of oscillation* ω_r is immediately seen to be

$$\omega_r = \omega_n \gamma\varsigma = \sqrt{[K - (a/2)^2]}\Leftarrow$$

i.e. the frequency of oscillation increases as the gain *K* increases.

The *predominant time constant* is the time constant associated with the envelope of the response (*Figure 13.12*) which is given by $\exp(-\zeta\omega_n t)$ and thus the predominant time constant is $1/\zeta\omega_n$:

$$\frac{1}{\zeta\omega_n} = \Leftarrow\frac{1}{(a/2\sqrt{K})\sqrt{K}} = \Leftarrow\frac{2}{a}$$

Note that this time constant is unaffected by the gain *K* and is associated with the 'plant parameter *a*', which will normally be unalterable, and so other means must be found to alter the predominant time constant should this prove necessary.

The *settling time* t_s is variously defined as the time taken for the system to reach 2–5% (depending on specification) of its final steady state and is approximately equal to four times the predominant time constant.

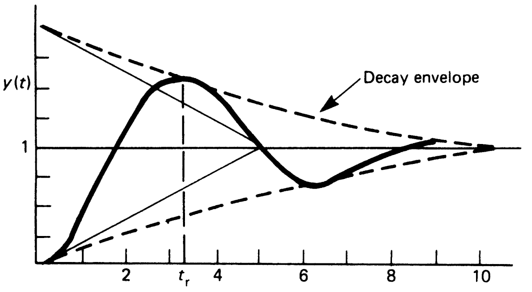


Figure 13.12 Step response of the system shown in *Figure 13.11*. The rise time t_r is the time taken to reach maximum overshoot. The predominant time constant is indicated by the tangents to the envelope curve

It should be obvious from the above that characteristics desired in plant dynamical behaviour may be conflicting (e.g. fast rise time with small overshoot) and it is up to the skill of the designer to achieve the best compromise. Overspecification can be expensive.

A number of the above items can be directly affected by the gain *K* and it may be that a suitable gain setting can be found to satisfy the design with no further attention. Unfortunately, the design is unlikely to be as simple as this, in view of the fact that the predominant time constant cannot be influenced by *K*. A particularly important method for influencing this term is the incorporation of so-called *velocity feedback*.

13.8.3 Velocity feedback

Given the prototype system shown in *Figure 13.11*, suppose that this is augmented by measuring the output $y(t)$, differentiating to form $\dot{y}(t)$, and feeding back in parallel with the normal feedback a signal proportional to $\dot{y}(t)$: say $T\dot{y}(t)$. The schematic of this arrangement is shown in *Figure 13.13*. Then, by simple manipulation, the modified transfer function becomes

$$W'\Leftarrow(s) = \Leftarrow\frac{K}{s^2 + (a + KT)s + K}$$

whence the modified predominant time constant is given by $2/(a + TK)$. The designer effectively has another string to his bow in that manipulation of *K* and *T* is normally very much in his command.

A similar effect may be obtained by the incorporation of a *derivative* term to act on the error signal (*Figure 13.14*) and in this case the transfer function becomes

$$W'\Leftarrow(s) = \Leftarrow\frac{K(1 + Ts)\Leftarrow}{s^2 + (a + KT)s + K}$$

It may be demonstrated that this derivative term when correctly adjusted can both stabilise the system and increase the speed of response. The control shown in *Figure 13.14* is referred to as *proportional-plus-derivative* control and is very important.

13.8.4 Incorporation of integral control

Mention has previously been made of the effect of using integrators within the loop to reduce steady-state errors; a particular study with reference to input/output effects was given. In this section consideration is given to the effects of disturbances injected into the loop, and we consider again

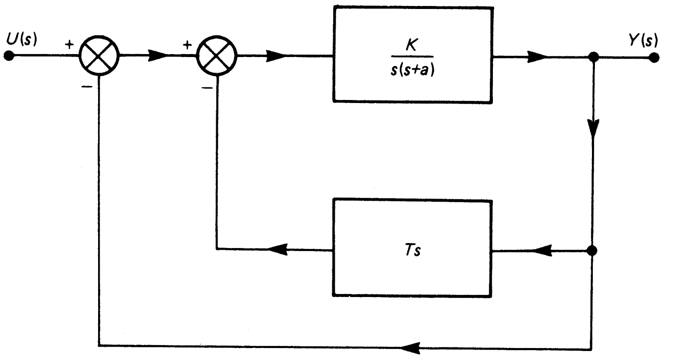


Figure 13.13 Schematic diagram showing the incorporation of velocity feedback

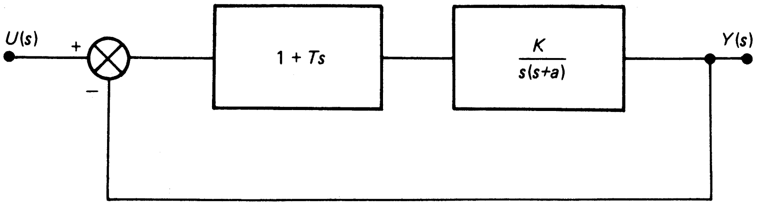


Figure 13.14 Schematic diagram of the proportional-plus-derivative control system

the simple second-order system shown in *Figure 13.11* but with a disturbance occurring between the amplifier and the plant dynamics. Appealing to superposition we can, without loss of generality, put $U(s) = 0$ and the transfer function between the output and the disturbance is then given by

$$\frac{Y(s)}{D(s)} = \frac{1}{s(s+a)+K}$$

Assuming that $d(t)$ is a unit step, $D(s) = 1/s$, and using the final-value theorem, $\lim_{t\to\infty} y(t)$ is obtained from

$$\lim_{t\to\infty} y(t) = \lim_{s\to\infty} \left[\frac{s}{[s(s+a)+K]s}\right] = \frac{1}{K}$$

and so the effect of this disturbance will always be present. By incorporating an integral control as shown in *Figure 13.15*, the output will, in the steady state, be unaffected by the disturbance, *viz*.

$$Y(s) = \frac{Ts}{Ts^2(s+a)+K(1+Ts)} D(s)$$

and so

$$y_{ss} \to 0$$

This controller is called a *proportional-plus-integral* controller.

An unfortunate side-effect of incorporating integral control is that it tends to destabilise the system, but this can be minimised by careful choice of T. In a particular case it might be that *proportional-plus-integral-plus-derivative (PID) control* may be called for, the amount of each particular control type being carefully proportioned.

In the foregoing discussions we have seen, albeit by using specific simple examples, how the behaviour of a plant might be modified by use of certain techniques. It is hoped that this will leave the reader with some sort of feeling for what might be done before embarking on more general tools, which tend to appear rather rarefied and isolated unless a basic physical feeling for system behaviour is present.

13.9 Root-locus method

The root locus is merely a graphical display of the *variation of the poles of the closed-loop system* when some parameter, often the gain, is varied. The method is useful since the loci may be obtained, at least approximately, by straightforward application of simple rules, and possible modification to reshape the locus can be assessed.

Considering once again the unity-feedback system with the open-loop transfer function $KG(s) = Kb(s)/a(s)$, where $b(s)$ and $a(s)$ represent mth- and nth-order polynomials, respectively, and $n > qn$, then the closed-loop transfer function may be written as

$$W(s) = \frac{KG(s)}{1+KG(s)} = \frac{Kb(s)}{a(s)+Kb(s)}$$

Note that the system is nth order and the zeros of the closed loop and the open loop are identical for unity feedback. The characteristic behaviour is determined by the roots of $1 + KG(s) = 0$ or $a(s) + Kb(s) = 0$. Thus, $G(s) = -(1/K)$ or $b(s)/a(s) = -(1/K)$.

Let s_r be a root of this equation; then

$$\text{mod}\left[\frac{b(s_r)}{a(s_r)}\right] = \frac{1}{K}$$

and

$$\text{phase}\left[\frac{b(s_r)}{a(s_r)}\right] = 180^\circ + n360^\circ$$

where n may take any integer value, including $n = 0$. Let $z_1, \ldots, z_m$ be the roots of the polynomial $b(s) = 0$, and $p_1, \ldots, p_n$ be the roots of the polynomial $a(s) = 0$. Then

$$b(s) = \prod_{i=1}^{m} (s - z_i)$$

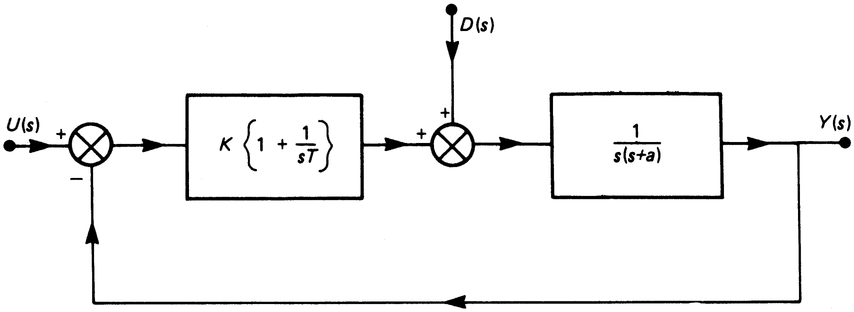


Figure 13.15 Schematic diagram of the proportional-plus-integral control system

and

$$a(s) = \prod_{i=1}^{n} (s - p_i)$$

Therefore

$$\frac{\prod_{i=1}^{m} |s_r - z_i|}{\prod_{i=1}^{n} |s_r - p_i|} = \frac{1}{K}, \textit{ the magnitude condition}$$

and

$$\sum_{i=1}^{m} \text{phase}(s_r - z_i) - \sum_{i=1}^{n} \text{phase}(s_r - p_i) = 180^\circ + n360^\circ, \quad \textit{the angle or phase condition}$$

Now, given a complex number p_j, the determination of the complex number $(s - p_j)$, where s is some point in the complex plane, is illustrated in *Figure 13.16*, where the $\text{mod}(s - p_j)$ and $\text{phase}(s - p_j)$ are also illustrated. The determination of the magnitudes and phase angles for all the factors in the transfer function, for any s, can therefore be done graphically.

The complete set of all values of s, constituting the root locus may be constructed using the angle condition alone; once found, the gain K giving particular values of s_r may be easily determined from the magnitude condition.

Example Suppose that $G(s) = K/[(s + a)(s + b)]$, then it is fairly quickly established that the only sets of points satisfying the angle condition

$$-\text{phase}(s_r + a) - \text{phase}(s_r + b) = 180 + n360^\circ$$

are on the line joining $-a$ to $-b$ and the perpendicular bisector of this line (*Figure 13.17*).

13.9.1 Rules for construction of the root locus

(1) The angle condition must be obeyed.
(2) The magnitude condition enables calibration of the locus to be carried out.
(3) The root locus on the real axis must be in sections to the left of an odd number of poles and zeros. This follows immediately from the angle condition.

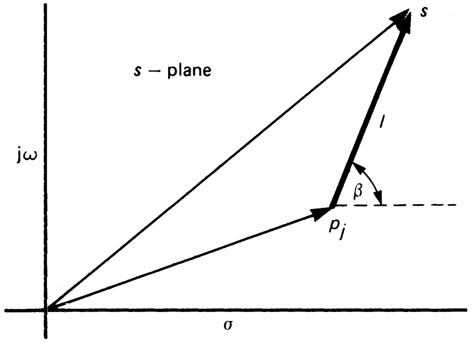


Figure 13.16 Representation of $(s - p_j)$ on the s plane ($l = |s - p_j|$; $\beta = \angle(s - p_j)$)

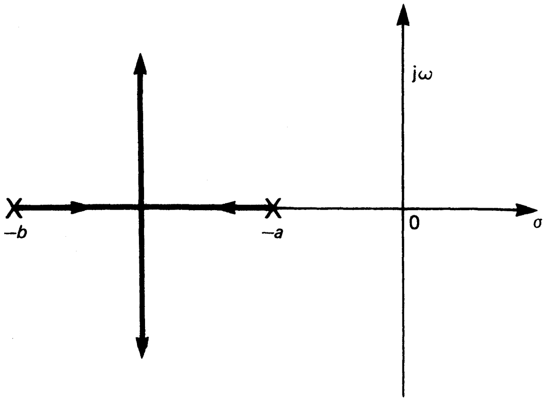


Figure 13.17 Root-locus diagram for $KG(s) = K/[(s + a)(s + b)]$

(4) The root locus must be symmetrical with respect to the horizontal real axis. This follows because complex roots must appear as complex conjugate pairs.
(5) Root loci always emanate from the poles of the open-loop transfer function where $K = 0$. Consider $a(s) + Kb(s) = 0$; then $a(s) = 0$ when $K = 0$ and the roots of this polynomial are the poles of the open-loop transfer function. Note that this implies that there will be n branches of the root locus.
(6) m of the branches will terminate at the zeros for $K \to \infty$. Consider $a(s) + Kb(s) = 0$, or $(1/K)a(s) + b(s) = 0$, whence as $K \to \infty$, $b(s) \to 0$ and, since this polynomial has m roots, these are where m of the branches terminate. The remaining $n - m$ branches terminate at infinity (in general, complex infinity).
(7) These $n - m$ branches go to infinity along asymptotes inclined at angles ϕ_i to the real axis, where

$$\phi_i = \frac{(2_i + 1)}{n - m} 180^\circ, \quad i = 0, 1, \ldots, (n - m - 1)$$

Consider a root s_r approaching infinity, $(s_r - a) \to s_r$ for all finite values of a. Thus, if ϕ_i is the phase s_r, then each pole and each zero term of the transfer function term will contribute approximately ϕ_i and $-\phi_i$, respectively. Thus,

$$\phi_i(n - m) = 180^\circ + i360^\circ$$

$$\phi_i = \frac{(2i + 1)}{n - m} 180^\circ, i = 0, 1, \ldots, (n - m - 1)$$

(8) The centre of these asymptotes is called the '*asymptote centre*' and is (with good accuracy) given by

$$\sigma_A = \left(\sum_{i=1}^{n} p_i - \sum_{j=1}^{m} z_i \right) \Big/ (n - m)$$

This can be shown by the following argument. For very large values of s we can consider that all the poles and zeros are situated at the point σ_A on the real axis. Then the characteristic equation (for large values of s) may be written as

$$1 + \frac{K}{(s + \sigma_A)^{n-m}} = 0$$

or approximately, by using the binomial theorem,

$$1 + \frac{K}{s^{n-m} + (n - m)s^{n-m-1}\sigma_A} = 0$$

Also, the characteristic equation may be written as

$$1 +\Leftarrow \frac{K\prod_{i=1}^{m}(s + z_i)\Leftarrow}{\prod_{i=1}^{m}(s + p_i)} = 0$$

Expanding this for the first two terms results in

$$1 + \frac{K}{s^{n-m} + (a_{n-1} - b_{m-1})s^{n-m-1}} = 0$$

where

$$b_{m-1} =\Leftarrow \sum_{i=1}^{m} z_i \quad \text{and} \quad a_{n-1} =\Leftarrow \sum_{i=1}^{n} p_j$$

whence

$$(a_{n-1} - b_{m-1}) = (n - m)\sigma_A$$

$$\sigma_A =\Leftarrow \frac{a_{n-1} - b_{m-1}}{n - m}$$

as required.

(9) When a locus breaks away from the real axis, it does so at the point where K is a local maximum. Consider the characteristic equation $1 + K[b(s)/a(s)] = 0$; then we can write $K = p(s)$, where $p(s) = -[a(s)/b(s)]$. Now, where two poles approach each other along the real axis they will both be real and become equal when K has the maximum value that will enable them both to be real and, of course, coincident. Thus, an evaluation of K around the breakaway point will rapidly reveal the breakaway point itself.

Example Draw the root locus for

$$KG(s) =\Leftarrow \frac{K(s + 1)\Leftarrow}{s(s + 2)(s + 3)\Leftarrow}$$

Procedure (*Figure 13.18*):

(1) Plot the poles of the open-loop system (i.e. at $s = 0$, $s = -2$, $s = -3$).
(2) Plot the zeros of the system (i.e. at $z = -1$).
(3) Determine the sections on the real axis at which closed-loop poles can exist. Obviously these are between 0 and −1 (this root travels along the real axis between these values as K goes from $0 \rightarrow \infty$), and between −2 and −3 (two roots are moving towards each other as K increases and, of course, will break away).
(4) Angle of asymptotes

$$\phi_1 = \frac{1}{2} + 180^\circ = 90^{\circ\Leftarrow}$$

$$\phi_2 = \frac{3}{2} \times 180^\circ = 270^{\circ\Leftarrow}$$

(5) Centroid σ_A is located at

$$\sigma_A =\Leftarrow \frac{-2 - 3 + 1}{2} = -2$$

(6) Breakaway point, σ_B

σ_B	−2.45	−2.465	−2.48
K	0.418	0.4185	0.418

(7) Modulus. For a typical root situated at, for example, point A, the gain is given by $K = l_2 l_3 l_4 / l_1$.

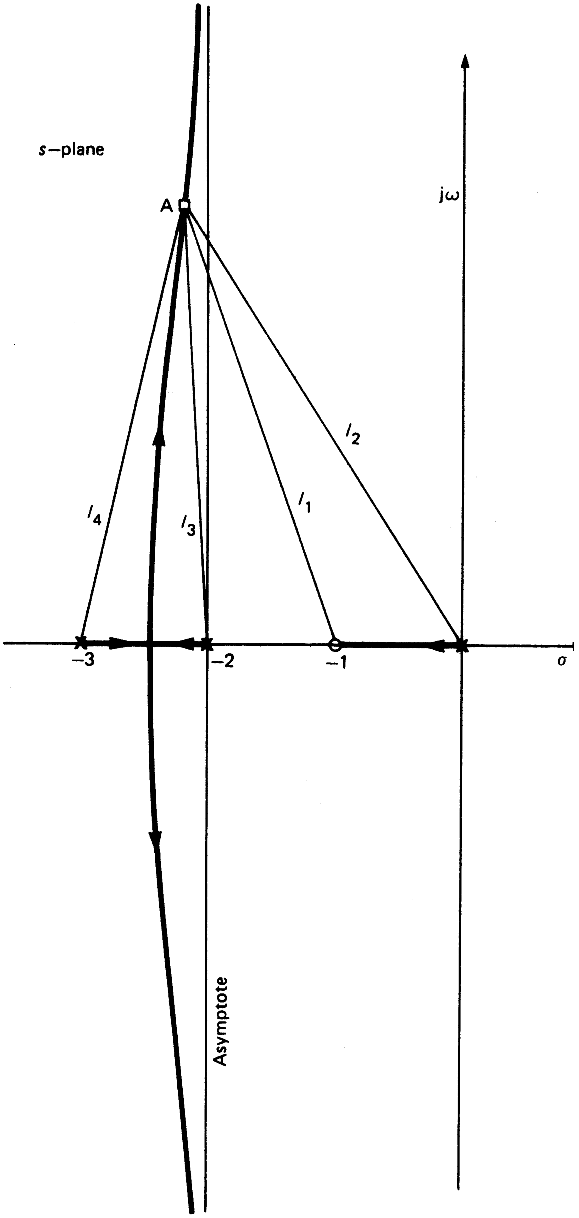


Figure 13.18 Root-locus construction for $KG(s) = [K(s+1)]/[s(s+2)(s+3)]$

After a little practice the root locus can be drawn very rapidly and compensators can be designed by pole-zero placement in strategic positions. A careful study of the examples given in the table will reveal the trends obtainable for various pole-zero placements.

13.10 Frequency-response methods

Frequency-response characterisation of systems has led to some of the most fruitful analysis and design methods in the whole of control system studies. Consider the situation of a linear, autonomous, stable system, having a transfer function $G(s)$, and being subjected to a unit-magnitude sinusoidal input signal of the form $\exp(j\omega t)$, starting at $t = 0$. The Laplace transformation of the resulting output of the system is

$$C(s) = G(s)/(s - j\omega)$$

and the time domain solution will be

$$c(t) = G(j\omega)\exp(j\omega t) + \left(\begin{array}{l}\text{Terms whose exponential terms}\\ \text{correspond to the roots}\\ \text{of the denominator of } G(s)\end{array}\right)$$

Since a stable system has been assumed, then the effects of the terms in the parentheses will decay away with time and so, after a sufficient lapse of time, the steady-state solution will be given by

$$c_{\mathrm{ss}}(t) = G(j\omega)\exp(j\omega t)$$

The term $G(j\omega)$, obtained by merely substituting $j\omega$ for s in the transfer function form, is termed the *frequency-response function*, and may be written

$$G(j\omega) = |G(j\omega)|\angle G(j\omega)$$

where $|G(j\omega)| = \text{mod } G(j\omega)$ and $\angle G(j\omega) = \text{phase } G(j\omega)$. This implies that the output of the system is also sinusoidal in magnitude $|G(j\omega)|$ with a phase-shift of $\angle G(j\omega)$ with reference to the input signal.

Example Consider the equation of motion

$$m\ddot{y} + bz + ky = f(t)$$

$$\frac{Y(s)}{F(s)} = G(s) = \frac{1}{ms^2 + bs + k}$$

If $f(t) = F_0 \exp(j\omega t)$, then

$$y_{\mathrm{ss}}(t) = \frac{F_0 \exp(j\omega t)}{(k - \omega^2 m) + j\omega b}$$

whence

$$y_{\mathrm{ss}}(t) = \frac{F_0 \exp[j(\omega t - \phi)]}{\sqrt{(k - \omega^2 m)^2 + (b\omega)^2}}$$

where $\phi = \arctan b\omega/(k - m\omega^2)$.

Within the area of frequency-response characterisation of systems three graphical techniques have been found to be particularly useful for examining systems and are easily seen to be related to each other. These techniques are based upon:

(1) The *Nyquist plot*, which is the locus of the frequency-response function plotted in the complex plane using ω as a parameter. It enables stability, in the closed-loop condition, to be assessed and also gives an indication of how the locus might be altered to improve the behaviour of the system.
(2) The *Bode diagram*, which comprises two plots, one showing the amplitude of the output frequency response (plotted in decibels) against the frequency ω (plotted logarithmically) and the other of phase angle θ of the output frequency response plotted against the same abscissa.
(3) The *Nichols chart*, a direct plot of amplitude of the frequency response (again in decibels) against the phase angle, with frequency ω as a parameter, but further enables the closed-loop frequency response to be read directly from the chart.

In each of these cases it is the *open-loop* steady-state frequency response, i.e. $G(j\omega)$, which is plotted on the diagrams.

13.10.1 Nyquist plot

The closed-loop transfer function is given by

$$\frac{C(s)}{R(s)} = \frac{G(s)}{1 + H(s)G(s)}$$

and the stability is determined by the location of the roots of $1 + H(s)G(s) = 0$, i.e. for stability no roots must have positive-real parts and so must not lie on the positive-real half of the complex plane. Assume that the open-loop transfer function $H(s)G(s)$ is stable and consider the contour C, the so-called 'Nyquist contour' shown in *Figure 13.19*, which consists of the imaginary axis plus a semicircle of large enough radius in the right half of the s plane such that any zeros of $1 + H(s)G(s)$ will be contained within this contour. This contour C_n is mapped via $1 + H(s)G(s)$ into another curve γ into the complex plane s'. It follows immediately from complex variable theory that the closed loop will be stable if the curve γ does not encircle the origin in the s' plane and unstable if it encircles the origin or passes through the origin. This result is the basis of the celebrated Nyquist stability criterion. It is rather more usual to map not $1 + H(s)G(s)$ but $H(s)G(s)$; in effect this is merely a change of origin from (0, 0) to (−1, 0), i.e. we consider curve γ_n'.

The statement of the stability criterion is that the closed-loop system will be stable if the mapping of the contour C_n by the open-loop frequency-response function $H(j\omega)G(j\omega)$ does not enclose the so-called critical point (−1, 0). Actually further simplification is normally possible, for:

(1) $|H(s)G(s)| \to 0$ as $|s| \to \infty$, so that the very large semicircular boundary maps to the origin in the s' plane.
(2) $H(-j\omega)G(-j\omega)$ is the complex conjugate of $H(j\omega)G(j\omega)$ and so the mapping of $H(-j\omega)G(-j\omega)$ is merely the mirror image of $H(j\omega)G(j\omega)$ in the real axis.
(3) Note: $H(j\omega)G(j\omega)$ is merely the frequency-response function of the open loop and may even be directly measurable from experiments. Normally we are mostly interested in how this behaves in the vicinity of the (−1, 0) point and, therefore, only a limited frequency range is required for assessment of stability.

The mathematical mapping ideas stated above are perhaps better appreciated practically by the so-called *left-hand rule* for an open-loop stable system, which reads as follows: if the open-loop sinusoidal response is traced out going from low frequencies towards higher frequencies, the closed loop will be stable if the critical point (−1, 0) lies on the left of all points on $H(j\omega)G(j\omega)$. If this plot passes through the critical point, or if the critical point lies on the right-hand side of $H(j\omega)G(j\omega)$, the closed loop will be unstable.

If the open loop has poles that actually lie on the imaginary axis, e.g. integrator $1/s$, then the contour is indented as shown in *Figure 13.20* and the above rule still applies to this modification.

13.10.1.1 Relative stability criteria

Obviously the closer the $H(j\omega)G(j\omega)$ locus approaches the critical point, the more critical is the consideration of stability, i.e. we have an indication of relative stability,

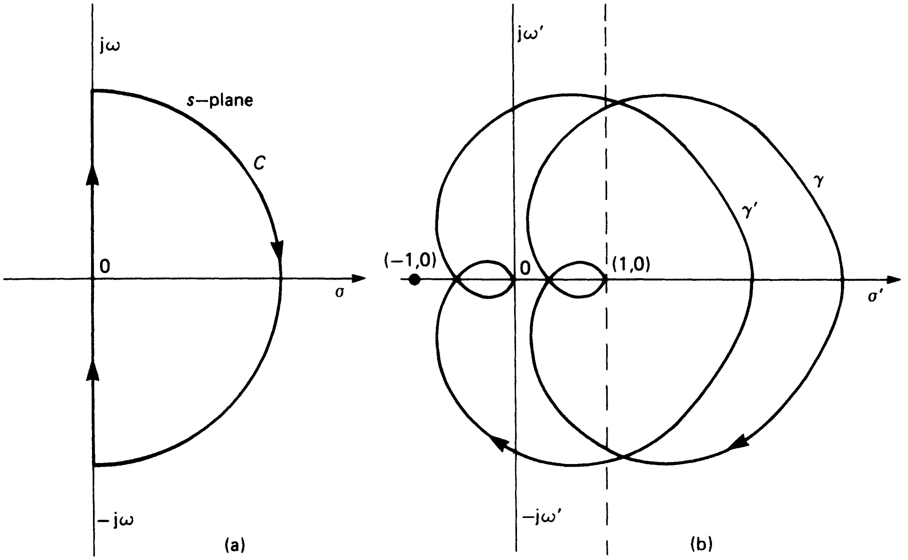


Figure 13.19 Illustration of Nyquist mapping: (a) mapping contour on the *s* plane; (b) resulting mapping of $1 + H(s)G(s) = 0$ and the shift of the origin

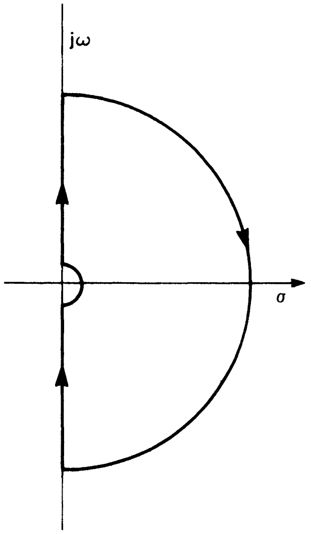


Figure 13.20 Modification of the mapping contour to account for poles appearing at the origin

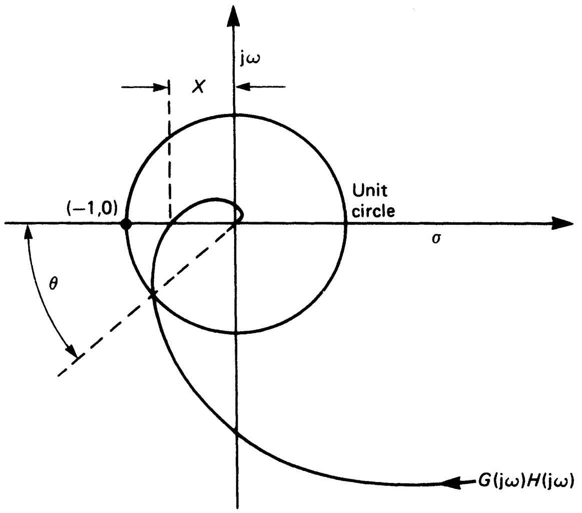


Figure 13.21 Illustration of the gain and phase margins. Gain margin $= 1/X$; phase margin $= 0$

given a measure by the gain and phase margins of the system.

If the modulus of $H(j\omega)G(j\omega) = X$ with a phase shift of 180°, then the *gain margin* is defined as

$$\text{Gain margin} = 1/X$$

The gain margin is usually specified in decibels, where we have

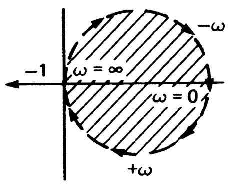
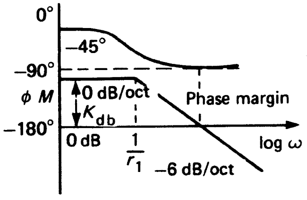
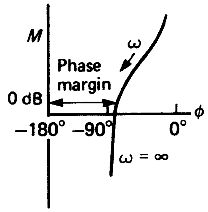
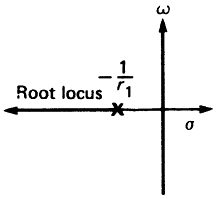
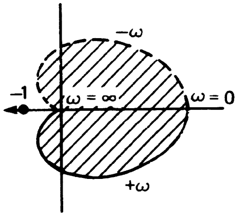
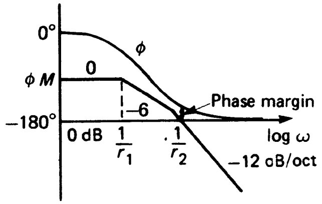
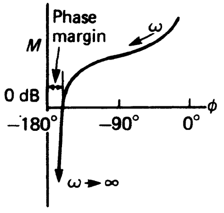
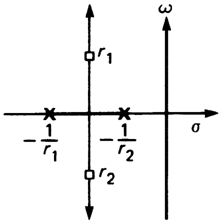
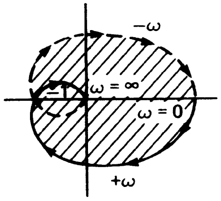
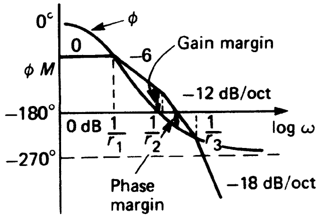
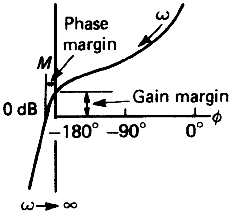
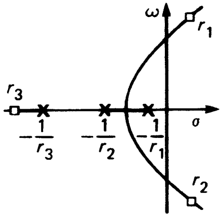
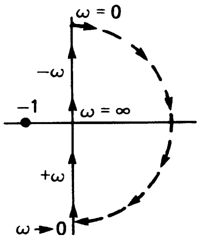
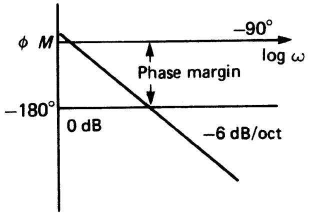
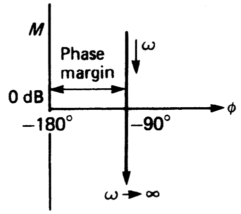
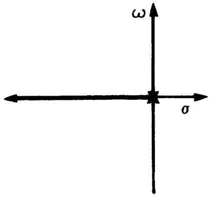
$$\text{Gain margin (dB)} = 20\log(1/X) = -20\log X$$

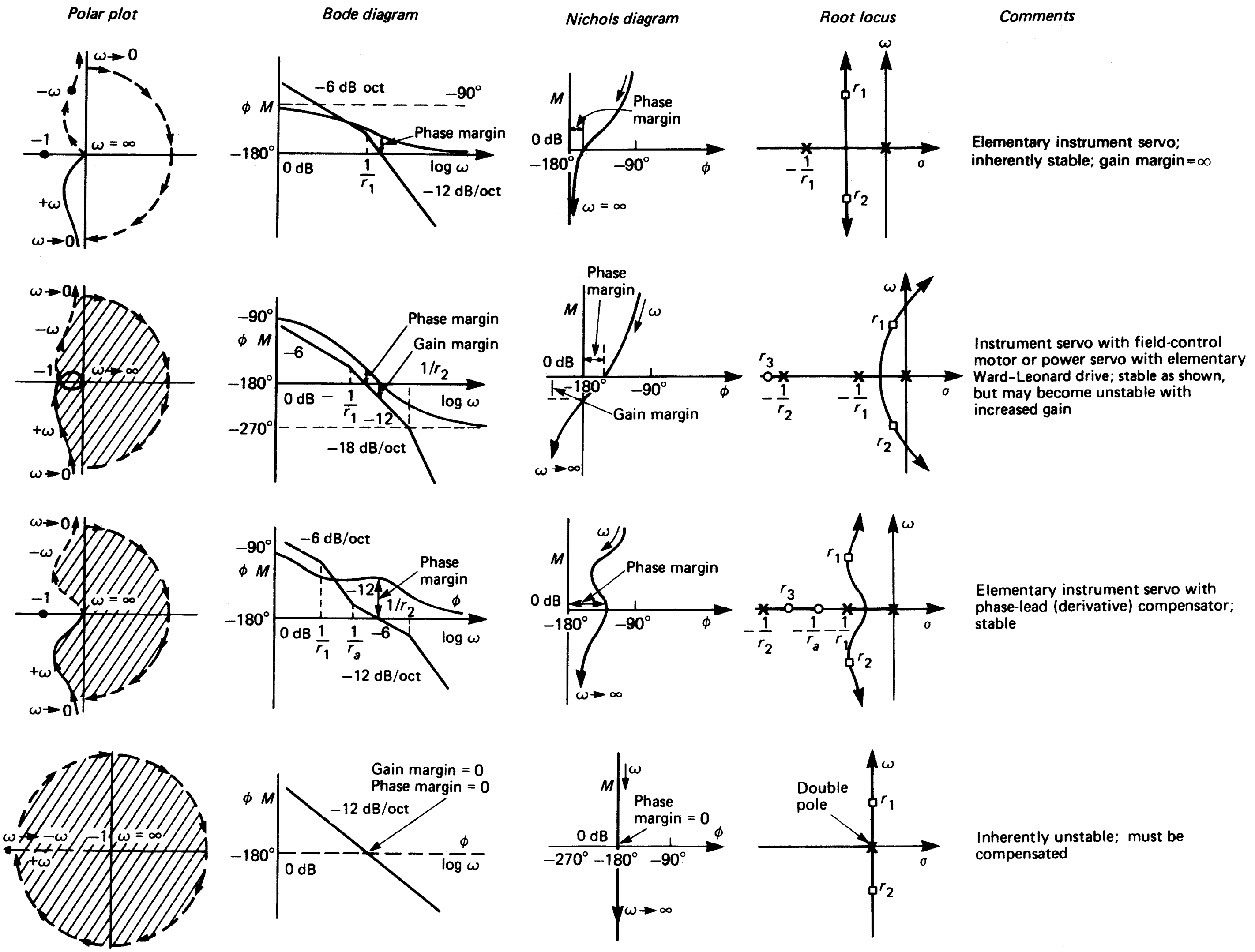
The *phase margin* is the angle which the line joining the origin to the point on the open-loop response locus corresponding to unit modulus of gain makes with the negative-real axis. These margins are probably best appreciated diagrammatically (*Figure 13.21*). They are useful, since a rough working rule for reasonable system damping and stability is to shape the locus so that a gain margin of at least 6 dB is achieved and a phase margin of about 40°.

Examples of the Nyquist plot are shown in *Figure 13.22*. Although from such plots the modifications necessary to achieve more satisfactory performance can be easily appreciated, precise compensation arrangements are not easily determined, since complex multiplication is involved and an appeal to the Bode diagram can be more valuable.

13.10.2 Bode diagram

As mentioned above, the Bode diagram is a logarithmic presentation of the frequency response and has the advantage

	G(s)	Polar plot	Bode diagram	Nichols diagram	Root locus	Comments
1.	$\frac{K}{sr_1 + 1}$		 0° −45° −90° ϕ M −180° 0 dB/oct K_{db} 0 dB $\frac{1}{r_1}$ Phase margin log ω −6 dB/oct	 M Phase margin ω 0 dB −180° −90° 0° ϕ ω = ∞	 ω Root locus $-\frac{1}{r_1}$ σ	Stable; gain margin = ∞
2.	$\frac{K}{(sr_1 + 1)(sr_2 + 1)}$		 0° ϕ 0 ϕ M −6 Phase margin −180° 0 dB $\frac{1}{r_1}$ $\frac{1}{r_2}$ log ω −12 dB/oct	 Phase margin M ω 0 dB −180° −90° 0° ϕ ω → ∞	 ω r_1 $-\frac{1}{r_1}$ $-\frac{1}{r_2}$ r_2 σ	Elementary regulator; stable; gain margin = ∞
3.	$\frac{K}{(sr_1 + 1)(sr_2 + 1)(sr_3 + 1)}$		 0° ϕ 0 −6 Gain margin ϕ M −12 dB/oct −180° 0 dB $\frac{1}{r_1}$ $\frac{1}{r_2}$ $\frac{1}{r_3}$ log ω −270° Phase margin −18 dB/oct	 Phase margin ω M 0 dB Gain margin −180° −90° 0° ϕ ω → ∞	 ω r_1 r_3 $-\frac{1}{r_3}$ $-\frac{1}{r_2}$ $-\frac{1}{r_1}$ σ r_2	Regulator with additional energy-storage component; unstable, but can be made stable by reducing gain
4.	$\frac{K}{s}$		 ϕ M −90° log ω Phase margin −180° 0 dB −6 dB/oct	 M Phase margin ω 0 dB −180° −90° ϕ ω → ∞		Ideal integrator; stable


Polar plot
Bode diagram
Nichols diagram
Root locus
ω→0
−ω
−1
ω = ∞
+ω
ω→∞
−6 dB oct
−90°
ϕ M
−180°
−270°
0 dB
Phase margin
Gain margin
log ω
−12 dB/oct
−18 dB/oct
−6 dB/oct
−6
−12
1/r₂
1/r₁
1/rₐ
ϕ
Gain margin = 0
Phase margin = 0
M
0 dB
−180°
−90°
ω = ∞
ω→∞
Phase margin = 0
−270°
ω
σ
r₁
r₂
r₃
−1/r₁
−1/r₂
−1/rₐ
Double pole

	G(s)	Comments
5.	$\frac{K}{s(sr_1+1)}$	Elementary instrument servo; inherently stable; gain margin = ∞
6.	$\frac{K}{s(sr_1+1)(sr_2+1)}$	Instrument servo with field-control motor or power servo with elementary Ward-Leonard drive; stable as shown, but may become unstable with increased gain
7.	$\frac{K(sr_a+1)}{s(sr_1+1)(sr_2+1)}$ $r_a < \frac{r_1r_2}{r_1+r_2}$	Elementary instrument servo with phase-lead (derivative) compensator; stable
8.	$\frac{K}{s^2}$	Inherently unstable; must be compensated

Figure 13.22 Transfer function plots for typical transfer functions

over the Nyquist diagram that individual factor terms may be added rather than multiplied, the diagram can usually be quickly sketched using asymptotic approximations and several decades of frequency may be easily considered.

Now suppose that

$$H(s)G(s) = H(s)G_1(s)G_2(s)G_3(s)\ldots$$

i.e. the composite transfer function may be thought of as being composed of a number of simpler transfer functions multiplied together, so

$$|H(j\omega)G(j\omega)| = |H(j\omega)||G_1(j\omega)||G_2(j\omega)||G_3(j\omega)|\ldots$$

$$20\log|H(j\omega)G(j\omega)| = 20\log|H(j\omega)| + 20\log|G_1(j\omega)| + 20\log|G_2(j\omega)| + 20\log|G_3(j\omega)| + \cdots$$

This is merely each individual factor (in decibels) being *added* algebraically to a grand total. Further,

$$\angle H(j\omega)G(j\omega) = \angle H(j\omega) + \angle G_1(j\omega) + \angle G_2(j\omega) + \angle G_3(j\omega) + \cdots$$

i.e. the individual phase shift at a particular frequency may be *added* algebraically to give the total phase shift.

It is possible to construct Bode diagrams from elemental terms including gain (K), differentiators and integrators (s and $1/s$), lead and lag terms ($(as+1)$ and $(1+as)^{-1}$), quadratic lead and lag terms ((bs^2+cs+1) and $(bs^2+cs+1)^{-1}$), and we consider the individual effects of their presence in a transfer function on the shape of the Bode diagram.

(*a*) *Gain term*, K The gain in decibels is simply 20 log K and is *frequency independent*; it merely raises (or lowers) the combined curve 20 log K dB.

(*b*) *Integrating term*, $1/s$ Now $|G(j\omega)| = 1/\omega$ and $\angle G(j\omega) = -90°$ (a constant) and so the gain in decibels is given by $20\log(1/\omega) = -20\log\omega$. On the Bode diagram this corresponds to a straight line with slope −20 dB/decade (or −6 dB/octave) of frequency and passes through 0 dB at $\omega = 1$ (see plot 4 in *Figure 13.22*).

(*c*) *Differentiating term, s* Now $|G(j\omega)| = \omega$ and $\angle G(j\omega) = 90°$ (a constant) and so the gain in decibels is given by $20\log\omega$. On the Bode diagram this corresponds to a straight line with slope 20 dB/decade of frequency and passes through 0 dB at $\omega = 1$.

(*d*) *First-order lag term*, $(1+s\tau)^{-1}$ The gain in decibels is given by

$$20\log\left(\frac{1}{1+\omega^2\tau^2}\right)^{1/2} = -10\log(1+\omega^2\tau^2)$$

and the phase angle is given by $\angle G(j\omega) = -\tan^{-1}\omega\tau$. When $\omega^2\tau^2$ is small compared with unity, the gain will be approximately 0 dB, and when $\omega^2\tau^2$ is large compared with unity, the gain will be $-20\log\omega\tau$. With logarithmic plotting this specifies a straight line having a slope of −20 dB/decade of frequency (6 dB/octave) intersecting the 0 dB line at $\omega = 1/\tau$. The actual gain at $\omega = 1/\tau$ is −3 dB and so the plot has the form shown in plot 1 of *Figure 13.22*. The frequency at which $\omega = 1/\tau$ is called the *corner or break frequency*. The two straight lines, i.e. those with 0 dB and −20 dB/decade, are called the 'asymptotic approximations' to the Bode plot. These approximations are often good enough for not too demanding design purposes.

The phase plot will lag a few degrees at low frequencies and fall to −90° at high frequency, passing through −45° at the break frequency.

(*e*) *First-order lead term*, $1+\omega\tau$ The lead term properties may be argued in a similar way to the above, but the gain, instead of falling, rises at high frequencies at 20 dB/decade and the phase, instead of lagging, leads by nearly 90° at high frequencies.

(*f*) *Quadratic-lag term*, $1/(1+2\tau\zeta s+\tau^2s^2)$ The gain for the quadratic lag is given by

$$-10\log\left[\left(1-\left(\frac{\omega}{\omega_n}\right)^2\right)^2 + \left(2\zeta\frac{\omega}{\omega_n}\right)^2\right]$$

and the phase angle by

$$\angle G(j\omega) = \arctan\left[-\frac{2\zeta(\omega/\omega_n)}{1-(\omega/\omega_n)^2}\right]$$

where $\tau = 1/\omega_n$. At low frequencies the gain is approximately 0 dB and at high frequencies falls at −40 dB/decade. At the break frequency $\omega = 1/\tau$ the actual gain is $20\log(1/2\zeta)$. For low damping (say $\zeta < 0.5$) an asymptotic plot can be in considerable error around the break frequency and more careful evaluation may be required around this frequency. The phase goes from minus a few degrees at low frequencies towards −180° at high frequencies, being −90° at $\omega = 1/\tau$.

(*g*) *Quadratic lead term*, $1+2\tau\zeta s+\tau^2s^2$ This is argued in a similar way to the lag term with the gain curves inscribed and the phase going from plus a few degrees to 180° in this case.

Example Plot the Bode diagram of the open-loop frequency-response function

$$G(j\omega) = \frac{10(1+j\omega)}{j\omega(j\omega+2)(j\omega+3)}$$

and determine the gain and phase margins (see *Figure 13.23*). Note: *Figure 13.22* shows a large number of examples and also illustrates the gain and phase margins.

13.10.3 Nichols chart

This is a graph with the open-loop gain in decibels as co-ordinate and the phase as abscissa. The open-loop frequency response is for a particular system and is plotted with frequency ω as parameter. Now the closed-loop frequency response is given by

$$W(j\omega) = \frac{G(j\omega)}{1+G(j\omega)}$$

and corresponding lines of constant magnitude and constant phase of $W(j\omega)$ are plotted on the Nichols chart as shown in *Figure 13.24*.

When the open-loop frequency response of a system has been plotted on such a chart, the closed-loop frequency response may be immediately deduced from the contours of $W(j\omega)$.

13.11 State-space description

Usually in engineering, when analysing time-varying physical systems, the resulting mathematical models are in differential equation form. Indeed, the introduction of the Laplace transformation, and similar techniques, leading to the whole edifice of transfer-function-type analysis and design methods are, essentially, techniques for solving, or manipulating to

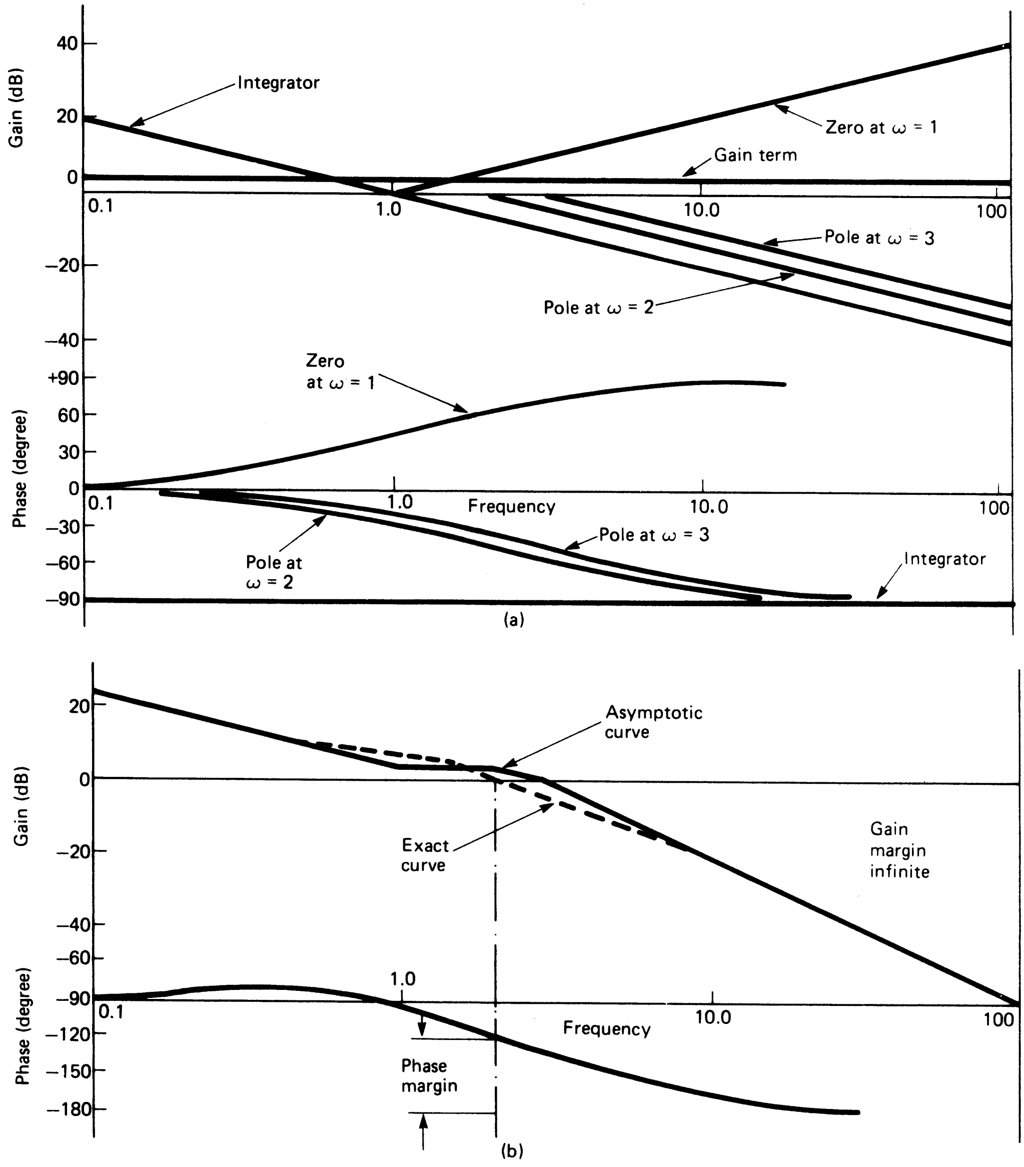


Figure 13.23 (a) Gain and phase curves for individual factors (see *Figure 13.18*); (b) Composite gain and phase curves. Note that the phase margin is about 60°, and the gain margin is infinite because the phase tends asymptotically to −180°⇐

advantage, differential equation models. In the state-space description of systems, which is the concern of this section, the models are left in the differential equation form, but rearranged into the form of a set of first-order simultaneous differential equations. There is nothing unique to systems analysis in doing this, since this is precisely the required form that differential equations are placed in if they are to be integrated by means of many common numerical techniques, e.g. the Runge–Kutta methods. Most of the interest in the state-space form of studying control systems stems from the 1950s, and intensive research work in this area has continued since then; however, much of it is of a highly theoretical nature. It is arguable that these methods have yet to fulfill the hopes and aspirations of the research workers who developed them. The early expectation was that they would quickly supersede classical techniques. This has been very far from true, but they do have a part to play, particularly if there are good mathematical models of the plant available and the real plant is well instrumentated.

Consider a system governed by the *n*th order linear constant-coefficient differential equation

$$\frac{d^n y}{dt^n} + \cdots \Leftarrow a_1 \frac{dy}{dt} + a_0 y = \Leftarrow ku(t) \Leftarrow$$

where y is the dependent variable and $u(t)$ is a time-variable forcing function.

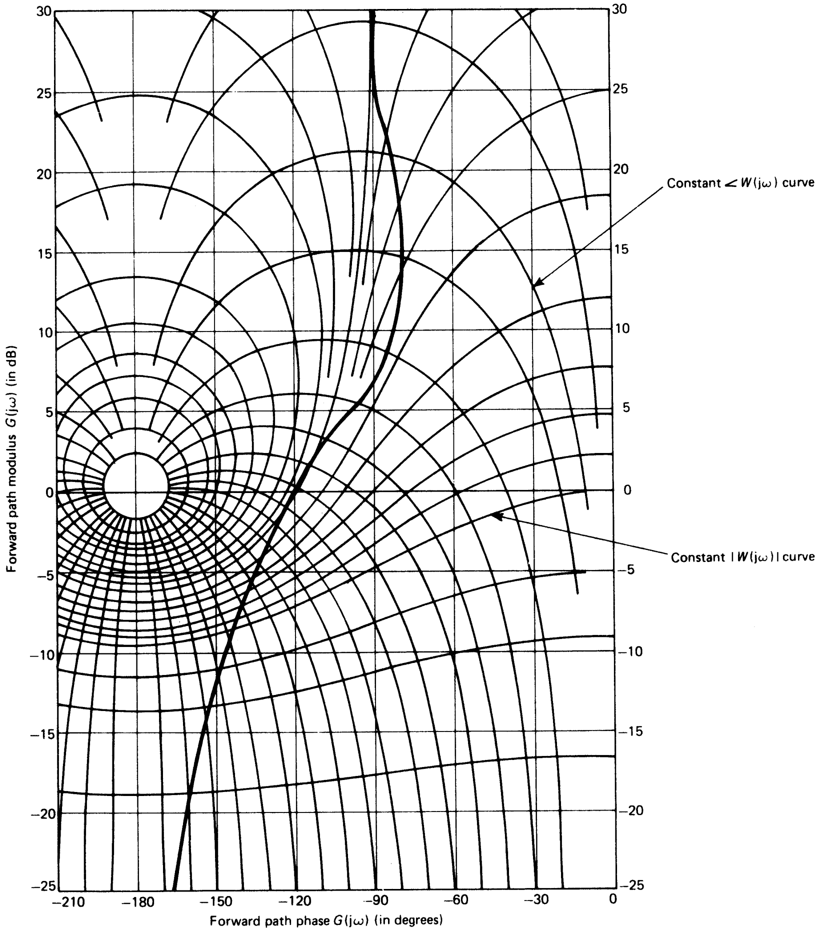


Figure 13.24 Nichols chart and plot of the system shown in *Figure 13.23*. Orthogonal families of curves represent constant $W(j_\omega)$ and constant $\angle W(j_\omega)$

Let $y = x_1$, then

$$\frac{dy}{dt} = \Leftarrow \frac{dx_1}{dt} = \Leftarrow x_2$$

say, and

$$\frac{d^2y}{dt^2} = \frac{dx_2}{dt} = x_3$$

$$\frac{d^{n-1}y}{dt^{n-1}} = \frac{dx_{n-1}}{dt} = \Leftarrow x_n$$

From the governing differential equation we can write

$$\frac{d^ny}{dt^n} = \Leftarrow \frac{dx_n}{dt} = -a_0x_1 - a_1x_2 \cdots - a_{n-1}x_n + ku(t) \Leftarrow$$

i.e. the *n*th order differential equation has been transformed into *n* first-order equations. These can be arranged into matrix form:

$$\begin{bmatrix} \dot{x}_1 \\ \dot{x}_2 \\ \vdots \\ \dot{x}_{n-1} \\ \dot{x}_n \end{bmatrix} = \Leftarrow \begin{bmatrix} 0 & 1 & 0 & \ldots & \varsigma \\ 0 & 0 & 1 & 0 \ldots & \varsigma \\ & & & 0 & 1 \\ -a_0 & -a_1 & \ldots & & -a_{n-1} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix} + \Leftarrow \begin{bmatrix} 0 \\ 0 \\ 0 \\ k \end{bmatrix} u(t) \Leftarrow \qquad (13.3) \Leftarrow$$

which may be written in matrix notation as

$$\dot{\mathbf{x}} = \Leftarrow \mathbf{A}\mathbf{x} + \mathbf{b}u(t) \Leftarrow$$

where $\mathbf{x} = [x_1, \ldots, x_n]^T$ and is called the 'state vector', $\mathbf{b} = [0, 0, \ldots, k]^T$ and $\mathbf{A}$ is the $n \times n$ matrix pre-multiplying $\mathbf{x}$ on the right-hand side of Equation (13.3).

It can be shown that the eigenvalues of **A** are equal to the characteristic roots of the governing differential equation which are also equal to the poles of the transfer function $Y(s)/U(s)$. Thus the time behaviour of the matrix model is essentially governed by the position of the eigenvalues of the **A** matrix (in the complex plane) in precisely the same manner as the poles govern the transfer function behaviour. Hence, if these eigenvalues do not lie in acceptable positions in this plane, the design process is to somehow modify the **A** matrix so that the corresponding eigenvalues do have acceptable positions (cf. the placement of closed-loop poles in the *s* plane).

Example Consider a system governed by the general second-order linear differential equation

$$\frac{d^2y}{dt^2} + 2\zeta\omega_n\frac{dy}{dt} + \omega_n^2y = \Leftarrow \omega_n^2u$$

Let $y = x_1$, then

$$\frac{\mathrm{d}y}{\mathrm{d}t} = \frac{\mathrm{d}x_1}{\mathrm{d}t} = x_2$$

and so

$$\frac{\mathrm{d}x_2}{\mathrm{d}t} = -\omega_n^2 x_1 - 2\zeta\omega_n x_2 + \omega_n^2 u$$

or

$$\begin{bmatrix}\dot{x}_1\\ \dot{x}_2\end{bmatrix} = \begin{bmatrix}0 & 1\\ -\omega_n^2 & -2\zeta\omega_n\end{bmatrix}\begin{bmatrix}x_1\\ x_2\end{bmatrix} + \begin{bmatrix}0\\ \omega_n^2\end{bmatrix}u \qquad (13.4)$$

The eigenvalues of the **A** matrix are given by the solution to the equation $\lambda^2 + 2\zeta\omega_n\lambda + \omega_n^2 = 0$, i.e.

$$\lambda_{1,2} = -\zeta\omega_n \pm \omega_n\sqrt{\zeta^2 - 1}$$

Now let $u = r - k_1x_1 - k_2x_2$ where r is an arbitrary, or reference, value or input, and k_1 and k_2 are constants. Note this is a feedback arrangement, since u has become a linear function of the state variables which, in a dynamic system, might be position and velocity. Substituting for u in equation (13.3), gives

$$\begin{bmatrix}\dot{x}_1\\ \dot{x}_2\end{bmatrix} = \begin{bmatrix}0 & 1\\ -\omega_n^2(1+k_1) & -\omega_n(2\zeta + \omega_n k_2)\end{bmatrix}\begin{bmatrix}x_1\\ x_2\end{bmatrix} + \begin{bmatrix}0\\ \omega_n^2\end{bmatrix}r$$

The eigenvalues of the **A** matrix are given by the roots of $\lambda^2 + (2\zeta\omega_n + \omega_n^2 k_2)\lambda + \omega_n^2(1 + k_1) = 0$ and, by choosing suitable values for k_1 and k_2 (the feedback factors), the eigenvalues can be made to lie in acceptable positions in the complex plane. Note that, in this case, k_1 alters the effective undamped natural frequency, and k_2 alters the effective damping of the second-order system.

If the governing differential equation has derivatives on the right-hand side, then the derivation of the first-order set involves a complication. Overcoming this is easily illustrated by an example. Suppose that

$$\frac{\mathrm{d}^2y}{\mathrm{d}t^2} + a_1\frac{\mathrm{d}y}{\mathrm{d}t} + a_2y = b_0u + b_1\frac{\mathrm{d}u}{\mathrm{d}t}$$

Let $y = x_1$, and

$$\frac{\mathrm{d}y}{\mathrm{d}t} = \frac{\mathrm{d}x_1}{\mathrm{d}t} = x_2 + b_1u$$

then

$$\frac{\mathrm{d}^2y}{\mathrm{d}t^2} = \frac{\mathrm{d}x_2}{\mathrm{d}t} + b_1\frac{\mathrm{d}u}{\mathrm{d}t} = -a_1(x_2 + b_1u) - a_0x + b_0u + b_1\frac{\mathrm{d}u}{\mathrm{d}t}$$

$$\begin{bmatrix}\dot{x}_1\\ \dot{x}_2\end{bmatrix} = \begin{bmatrix}0 & 1\\ -a_0 & -a_1\end{bmatrix}\begin{bmatrix}x_1\\ x_2\end{bmatrix} + \begin{bmatrix}b_1\\ b_0 - a_1b_1\end{bmatrix}u$$

Note that care may be necessary in interpreting the x derivatives in a physical sense.

The state-space description is also a convenient way of dealing with multi-input/multi-output systems. A simple example is shown in *Figure 13.25*, where $U_1(s)$ and $U_2(s)$ are the inputs and $Y_1(s)$ and $Y_2(s)$ are the corresponding outputs, and so

$$Y_1(s) = \frac{k_1}{s + a_1}U_1(s) + \frac{k_3}{s + a_3}U_2(s)$$

and

$$Y_2(s) = \frac{k_2}{s + a_1}U_1(s) + \frac{k_4}{s + a_4}U_2(s)$$

The first of these two equations may be written as

$$[s^2 + s(a_1 + a_2) + a_1a_3]Y_1(s) = k_1U_1(s) + k_3U_2(s)$$

or

$$\frac{\mathrm{d}^2y_1}{\mathrm{d}t^2} + (a_1 + a_2)\frac{\mathrm{d}y_1}{\mathrm{d}t} + a_1a_3y_1 = k_1u_1 + k_zu_2$$

let $y_1 = x_1$, then

$$\frac{\mathrm{d}y_1}{\mathrm{d}t} = \frac{\mathrm{d}x_1}{\mathrm{d}t} = x_2$$

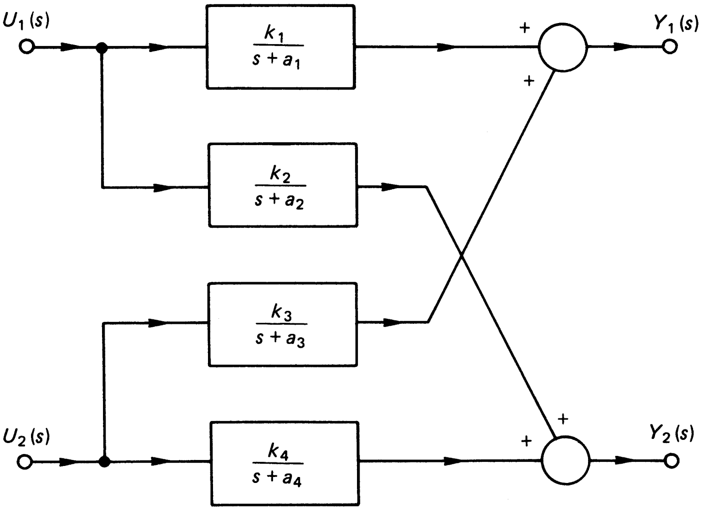


Figure 13.25 Block diagram of a two-input/two-output multi-variable system

and

$$\frac{d^2y_1}{dt^2} = \frac{dx_2}{dt} = -(a_1 + a_2)x_2 - a_1a_3x_1 + k_1u_1 + k_3u_2$$

Similarly, for the second of the two equations, writing

$$y_2 = x_3 \text{ and } \frac{dy_2}{dt} = \frac{dx_3}{dt} = x_4$$

leads to

$$\frac{d^2y_2}{dt^2} = \frac{dx_4}{dt} = -(a_2 + a_4)x_4 - a_2a_4x_3 + k_2u_1 + k_4u_2$$

Whence the entire set may be written as

$$\begin{bmatrix} \dot{x}_1 \\ \dot{x}_2 \\ \dot{x}_3 \\ \dot{x}_4 \end{bmatrix} = \begin{bmatrix} 0 & 1 & 0 & 0 \\ -a_1a_3 & -(a_1 - a_3) & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & -a_2a_4 & -(a_2 + a_4) \end{bmatrix} \begin{bmatrix} \dot{x}_1 \\ \dot{x}_2 \\ \dot{x}_3 \\ \dot{x}_4 \end{bmatrix} + \begin{bmatrix} 0 & 0 \\ k_1 & k_3 \\ 0 & 0 \\ k_2 & k_4 \end{bmatrix} \begin{bmatrix} u_1 \\ u_2 \end{bmatrix}$$

The problem is now how to specify u_1 and u_2 (e.g. a linear combination of state variables similar to the simple second-order system above), so as to make the plant behave in an acceptable manner. It must be pointed out that the theory of linear matrix-differential equations is an extremely well developed mathematical topic and has been extensively plundered in the development of state-space methods. Thus a vast literature exists, and this is not confined to linear systems. Such work has, among other things, discovered a number of fundamental properties of systems (for example, controllability and observability); these are well beyond the scope of the present treatment. The treatment given here is a very short introduction to the fundamental ideas of the state-space description.

13.12 Sampled-data systems

Sampled-data systems are ones in which signals within the control-loop are sampled at one or more places. Some sort of sampling action may be inherent in the very mode of operation of some of the very components comprising the plant, e.g. thyristor systems, pulsed-radar systems and reciprocating internal combustion engines. Moreover, sampling is inevitable if a digital computer is used to implement the control laws, and/or used in condition monitoring operations. Nowadays, digital computers are routinely used in control-system operation for reasons of cheapness and versatility, e.g. they may be used not only to implement the control laws, which can be changed by software alterations alone, but also for sequencing control and interlocking in, say, the start up and safe operation of complex plant. Whatever the cause, sampling complicates the mathematical analysis and design of systems.

Normally most of the components, comprising the system to be controlled, will act in a continuous (analogue) manner, and hence their associated signals will be continuous. With the introduction of a digital computer it is necessary to digitise the signal, by an analogue-to-digital converter before the signal enters the computer. The computer processes this digital sequence, and then outputs another digital sequence which, in turn, passes to a digital-to-analogue converter. This process is shown schematically in *Figure 13.26*.

In this diagram the sampling process is represented by the periodic switch (period *T*), which at each sampling instant is closed for what is regarded as an infinitesimal time. The digital-to-analogue process is represented by the hold block. Thus the complete system is a hybrid one, made up of an interconnection of continuous and discrete devices. The most obvious way of representing the system mathematically is by a mixed difference-differential equation set. However, this makes a detailed analysis of the complete system difficult.

Fortunately, provided the investigator or system designer is prepared to accept knowledge of the system's behaviour at the instants of sampling only, a comparatively simple approach having great similarity to that employed for wholly continuous systems is available. At least for early stages of the analysis or design proposal, the added complications involved in this process are fairly minor. Further, the seemingly severe restriction of knowing the system's behaviour at the instants of sampling only is normally quite acceptable; for example, the time constants associated with the plant will generally be much longer than the periodic sampling time, so the plant effectively does not change its state significantly in the periodic time. The sampling time period is a parameter which often can be chosen by the designer, who will want sampling to be fast enough to avoid aliasing problems; however, the shorter the sampling period the less time the computer has available for other loops. Suffice it to say that the selection of the sampling period is normally an important matter.

If we take a continuous signal $y(t)$, say, and by the periodic sampling process convert it into a sequence of values $y(n)$, where n represents the nth sampling period, then the sequence $y(n)$ becomes the mathematical entity we manipulate, and the values of $y(t)$ between these samples will not be known. However, if at an early stage it is essential to know the inter-sample behaviour of the system with some accuracy, then advance techniques are available for this purpose.[1] In addition, it is now fairly routine to simulate control system behaviour before implementation, and a

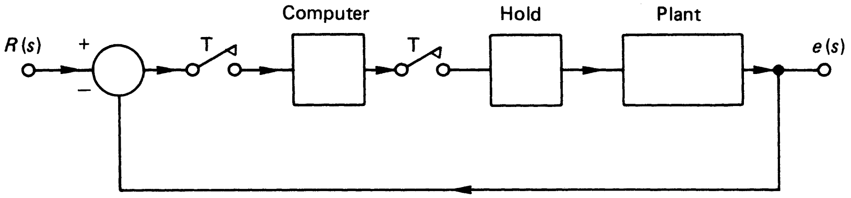


Figure 13.26 General arrangement of a sampled-data system

good simulation package should be capable of illustrating the inter-sample behaviour.

We need techniques for mathematically manipulating sequences, and these are discussed in the following section.

13.13 Some necessary mathematical preliminaries

13.13.1 The *z* transformation

This transformation plays the equivalent role in sampled-data system studies as the Laplace transformation does in the case of continuous systems; indeed, these two transformations are mathematically related to each other. It is demonstrated below that the behaviour of sampled-data systems at the sampling instant is governed mathematically by difference equations, e.g. a linear system might be governed by the equation

$$y(n) + a_1 y(n-1) + a_2 y(n-2) = b_1 x(n) + b_2 x(n-1) \Leftarrow$$

where, in the case of $y(n)$, the value of a variable at instant n is in fact dependent on a linear combination of its previous two values and the current and previous values of an independent (forcing variable) $x(n)$. In a similar way to using the Laplace transformation to convert linear differential equations to transfer-function form, the z transformation is used to convert linear difference equations into the so-called ‘pulse transfer-function form’. The definition of the z transformation of a sequence $y(n)$, $n = 0, 1, 2, \ldots$, is

$$Z[y(n)] = Y(z) = \sum_{n=0}^{\infty} y(n) z^{-n}$$

The z transformations of commonly occurring sequences are listed in *Table 13.1*, and a simple example will illustrate how such transformations may be found.

Suppose $y(n) = nT(n = 0, 1, 2, \ldots)$ such a sequence would be obtained by sampling the continuous ramp function $y(t) = t$, at intervals of time T. Then, by definition,

$$\begin{aligned} Z[y(n)] &= Y(z) = 0 + Tz^{-1} + 2Tz^{-2} + \cdots \Leftarrow \\ &= T(z^{-1} + 2z^{-2} + \cdots) \Leftarrow \\ &= \frac{Tz}{(z-1)^2} \end{aligned}$$

It can also be shown that

$$Z[y(n-1)] = z^{-1} Z[y(n)] = z^{-1} Y(z) \Leftarrow$$

and

$$Z[y(n-2)] = z^{-2} Z[y(n)] = z^{-2} Y(z) \Leftarrow$$

Then, applying this to the difference equation above, we have

$$Y(z) = -(a_1 z^{-1} + a_2 z^{-2}) Y(z) + (b_0 + b_1 z^{-1}) X(z) \Leftarrow$$

or

$$Y(z) = \Leftarrow \frac{(b_0 + b_1 z^{-1}) X(z) \Leftarrow}{(1 + a_1 z^{-1} + a_2 z^{-2}) \Leftarrow}$$

So that, if $x(n)$ or $X(z)$ is given, $Y(z)$ can be rearranged into partial fraction form, and $y(n)$ determined from the table. For example, suppose that

$$Y(z) = \Leftarrow \frac{z(z-0.25) \Leftarrow}{(z-1)(z-0.5) \Leftarrow}$$

then

$$\frac{Y(z) \Leftarrow}{z} = \Leftarrow \frac{1.5}{z-1} - \Leftarrow \frac{0.5}{z-0.5}$$

or

$$Y(z) = \Leftarrow \frac{1.5z}{z-1} - \Leftarrow \frac{0.5z}{z-0.5}$$

Whence, from the tables we see that

$$y(n) = 1.5 - 0.5 \exp(-0.60n) \Leftarrow$$

The process of dividing $Y(z)$ by z before taking partial fractions is important, as most tabulated values of the transformation have z as a factor in the numerator, and the partial function expansion process needs the order of the denominator to exceed that of the numerator.

An alternative method of approaching the z transform is to assume that the sequence to be transformed is a direct consequence of sampling a continuous signal using an impulse modulator. Thus a given signal $y(t)$ is sampled with periodic time T, to give the assumed signal $y^*(t)$, where

$$y^*(t) = y(\mathrm{o})\delta(t) + y(T)\delta(t-T) + y(2T)\delta(t-2T) + \cdots \Leftarrow$$

where $\delta(t)$ is the delta function.

Taking the Laplace transformation of $y^*(t)$ gives the series

$$\mathscr{L}[y^*(t)] = y(\mathrm{o}) + y(T)\mathrm{e}^{-sT} + y(2T)\mathrm{e}^{-2sT} + \cdots \Leftarrow$$

On making the substitution $\mathrm{e}^{sT} = z$, then the resulting series is identical to that obtained by taking the z transformation of the sequence $y(n)$. For convenience, we often write $Y(z) = Z[y^*(t)]$.

$z = \mathrm{e}^{sT}$ may be regarded as constituting a transformation of points in an s plane to those in a z plane, and this has exceedingly important consequences. If, for example, we map lines representing constant damping ζ, and constant natural frequency ω_n, for a system represented in an s plane onto a z plane, we obtain *Figure 13.27*.

There are important results to be noted from this diagram.

(1) The stability boundary in the s plane (i.e. the imaginary axis) transforms into the unit circle $|z| = 1$ in the z plane.
(2) Points in the z plane indicate responses relative to the periodic sampling time T.
(3) The negative real axis of the z plane always represents half the sampling frequency ω_s, where $\omega_s = 2\pi/T$.
(4) Vertical lines (i.e. those with constant real parts) in the left-half plane of the s plane map into circles *within* the unit circle in the z plane.
(5) Horizontal lines (i.e. lines of constant frequency) in the s plane map into radial lines in the z planes.
(6) The mapping is not one-to-one; and frequencies greater than $\omega_s/2$ will coincide on the z plane with corresponding points below this frequency. Effectively this is a consequence of the Nyquist sampling theorem which states, essentially, that faithful reconstruction of a sampled signal cannot be achieved if the original continuous signal contained frequencies greater than one-half the sampling frequency.

A vitally important point to note is that *all the roots* of the denominator of a pulse transfer function of a system must fall *within* the unit circle, on the z plane, if the system is to be stable; this follows from (1) above.

13.14 Sampler and zero-order hold

The sampler produces a series of discrete values at the sampling instant. Although in theory these samples exist for

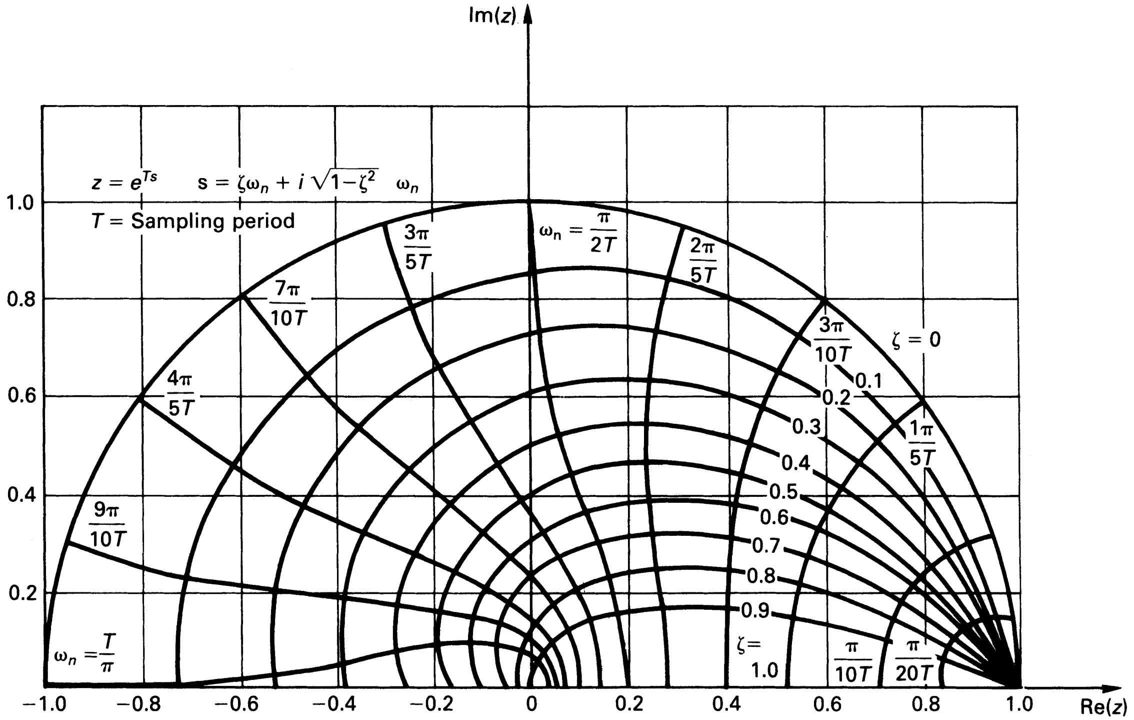


Figure 13.27 Natural frequency and damping loci in the *z* plane. The lower half is the mirror image of the half shown. (Reproduced from Franklin *et al.*,[2] courtesy of Addison-Wesley)

zero time, in practice they can be taken into the digital computer and processed. The output from the digital computer will be a sequence of samples with, again in theory, each sample existing for zero time. However, it is necessary to have a continuous signal constructed from this output, and this is normally done using a *zero-order hold.* This device has the property that, as each sample (which may be regarded as a delta function) is presented to its input, it presents the strength of the delta function at its output until the next sample arrives, and then changes its output to correspond to this latest value, and so on.

This is illustrated diagrammatically in *Figure 13.28*. Thus a unit delta function $\delta(t)$ arriving produces a positive unit-value step at the output at time t. At time $t = \Leftarrow T$, we may regard a negative unity-value step being superimposed on the output. Since the transfer function of a system may be regarded as the Laplace transformation of the response of that system to a delta function, the zero-order hold has the transfer function

$$\frac{1}{s}[1 - \Leftarrow e^{-sT}] \Leftarrow$$

13.15 Block diagrams

In a similar way to their use in continuous-control-system studies, block diagrams are used in sampled-data-system studies. It is convenient to represent individual pulse transfer functions in individual boxes. The boxes are joined together by lines representing their *z* transformed input/output sequences to form the complete block diagrams. The manipulation of

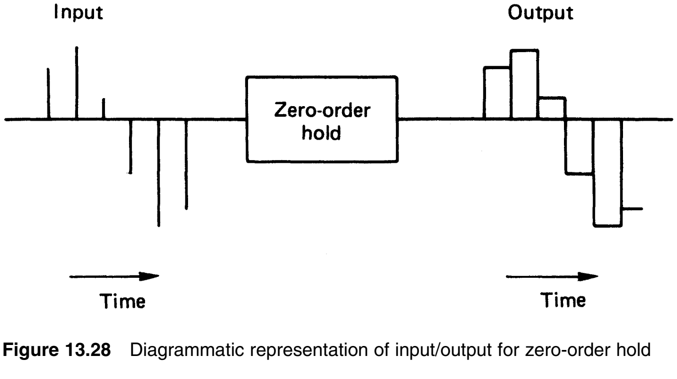


Figure 13.28 Diagrammatic representation of input/output for zero-order hold

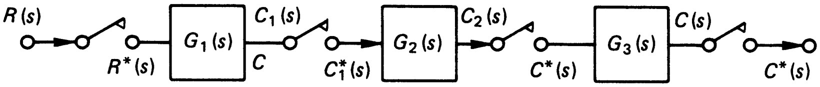


Figure 13.29 Cascade transfer functions with sampling between connections

the block diagrams may be conducted in a similar fashion to that adopted for continuous systems. Again, it must be stressed that such manipulation breaks down if the boxes load one another.

Consider the arrangement shown in *Figure 13.29*. Here we have a number of continuous systems, represented by their transfer functions, in cascade. However, a sampler has been placed in each signal line, and so for each box we may write

$$C_1(s) = G_1(s)R^*(s) \rightarrow C_1^*(s) = G_1^*(s)R^*(s)\Leftarrow$$

$$C_2(s) = G_2(s)C_1^*(s) \rightarrow C_2^*(s) = G_2^*(s)C_1^*(s)\Leftarrow$$

$$C(s) \;= G_3(s)C_2^*(s) \rightarrow C^*(s) = G_1(s)C_2^*(s)\Leftarrow$$

Thus

$$C^*(s) = G_1^*(s)G_2^*(s)G_3^*(s)R^*(s)\Leftarrow$$

i.e.

$$\frac{C(z)\Leftarrow}{R(z)\Leftarrow} = G_1(z)G_2(z)G_3(z)\Leftarrow$$

This, of course, generalises for n similar pulse transfer functions in series to give

$$\frac{C(z)\Leftarrow}{R(z)\Leftarrow} = \Leftarrow\prod_{i=1}^{n} G_i(z)\Leftarrow$$

It is necessary to realise that this result does not apply if there is no sampler between two or more boxes. As an illustration, *Figure 13.30*(*a*) shows the arrangement for which the above result applies. We have

$$G_1(s) = \frac{1}{\Leftarrow s}$$

whence (see *Table 13.1*)

$$G_1(z) = \frac{z}{\Leftarrow z-1}$$

and

$$G_2(s) = \frac{1}{\Leftarrow s+1}$$

whence (see *Table 13.1*)

$$G_2(z) = \frac{z}{\Leftarrow z-e^{-T}}$$

Therefore,

$$\frac{C(z)\Leftarrow}{R(z)} = G_1(z)G_2(z) = \frac{z^2}{\Leftarrow (z-1)(z-e^{-T})\Leftarrow}$$

Figure 13.30(*b*) shows the arrangement *without* a sampler between $G_1(s)$ and $G_2(s)$, and so

$$\frac{C(z)\Leftarrow}{R(z)\Leftarrow} = Z\left[\frac{1}{s(s+1)}\right]\Big(\Leftarrow \frac{z(1-e^{-T})}{(z-1)(z-e^{-T})}$$

Note that $Z[G_1(s)G_2(s)]$ is often written $G_1G_2(z)$, and thus, in general, $G_1(z)G_2(z) \neq G_1G_2(z)$.

13.16 Closed-loop systems

Figure 13.31 shows the sampler in the error channel of an otherwise continuous system. We may write

$$C(s) = G(s)E^*(s)\Leftarrow$$

and

$$E(s) = R(s) - H(s)C(s)\Leftarrow$$

or

$$E(s) = R(s) - H(s)G(s)E^*(s)\Leftarrow$$

and

$$E^*(s) = R^*(s) - HG^*(s)E^*(s)\Leftarrow$$

and so

$$E^*(s) = \frac{R^*(s)\Leftarrow}{\Leftarrow 1 + HG^*(s)\Leftarrow}$$

Thus

$$\frac{C^*(s)}{R^*(s)\Leftarrow} = \frac{G^*(s)\Leftarrow}{\Leftarrow 1 + HG^*(s)\Leftarrow}$$

or

$$\frac{C(z)\Leftarrow}{R(z)\Leftarrow} = \frac{G(z)\Leftarrow}{\Leftarrow 1 + HG(z)\Leftarrow}$$

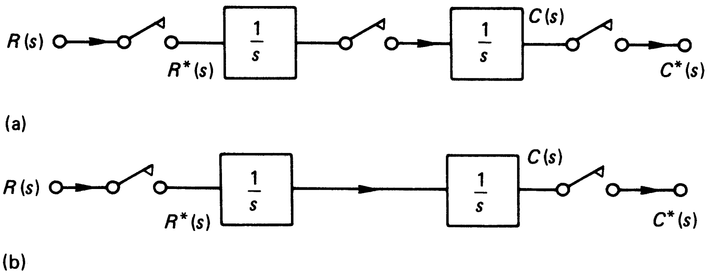


Figure 13.30 Two transfer functions with (a) sampler interconnection and (b) with continuous signal connecting transfer functions

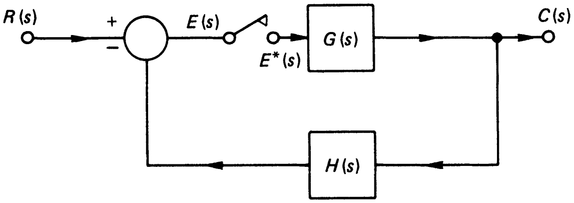


Figure 13.31 Prototype sampled system with a sampler in the error channel

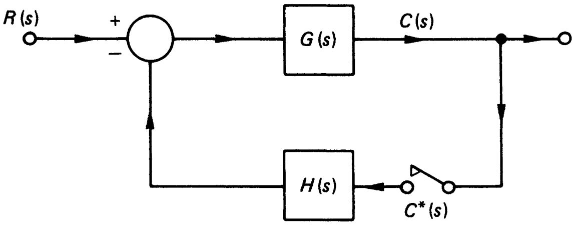


Figure 13.32 Prototype sampled system with a sampler in the feedback channel

If the sampler is in the feedback loop, as shown in *Figure 13.32*, a similar analysis would show that

$$C(z) = \Leftarrow\frac{GR(z)\Leftarrow}{1 + HG(z)\Leftarrow}$$

Note that, in this case, it is not possible to take the ratio $C(z)/R(z)$. We may conclude that the position of the sampler(s) within the loop has a vitally important effect on the behaviour of the system.

Example Consider the arrangement shown in *Figure 13.33*. To calculate the pulse transfer function it is necessary to determine

$$L\left[\frac{1}{s}\,\frac{1}{s}\,(1/ - e^{-sT})\right]\Big(\Big($$

Consider (from *Table 13.1*)

$$Z\left[\frac{1}{s^2}\right] = \Big(\Leftarrow\frac{Tz}{(z-1)^2}$$

and, therefore,

$$Z\left[e^{-sT}\frac{1}{s^2}\right]\Big(= z^{-1}Z\left[\frac{1}{s^2}\right] = \Big(\Leftarrow\frac{T}{(z-1)^2}$$

Thus

$$Z\left[\frac{1}{s^2}(1 - e^{-Ts})\right]\Big(= \Leftarrow\frac{T}{(z-1)\Leftarrow} = G(z)\Leftarrow$$

and

$$\frac{C(z)\Leftarrow}{R(z)\Leftarrow} = \frac{G(z)\Leftarrow}{1 + G(z)\Leftarrow} = \Leftarrow\frac{T}{z + (T-1)\Leftarrow}$$

13.17 Stability

It should be appreciated from the above that, in general, $C(z)/R(z)$ results in a ratio of polynomials in z in a similar way as, for continuous systems, $C(s)/R(s)$ results in a ratio of polynomials in s. Thus, just as the equation $1 + G(s)H(s) = 0$ is called the 'characteristic equation' for the continuous system, $1 + GH(z) = 0$ is the characteristic equation for the sampled-data system. Both of these characteristic equations are polynomials in their respective variables, and the positions of the roots of these equations determine the characteristic behaviour of the corresponding closed-loop systems. Mathematically, the process of determining the roots is identical in the two cases. The difference between the two characteristic equations arises because of the need to interpret the effects of the location of the roots, when they are plotted in their respective s and z planes, on the two plants. For continuous systems, if any of these poles are located in the right-half s plane, then the system is unstable. Similarly, since the whole of the left-hand s plane maps into the unit-circle of the z plane under the transformation $z = e^{sT}$, then in the simple-data case, for stability *all* of the roots of $1 + GH(z) = 0$ must lie within the unit circle.

Much of the design process of control systems is to arrange for the roots of the characteristic equation to locate at desired positions in either the s or z plane. It will be recalled, from continuous theory, that the locus of these roots, as a particular parameter is varied, may be determined by using the root-locus technique. Thus, since the characteristic equation of the sample-data system has a similar form (i.e. a polynomial), the root-locus technique may be applied to $1 + GH(z) = 0$ in exactly the same way. Only once the root-locus has been determined is there a difference in interpreting the effects of pole positions between the two cases.

13.18 Example

Consider the system shown in *Figure 13.34*, and suppose that the requirement is to draw the root-locus diagrams for, say, sampling periods of 1 and 0.5 secs.

The first requirement is to determine the pulse-transfer function for the open loop, i.e. $G(z)$:

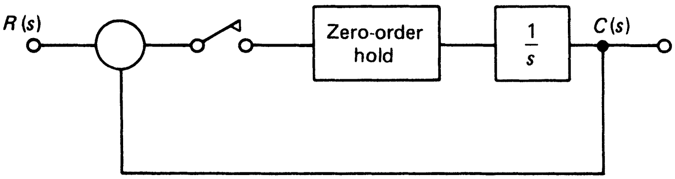


Figure 13.33 Arrangement used in the example in Section 13.16

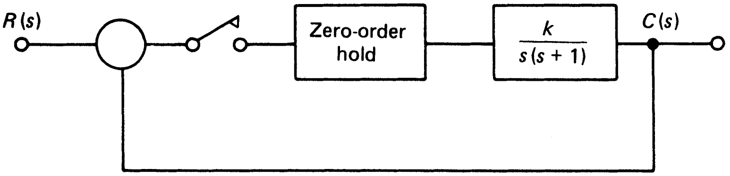


Figure 13.34 Arrangement used in the example in Section 13.18

$$G(z) = Z\left[\frac{K(1-e^{-Ts})}{s^2(s+1)}\right] = K(1-z^{-1})Z\left[\frac{1}{s^2(s+1)}\right]$$

Consider

$$Z\left[\frac{1}{s^2(s+1)}\right] = Z\left[\frac{1}{s^2} - \frac{1}{s} + \frac{1}{s+1}\right]$$

where, from *Table 13.2*, we have

$$Z\left[\frac{1}{s^2(s+1)}\right] = \frac{T_z}{(z-1)^2} - \frac{z}{z-1} + \frac{z}{(z-e^{-T})} = z\left[\frac{z(T+e^{-T}-1)+1-e^{-T}(1+T)}{(z-1)^2(z-e^{-T})}\right]$$

and so

$$G(z) = \frac{K[z(T+e^{-T}-1)+1-e^{-T}(1+T)]}{(z-1)(z-e^{-T})}$$

Thus, when $T=1$ secs,

$$G_1(z) = \frac{0.368K(z+0.718)}{(z-1)(z-0.368)}$$

and when $T=0.5$ secs

$$G_2(s) = \frac{0.107K(z+0.841)}{(z-1)(z-0.606)}$$

Both of these equations have two real poles and one zero pole, and the root loci are as shown in *Figures 13.35* and *13.36*. It can be seen that the difference in the two sampling times T causes fairly dramatic changes; when $T=1$ secs the system becomes unstable at $K=1.9$, and when $T=0.5$ secs the system becomes unstable at $K=3.9$. The process of drawing the root locus for either a continuous plant or a sampled-data plant is identical. It is the interpretation of the positions of the roots that is different, although in both cases the design is to place the roots in acceptable locations in the two planes. It is possible to use Bode diagrams in sampled-data design work and this is explained in many of the references given in the Bibliography at the end of this chapter.

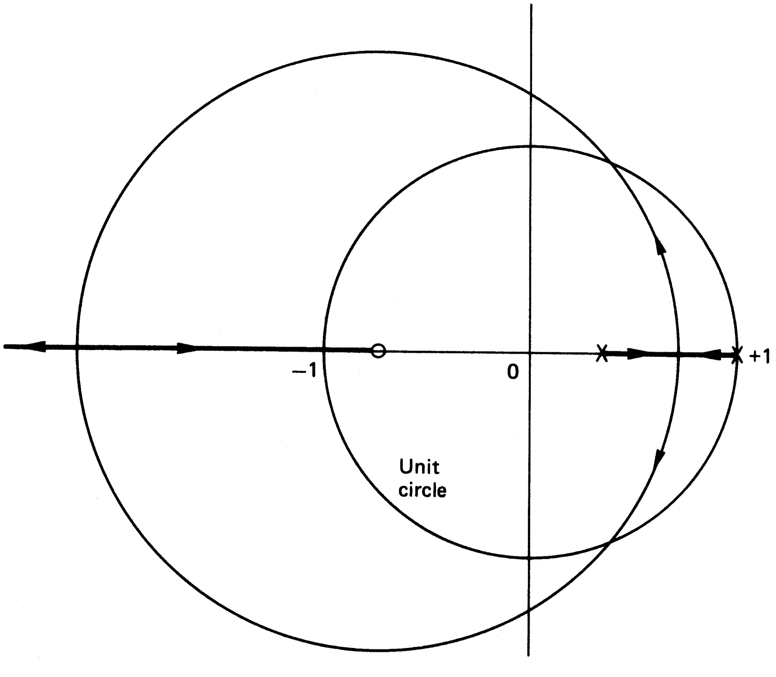


Figure 13.35 Root locus plot: $G(z) = [0.368K(z+0.718)]/[(z-1)(z-0.368)]$

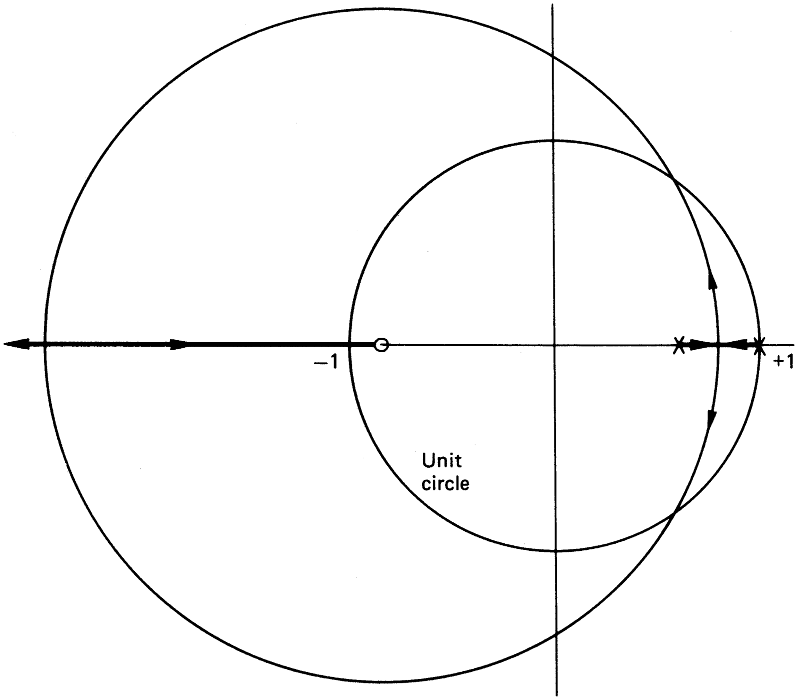


Figure 13.36 Root-locus plot: $G(z) = [0.107K(z + 0.841)]/[(z - 1)(z - 0.606)]$

13.19 Dead-beat response

Consider the system shown above where $T = 1$ secs and $K = 1$; suppose that compensation of the form

$$D(z) = \frac{1.582(z - 0.368)}{(z + 0.418)}$$

is inserted immediately after the sampler. Then it is easy to show that

$$\frac{C(z)}{R(z)} = \frac{0.582(z + 0.71)}{z^2}$$

If

$$R(z) = \frac{z}{z - 1}$$

i.e. $r(t)$ is a unit step function, then

$$\begin{aligned} C(z) &= \frac{0.582(z + 0.718)}{z(z - 1)} \\ &= \frac{1}{z}\left[\frac{0.582z + 0.418}{z - 1}\right] \\ &= \frac{1}{z}\left[0.582 + \frac{1}{z} + \frac{1}{z^2} + \cdots\right] \end{aligned}$$

i.e. $c(0) = 0$, $c(1) = 0.582$ and $c(n) = 1$, for $n = 2, 3, \ldots$.

The implication is that $c(t)$ has reached its target position after two sample periods. If an nth order system reaches its target position in, at most, nth sampling instants, then this is called a 'dead-beat response'; a controller that achieves this, such as $D(z)$ above, is called a 'dead-beat controller' for this system. This is an interesting response, for it is not possible to achieve this with a continuous control system.

At least two dangers are inherent in dead-beat controllers:

(1) the demanded controller outputs during the process may be excessive; and
(2) there may be an oscillation set up which is not detected without further analysis.

In fact, the system is only 'dead beat' at the sampling instants. Indeed, in the above example, there is an oscillation between sampling instants of about 10% of the step value. However, theoretically it is possible for a sampled-data system to complete a transient of the above nature in finite time.

13.20 Simulation

13.20.1 System models

Regardless of the simulation language to be used, a necessary prerequisite is a description of the system of interest by a mathematical model. Some physical systems can be described in terms of models that are of the state transition type. If such a model exists, then given a value of the system variable of interest, e.g. voltage, charge position, displacement, etc., at time t, then the value of the variable (state) at some future time $t + \Delta t$ can be predicted. The prediction of the variable of interest (state variable) $x(t)$ at time $t + \Delta t$, given a state transition model S, can be expressed by the state equation:

$$x(t + \Delta t) = S[x(t), t, \Delta t] \tag{13.5}$$

Equation (13.5) shows that the future state is a function of the current state $x(t)$ at the current time t and the time increment Δt. Thus, once the model is known, from either empirical or theoretical considerations, Equation (13.5), given an initial condition (value), allows for the recursive computation of $x(t)$ for any number of future steps in time. For an initial value of the state variable $\bar{x} = x(t_1)$ at time t_1, then

$$x(t_1 + \Delta t) = S[\bar{x}, t_1, \Delta t] \Leftarrow$$

then letting $t_2 = t_1 + \Delta t$, Equation (13.5) for the next time step Δt, becomes

$$x(t_2 + \Delta t) = S[x(t_2), t_2, \Delta t] \Leftarrow$$

Obviously, this operation is continued until the calculation of the state variable has been performed for the total time period of interest.

Systems of interest will clearly not be characterised only by a single state variable but by several state variables. *Figure 13.37* is a schematic representation of a multivariable system that has r inputs, n states and m outputs.

In general, the simulation will involve calculation of all of the state variables, even though the response of only a selected number of output variables is of interest. For many systems, the output variables may well exhibit a simple one-to-one correspondence to the state variables. As shown by the representation in *Figure 13.37*, the values of the state variables depend on the inputs to the system. For a single interval, between the k and $k+1$ time instants, the state equations for the n state variable system for a change in the jth input $(j \leq r) u_j(t)$ is written as

$$\begin{aligned} x_1(t_k + \Delta t) &= S_1[x_1(t_k), u_j(t_k), t_k, \Delta t] \Leftarrow \\ x_2(t_k + \Delta t) &= S_2[x_2(t_k), u_j(t_k), t_k, \Delta t] \Leftarrow \\ &\vdots \qquad \vdots \\ x_n(t_k + \Delta t) &= S_n[x_n(t_k), u_j(t_k), t_k, \Delta t] \Leftarrow \end{aligned} \qquad (13.6) \Leftarrow$$

The above system of equations, a collection of difference equations, would be used to predict the state variables $x_1, x_2, \ldots, x_n$ at time intervals of Δt from the initial time t_0 until the total time duration of interest $T = t_0 + K\,\Delta t$. For engineering systems, the dependent variable will generally be a continuous variable. In this case the system description will be in terms of a differential equation of the form

$$\mathrm{d}x/\mathrm{d}t = g(x, t) \qquad (13.7) \Leftarrow$$

Recalling basic calculus for a small time increment, the left-hand side of Equation (13.7) can be expressed as

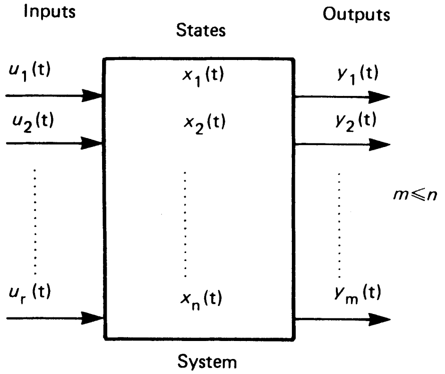


Figure 13.37 Schematic representation of a multi-variable system

$$\lim_{\Delta t \to 0} \frac{x(t + \Delta t) - x(t) \Leftarrow}{\Delta t}$$

so, for a small time increment Δt, Equation (13.7) can be written as

$$x(t + \Delta t) = x(t) + [g(x, t)]\Delta t$$

or $\qquad (13.8) \Leftarrow$

$$x(t + \Delta t) = G[x(t), t, \Delta t)] \Leftarrow$$

Equation (13.8) is a form of Equation (13.5), so for a small time increment, a first-order ordinary differential equation can be approximated by a state transition representation.

It thus follows from the preceding discussion that, in digital continuous system simulation, the principal numerical task is the approximate integration of Equation (13.7). For a small time increment DT, the integration step size, the computation involves the evaluation of the difference equation

$$x(t + DT) = x(t) + [g(xt)]DT) \qquad (13.9a) \Leftarrow$$

which can be written explicitly as

$$x(t_{k+1}) = x(t_k) + \int_{t_k}^{t_{k+1}} \Leftarrow g[x(t_k), t_k]DT) \qquad (13.9b) \Leftarrow$$

where $DT = t_{k+1} - t_k$. The calculation starts with a known value of the initial state $x(0)$ at time t_0 and proceeds successively to evaluate $x(t_1)$, $x(t_2)$, etc. The computation involves successive computation of $x(t_{k+1})$ by alternating calculation of the derivative $g[x(t_k), t_k]$ followed by integration to compute $x(t_{k+1})$ at time $t_{k+1} = t_k + DT$.

Obviously, most physical systems will be described by second or higher order ordinary differential equations so the higher order equation must be re-expressed in terms of a group of first-order ordinary differential equations by introducing state variables. For an nth order equation,

$$\frac{\mathrm{d}^n z}{\mathrm{d}t^n} = f\left[z, \frac{\mathrm{d}z}{\mathrm{d}t}, \varsigma \frac{\mathrm{d}^2 z}{\mathrm{d}t^2} \cdots \Leftarrow \frac{d^{n-1} z}{\mathrm{d}t^{n-1}}; t\right] \Bigg(\qquad (13.10) \Leftarrow$$

the approach involves the introduction of new variables as state variables to yield the following first-order differential equations

$$\begin{aligned} \frac{\mathrm{d}x_1}{\mathrm{d}t} &= x_2 \\ \frac{\mathrm{d}x_2}{\mathrm{d}t} &= x_3 \\ \frac{\mathrm{d}x_3}{\mathrm{d}t} &= x_4 \\ &\vdots \\ \frac{\mathrm{d}x_{n-1}}{\mathrm{d}t} &= x_n \\ \frac{\mathrm{d}x_n}{\mathrm{d}t} &= f(x_1, x_2, x_3, \ldots, x_n; t) \Leftarrow \end{aligned} \qquad (13.11) \Leftarrow$$

It should be noted that this equation can be expressed in shorthand notation as a vector-matrix differential equation. In an analogous manner, Equation (13.6) can be expressed as a vector different equation. There is no unique approach to the selection of state variables for system representation, but for many systems the choice of state variables will be obvious. In electric circuit problems, capacitor voltages and inductor currents would be logical choices, as would position, velocity and acceleration for mechanical systems.

13.20.2 Integration schemes

The simple integration step, embodied by the first-order Euler form in Equation (13.9a) only provides a satisfactory approximation of the solution of the differential equation, within specified error limits, for a very small integration step size DT. Since the small integration interval leads to substantial computing effort and to round-off error accumulation, all digital simulation languages use improved integration schemes. Despite the wide variety of different integration schemes that are available in the many different simulation languages, the calculational approach can be categorised into two groups. The types of algorithm are:

(1) *Multi-step formulae* In such algorithms, the value of $x(t+DT)$ is not calculated by the simple linear extrapolation of Equation (13.9a). Rather than use only $x(t)$ and one derivative value, the algorithms use a polynomial approximation based on past values of $x(t)$ and $g[x(t),t]$, that is at times $t-DT$, $t-2DT$, etc.
(2) *Runge–Kutta formulae* In Runge–Kutta type algorithms, the derivative value used for the calculation of $x(t+DT)$ is not the point value at time t. Instead, two or more approximate derivative values in the interval t, $t+DT$ are calculated and then a weighted average of these derivative values is used instead of a single value of the derivative to compute $x(t+DT)$.

13.20.3 Organisation of problem input

Most simulation language input is structured into three separate sections, although in some programs the statement can be used with limited sectioning of the program. A typical structure and the type of statements, functions or parts of the simulation program that appear are as follows.

(1) *Initialisation*
    Problem documentation (e.g. name, date, etc.).
    Initial conditions for state variables.
    Parameter values (problem variables that may not be constant, problem time, integration order, integration step size, etc.).
    Problem constants.
(2) *Dynamic*
    Derivative statements.
    Integration statements (including any control parameters not given in the initialisation section).
(3) *Terminal*
    Conditional statements (e.g. total time, variable(s), value(s), etc.).
    Multiple run parameters.
    Output (print/plot/display) option(s).
    Output format (e.g. designation of independent variable; increment for independent variable; dependent variable(s) to be output; maximum and minimum values of variable(s); or automatic scaling; total number of points for the independent variable or total length of time).

It should be understood that the specific form of the statements within each section is not exactly the same for all digital simulation languages. However, from the continuous system modelling package (CSMP) simulation programs presented in the next section, with the aid of the appropriate language manual, there should be no difficulty in formulating a simulation program using any continuous system simulation language (CSSL)-type digital simulation program.

13.20.4 Illustrative example

Simulation programs are presented, using the CSMP language, that would be suitable for investigating system dynamic behaviour. The system model, although relatively simple in nature, is typical of those used for system representation.

13.20.4.1 Example

Frequently, it will be found that system dynamic behaviour can be described by a differential equation of the form

$$\begin{aligned} y^n + a_1 y^{n-1} + a_2 y^{n-2} + a_{n-1} y^1 + a_n y \\ = b_0 r^m + b_1 r^{m-1} + b_{m-1} r^1 + b_m r \end{aligned} \qquad (13.12)\Leftarrow$$

where

$$y^n = \frac{\mathrm{d}^n y}{\mathrm{d}t^n} \text{ and } r^m = \frac{\mathrm{d}^m r}{\mathrm{d}t^m}$$

Use of CSMP for studying the dynamic behaviour of a system described by a high-order differential equation is illustrated here using a simulation program for the differential equation

$$y^3 + 2.5y^2 + 3.4y^1 + 0.8y = 7.3r \qquad (13.13)\Leftarrow$$

with the initial conditions

$$y^2(0) = 0; y^1(0) = -4.2; \; y^0 = 2.5$$

Development of the simulation program follows logically by rewriting Equation (13.13) as

$$\frac{\mathrm{d}^3 y}{\mathrm{d}t^3} = -2.5\frac{\mathrm{d}^2 y}{\mathrm{d}t^2} - 3.4\frac{\mathrm{d}y}{\mathrm{d}t} - 0.8y + 7.3r \qquad (13.14)\Leftarrow$$

$$\left.\frac{\mathrm{d}^2 y}{\mathrm{d}t^2}\right|_{t=0} = 0; \quad \left.\frac{\mathrm{d}y}{\mathrm{d}t}\right|_{t=0} = -4.2; \quad y|_{t=0} = 2.5$$

A block diagram showing the successive integrations to be solved for the dependent variable y is given in *Figure 13.38*. As can be seen from the labelling on the diagram, the output of the integration blocks is successive derivative values and

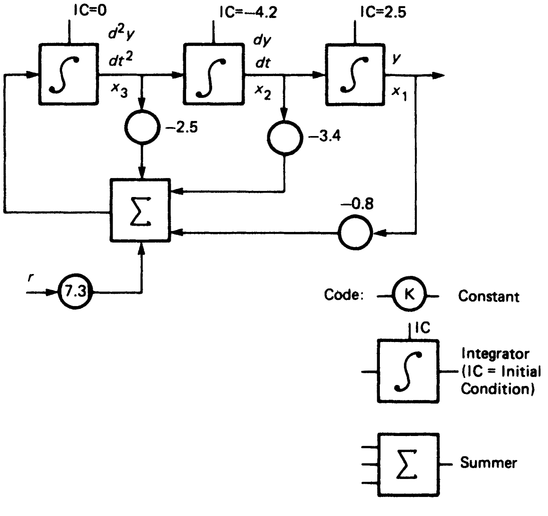


Figure 13.38 CSMP block diagram for a third-order differential equation

```
LABEL THIRD ORDER DIFFERENTIAL EQUATION
INITIAL
CONSTANT A1=-2.5,A2=-3.4,A3--0.8,B0=7.3,  ...
         X1INIT=2.5,X2INIT=-4.2,X3INIT-0.0
FUNCTION FCHG=(0.5,4.8),(1.0,6.3),(1.5,2.8),(2.0,3.9),  ...
              (2.5,4.8),(3.0,3.2),(3.5,2.1),(4.0,5.6),  ...
              (4.5,6.8),(5.0,3.7),(5.5,4.6),(6.0,3.4)
DYNAMIC
     R=NLFGEN(FCHG,TIME)
     X1=INTGRL(X1INIT,X2)
     X2=INTGRL(X2INIT,X3)
     X3=INTGRL(X3INIT,DHX3)
     DHX3=A1*X3+A2*X2+A3*X1+B0*R
TERMINAL
TIMER FINTIM=6.0,PRDEL=0.2
PRINT R,X1,X2,X3
END
STOP
ENDJOB
```

Figure 13.39 Simulation program for studying the dynamic behaviour of a system described by a third-order differential equation

the dependent variable. In fact, the output of each integration block is a state variable. This becomes obvious by introducing new variables, x_1, x_2, x_3 defined as:

$$x_1 = y$$

$$\frac{dx_1}{dt} = x_2$$

$$\frac{dx_2}{dt} = x_3$$

which allows Equation (13.14) to be expressed as

$$\frac{dx_1}{dt} = x_2 \qquad (13.15)\Leftarrow$$

$$\frac{dx_2}{dt} = x_3$$

$$\frac{dx_3}{dt} = -2.5x_3 - 3.4x_2 - 0.8x_1 + 7.3r$$

with the initial conditions

$$x_3(0) = 0; \quad x_2(0) = -4.2; \quad x_1(0) = 2.5$$

A program for solving equation (13.15) is given in *Figure 13.39*. Examination of the program shows that the value of the forcing function r is not constant but varies with time. The variation is provided using the quadratic interpolation function, NLFGEN. Total simulation time is set for 6 min with the interval for tabular output specified as 0.2 min. The time unit is determined by the problem parameters. It is to be noted that the program does not include any specification for the method of integration. The CSMP language does not require that a method of integration be given, but a particular method may be specified. If a method is not given, then by default the variable step size fourth-order Runge–Kutta method is used for calculation. The initial step size, by default, is taken as 1/16 of the PRDEL (or OUTDEL) value. Minimum step size can be limited by giving a value for DELMIN as part of the TIMER statement. If a DELMIN value is not given then, by default, the minimum step size is FINTIM $\times 10^{-7}$.

13.21 Multivariable control

Classical process control analysis is concerned with single loops having a single setpoint, single actuator and a single controlled variable. Unfortunately, in practice, plant variables often interact, leading to interaction between control loops. A typical interaction is shown on *Figure 13.40*, where a single combustion air fan feeds several burners in a multizone furnace. An increase in air flow, via V_1 say to raise the temperature in zone 1, will lead to a reduction in the duct air pressure P_d, and a fall in air flow to the other zones. This will lead to a small fall in temperature in the other zones which will cause their temperature controllers to call for increased air flow which affects the duct air pressure again. The temperature control loops interact via the air valves and the duct air pressure.

Where interaction between variables is encountered an attempt should always be made to remove the source of the interaction, as this leads to a simpler, more robust, system. In *Figure 13.40*, for example, the interaction could be reduced significantly by adding a pressure control loop which maintains duct pressure by using a VF to set the speed of the combustion airfan. Often, however, the interaction is inherent and cannot by removed.

Figure 13.41 is a general representation of two interacting control loops. The blocks C_1 and C_2 represent the controllers comparing setpoint R with process variable V to give a controller output U. The blocks K_{ab} represent the transfer function relating variable a to controller output b. Blocks K_{11} and K_{12} are the normal forward control path, with blocks K_{21} and K_{12} representing the interaction between the loops.

The process gain of process 1 can be defined as $\Delta V_1/\Delta U_2$ where Δ denotes small change. This process gain can be measured with loop 2 in open loop (i.e. U_2 fixed) or loop 2 in closed loop control (i.e. V_2 fixed) we can thus observe two gains

$$K_{2\mathrm{OL}} = \Leftarrow \frac{\Delta V_1}{\Delta U_1} \text{ for loop 2 open loop}$$

and

$$K_{2\mathrm{CL}} = \Leftarrow \frac{\Delta V_1}{\Delta U_1} \text{ for loop 2 closed loop}$$

The gains will, of course vary with frequency and have magnitude and phase shift components. We can now define a relative gain $\lambda\varsigma$for loop 1

$$\lambda\varsigma = \Leftarrow \frac{K_{2\mathrm{OL}}}{K_{2\mathrm{CL}}}$$

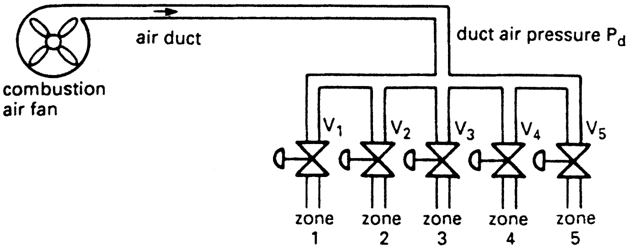


Figure 13.40 A typical example of interaction between variables in multi-variable control. The air flows interact via changes in the duct air pressure

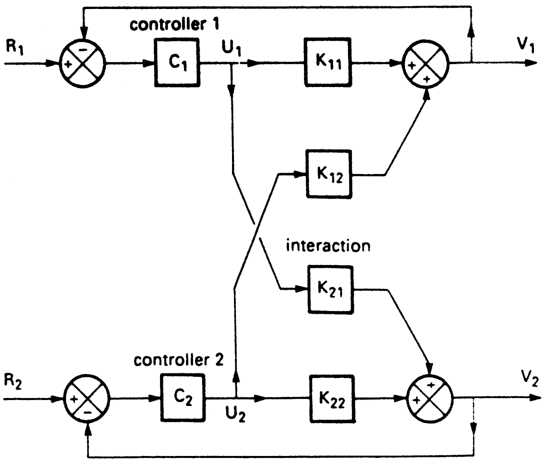


Figure 13.41 General representation of interacting loops

If λ is unity, changing from manual to auto in loop 2 does not affect loop 1, and there is no interaction between the loops.

If $\lambda < 1$, the interaction will apparently increase process 1 gain when loop 2 is switched to automatic. If $\lambda > 1$, process 1 gain will apparently be decreased when loop 2 is in automatic.

This apparent change in gain can be seen with loop 2 in manual, U_2 is fixed, so K_{2OL} is simply K_{11}. To find K_{2CL} we must consider what happens when loop 2 effectively shunts K_{11}. We have

$$V_1 = K_{11}U_1 + K_{12}U_2 \qquad (13.16)$$

and

$$V_2 = K_{22}U_2 + K_{21}U_1 \qquad (13.17)$$

Re-arranging Equation 13.17 gives

$$U_2 = \frac{V_2 - K_{21}U_1}{K_{22}}$$

which can be substituted in Equation 13.16 giving

$$V_1 = K_{11}U_1 + \frac{K_{12}}{K_{22}}(V_2 - K_{21}U_1)$$

The process 1 gain with loop 2 in auto is

$$K_{2CL} = \frac{dV_1}{dU_1}$$

$$= \frac{K_{11}K_{22} - K_{12}K_{21}}{K_{22}}$$

The relative gain, λ , is

$$\lambda = \frac{K_{2OL}}{K_{2CL}}$$

$$= \frac{1}{1 - K_{12}K_{21}/K_{11}K_{22}}$$

It should be remembered that the gains K_{ab} are dynamic functions, so λ will vary with frequency.

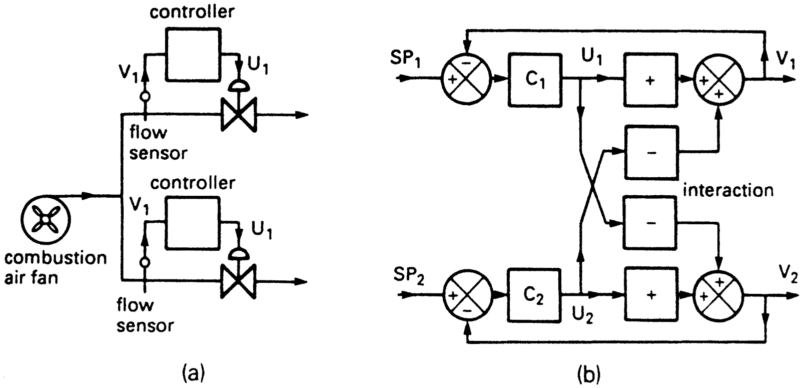
The term $(K_{12}K_{21}/K_{11}K_{22})$ is the ratio between the interaction and forward gains. This should be in the range 0 to 1. If the term is greater than unity, the interactions have more effect than the supposed process, and the process variables are being manipulated by the wrong actuators!

It is possible to determine the range of λ from the relationship $(K_{12}K_{21}/K_{11}K_{22})$. If this is positive, λ will be greater than unity, and loop 1 process gain will decrease when loop 2 is switched to auto. This will occur if there is an even number of K_{ab} blocks with negative sign (0, 2 or 4). If the relationship is negative, λ will be less than unity and loop 1 process gain will increase when loop 2 is closed. This occurs if there is an odd number of blocks with negative sign (1 or 3).

The combustion air flow system of *Figure 13.40* is redrawn on *Figure 13.42(a)*. Increasing U_1 obviously decreases V_2, and increasing U_2 similarly decreases V_1. The interaction block diagram thus has the signs of *Figure 13.42(b)*. There are two negative blocks, so λ is greater than unity.

If λ is greater than unity, the interaction can be considered benign as the reduced process gain will tend to increase the loop stability (albeit at the expense of response time). The loops can be tuned individually in the knowledge that they will remain stable with all loops in automatic control.

If λ is in the range $0 < \lambda < 1$, care must be taken as it is possible for loops to be individually stable but collectively unstable requiring a reduction in controller gains to maintain stability. The closer λ gets to zero, the greater the interaction and the more de-gaining will be required.



(a) (b)

Figure 13.42 The combustion air system redrawn to show interactions: (a) block diagram; (b) interaction diagram, with two negative blocks the interaction decreases the apparent process gain and the interaction is benign

The calculation of dynamic interaction is difficult, even for the two variable case.With more interacting variables, the analysis becomes exceedingly complex and computer solutions are best used. Ideally, though, interactions once identified, should be removed wherever possible.

13.22 Dealing with non-linear elements

13.22.1 Introduction

All systems are non linear to some degree. Valves have non linear transfer functions, actuators often have a limited velocity of travel and saturation is possible in every component. A controller output is limited to the range 4–20 mA, say, and a transducer has only a restricted measurement range.

One of the beneficial effects of closed loop control is the reduced effect of non linearities. The majority of non linearities are therefore simply lived with, and their effect on system performance is negligible. Occasionally, however, a non linear element can dominate a system and in these cases its effect must be studied.

Some non linear elements can be linearised with a suitable compensation circuit. Differential pressure flow meters have an output which is proportional to the square of flow. Following a non linear differential pressure flow transducer with a non linear square root extractor gives a linear flow measurement system.

Cascade control can also be used around a non linear element to linearise its performance as seen by the outer loop. Butterfly valves are notoriously non linear. They have an S shaped flow/position characteristic, suffer from backlash in the linkages and are often severely velocity limited. Enclosing a butterfly valve within a cascade flow loop, for example, will make the severely non linear flow control valve appear as a simple linear first order lag to the rest of the system.

There are two basic methods of analysing the behaviour of systems with non linear elements. It is also possible, of course, to write computer simulation programs and often this is the only practical way of analysing complex non linearities.

13.22.2 The describing function

If a non linear element is driven by a sine wave, its output will probably not be sinusoidal, but it will be periodic with the same frequency as the input, but of differing shape and possibly shifted in phase as shown on *Figure 13.43*. Often the shape and phase shift are related to the amplitude of the driving signal.

Fourier analysis is a technique that allows the frequency spectrum of any periodic waveform to be calculated. A simple pulse can be considered to be composed of an infinite number of sine waves.

The non linear output signals of *Figure 13.43* could therefore be represented as a frequency spectrum, obtained from Fourier analysis. This is, however, unnecessarily complicated. Process control is generally concerned with only dominant effects, and as such it is only necessary to consider the fundamental of the spectrum. We can therefore represent a non linear function by its gain and phase shift at the fundamental frequency. This is known as the *describing function*, and will probably be frequency and amplitude dependant.

Figure 13.44 shows a very crude bang/bang servo system used to control level in a header tank. The level is sensed by a capacitive probe which energises a relay when a nominal depth of probe is submerged. The relay energises a solenoid which applies pneumatic pressure to open a flow valve. This system is represented by *Figure 13.45*.

The level sensor can be considered to be a level transducer giving a 0–10 V signal over a 0.3 m range. The signal is filtered with a 2 sec time constant to overcome noise from splashing, ripples etc. The level transducer output is compared with the voltage from a setpoint control and the error signal energises or de-energises the relay. We shall assume no hysteresis for simplicity although this obviously would be desirable in a real system.

The relay drives a solenoid assumed to have a small delay in operation which applies 15 psi to an instrument air pipe to open the valve. The pneumatic signal takes a finite time to travel down the pipe, so the solenoid valve and piping are considered as a 0.5 sec transit delay. The valve actuator turns on a flow of 150 $\mathrm{m}^3/\mathrm{min}$ for an applied pressure of 15 psi. We shall assume it is linear for other applied pressures. The actuator/valve along with the inertia of the water in the pipe appear as a first order lag of 4 sec time constant. The tank itself appears as an integrator from flow to level.

This system is dominated by the non linear nature of the level probe and the solenoid. The rest of the system can be considered linear if we combine the level comparator, relay and solenoid into a single element which switches 0 to 15 psi according to the sign of the error signal (15 psi for negative error, i.e. low level).

This non linear element will therefore have the response of *Figure 13.46*. when driven with a sinusoidal error signal. The output will have a peak to peak amplitude of 15 psi regardless of the error magnitude.

From Fourier analysis, the fundamental component of the output signal is a sine wave with amplitude $4 \times 7.5/\pi$ psi as shown. The phase shift is zero at all frequencies. The non linear element of the comparator/relay/solenoid can thus be considered as an amplifier whose gain varies with the amplitude of the input signal.

For a 1 V amplitude error signal the gain is

$$(4 \times 7.5)/(\pi \times 1) = 9.55$$

For a 2 V amplitude error signal the gain is

$$(4 \times 7.5)/(\pi \times 2) = 4.78$$

In general, for an E volt error signal the gain is

$$(4 \times 7.5)/(\pi \times E) = 9.55/E \qquad (13.18)$$

Figure 13.47 is a Nichols chart for the linear parts of the system. This has 180° phase shift for $\omega = 0.3$ rads/sec, so if it was controlled by a proportional controller, it would oscillate at 0.3 rads/sec if the controller gain was sufficiently high. The linear system gain at this frequency is -7dB, so a proportional controller gain of 7dB would just sustain continuous oscillation.

Let us now return to our non linear level switch. This has a gain which varies inversely with error amplitude. If we are, for some reason, experiencing a large sinusoidal error signal the gain will be low. If we have a small sinusoidal error signal the gain will be high.

Intuitively we know the system of *Figure 13.44* will oscillate. The non linear element will add just sufficient gain to make the Nichols chart of *Figure 13.47* pass through the 0 dB/$-180°$ origin. Self sustaining oscillations will result at 0.3 rads/sec. If these increase in amplitude for some reason, the gain will decrease causing them to decay again. If they cease, the gain will increase until oscillations recommence.

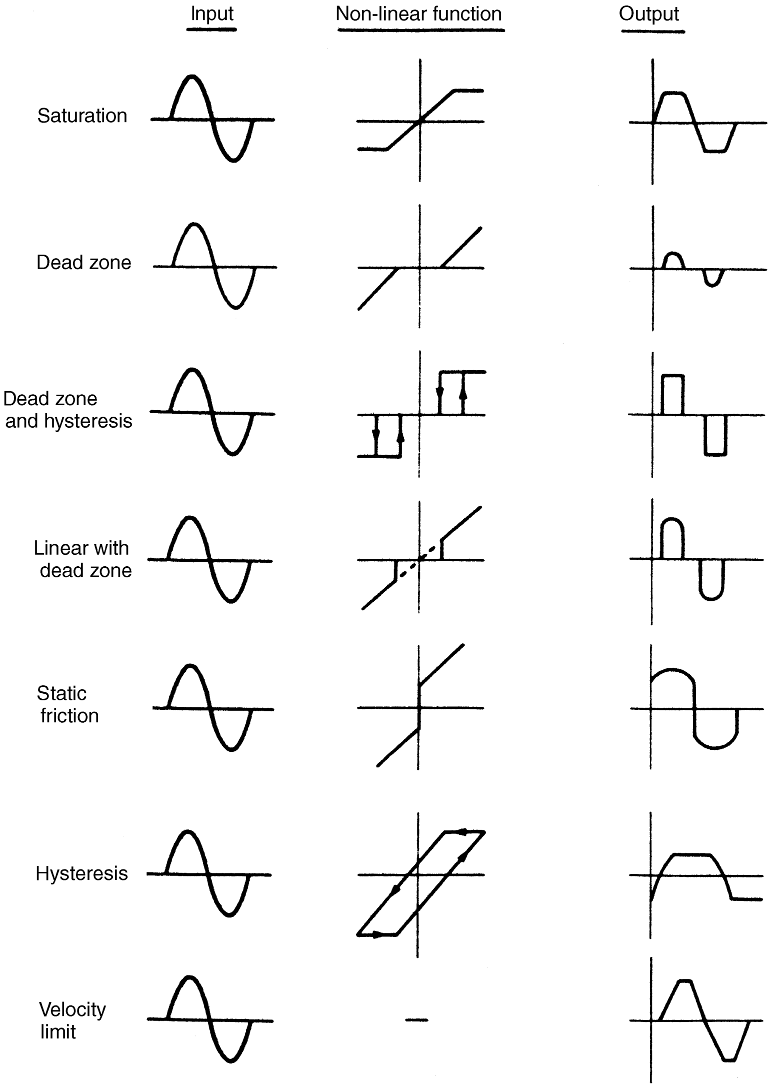

Input
Non-linear function
Output
Saturation
Dead zone
Dead zone and hysteresis
Linear with dead zone
Static friction
Hysteresis
Velocity limit
–

Figure 13.43 Common non-linearities

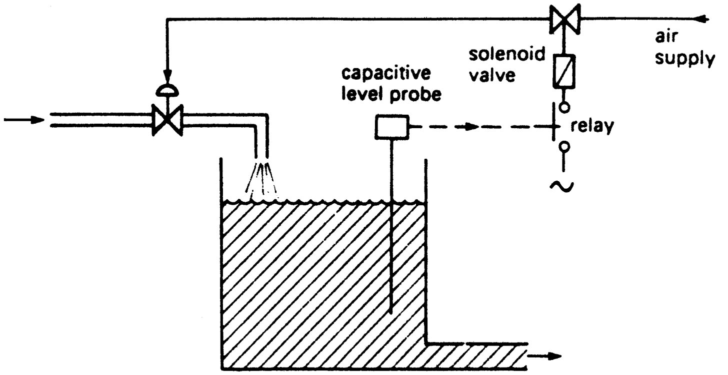

air supply
solenoid valve
capacitive level probe
relay
~

Figure 13.44 Bang-bang level control system

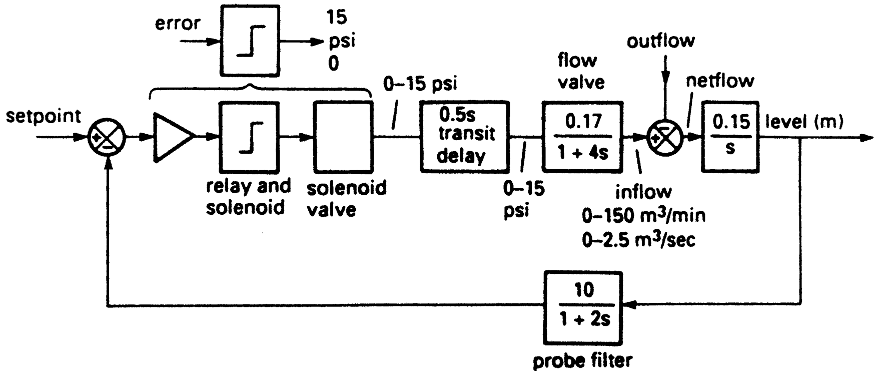

error
15
psi
0
0–15 psi
setpoint
relay and
solenoid
solenoid
valve
0.5s
transit
delay
0–15
psi
flow
valve
0.17
1 + 4s
outflow
netflow
inflow
0–150 m³/min
0–2.5 m³/sec
0.15
s
level (m)
10
1 + 2s
probe filter

Figure 13.45 Block diagram of bang-bang level control system

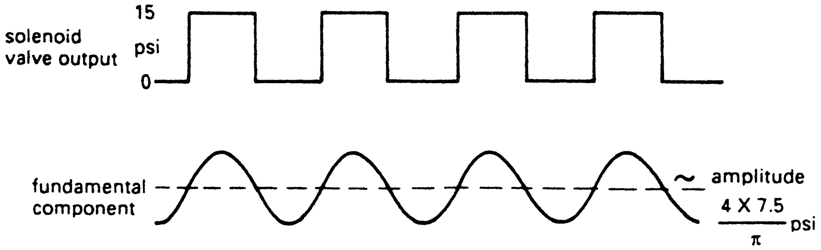

solenoid
valve output
15
psi
0
fundamental
component
~ amplitude
4 X 7.5
π
psi

Figure 13.46 Action of solenoid valve in level control system

The system stabilises with continuous constant amplitude oscillation.

To achieve this the non linear element must contribute 7 dB gain, or a linear gain of 2.24. From Equation 13.18 above, the gain is 9.55/*E* where *E* is the error amplitude. The required gain is thus given by an error amplitude of 9.55/2.24 = 4.26 V. This corresponds to an oscillation in level of 0.426 m.

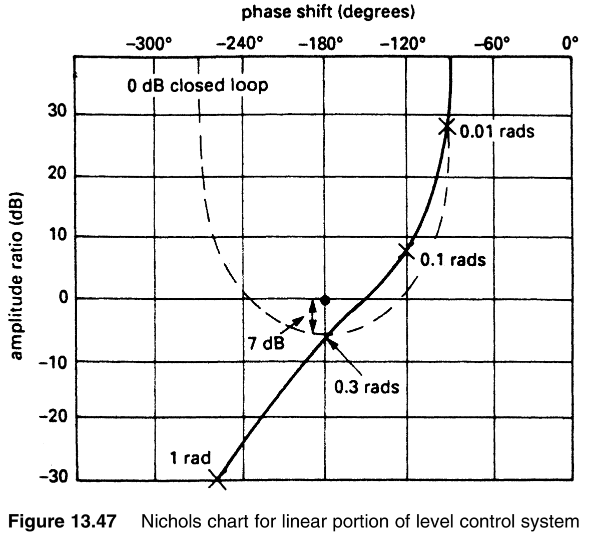


Figure 13.47 Nichols chart for linear portion of level control system

The system will thus oscillate about the set level with an amplitude of 0.4 m (the assumptions and approximations give more significant figures a relevance they do not merit) and an angular frequency of 0.4 rads/sec (period fractionally over 20 s).

There is a hidden assumption in the above analysis that the outgoing flow is exactly half the available ingoing flow to give equal mark/space ratio at the valve. Other flow rates will give responses similar to *Figure 13.48*, exhibiting a form of pulse width modulation. The relatively simple analysis however has told us that our level control system will sustain constant oscillation with an amplitude of around half a metre and a period of about 20 sec at nominal flow.

Similar techniques can be applied to other non linearities; a limiter, shown on *Figure 13.49*(*a*) and (*b*), for example, will have unity gain for input amplitudes less than the limiting level. For increasing amplitude the apparent gain will decrease. The describing function when limiting occurs has a gain dependent on the ratio between the input signal amplitude and the limiting value as plotted on *Figure 13.49*(*c*). There is no phase shift between input and output.

Hysteresis, shown on *Figure 13.50*, introduces a phase shift, and a flat top to the output waveform. This is not the same waveform as the limiter; the top is simply levelled off at 2*a* below the peak where *a* is half the dead zone width. If the input amplitude is large compared to the dead zone, the gain is unity and the phase shift can be approximated by

$$\phi = \sin^{-1}(a/V_i)$$

As the input amplitude decreases, the gain increases and becomes zero when the input peak to peak amplitude is

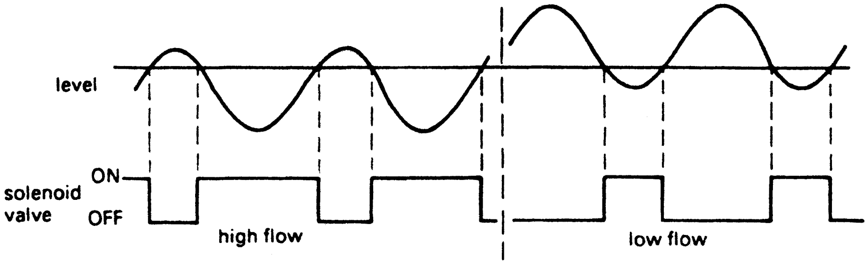


Figure 13.48 Response of level control system to changes in flow

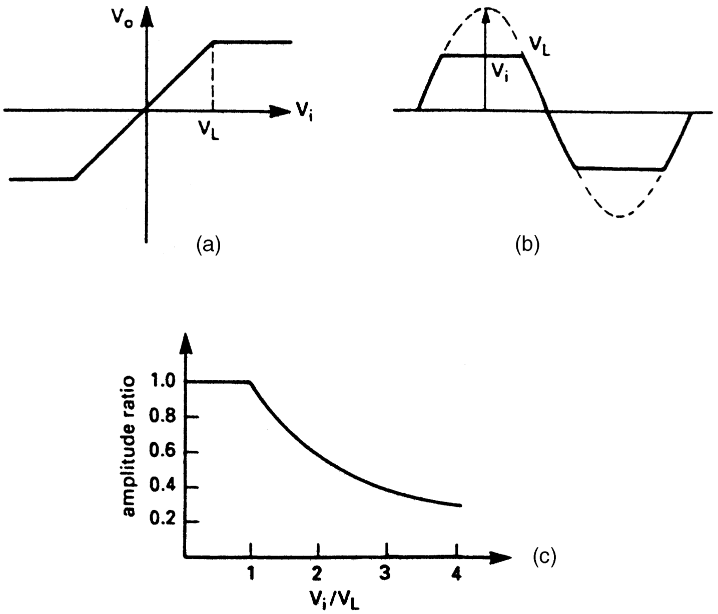


Figure 13.49 The limiter circuit: (a) relationship between input and output; (b) effect of limiting on a sine wave input signal; (c) 'Gain' of a limiter related to the input signal amplitude

less than the dead zone width. The exact relationship is complex, but is shown on *Figure 13.50*(*c*) and (*d*).

Non linear elements generally have gains and phase shift which increase or decrease with input amplitude (usually a representation of the error signal). *Figure 13.51* illustrates the two gain cases. For a loop gain of unity, constant oscillations will result. For loop gains greater than unity, oscillations will increase in amplitude, for loop gain less than unity oscillations will decay.

In *Figure 13.51*(*a*), the gain falls off with increasing amplitude. The system thus tends to approach point X as large oscillations will decay and small oscillations increase. The system will oscillate at whatever gain gives unity loop gain. This is called *limit cycling*. Most non linearities (bang-bang servo, saturation etc.) are of this form.

Where loop gain increases with amplitude as *Figure 13.51*(*b*), decreasing gain gives increasing damping as the amplitude decreases, so oscillations will quickly die away. This response is sometimes deliberately introduced into level controls. If, however, the system is provoked beyond Y by a disturbance, the oscillation will rapidly increase in amplitude and control will be lost.

13.22.3 State space and the phase plane

Figure 13.52(*a*) shows a simple position control system. The position is sensed by a potentiometer, and compared with a setpoint from potentiometer RV_1. The resulting error signal is compared with an error 'window' by comparators C_1, C_2. Preset RV_2 sets the deadband, i.e. the width of the window. The comparators energise relays RL_F and RL_R which drive the load to the forward and reverse respectively.

Initially, we shall analyse the system with RV_2 set to zero, i.e. no deadband. This has the block diagram of *Figure 13.52*(*b*), with a first order lag of time constant T arising from the inertia of the system, and the integral action converting motor velocity to load position.

The system is thus represented by

$$x = \frac{\pm K}{s(1 + sT)} \qquad (13.19)$$

where K represents the acceleration resulting from the motor torque and inertia with the sign of K indicating the sign of the error. This has the solution

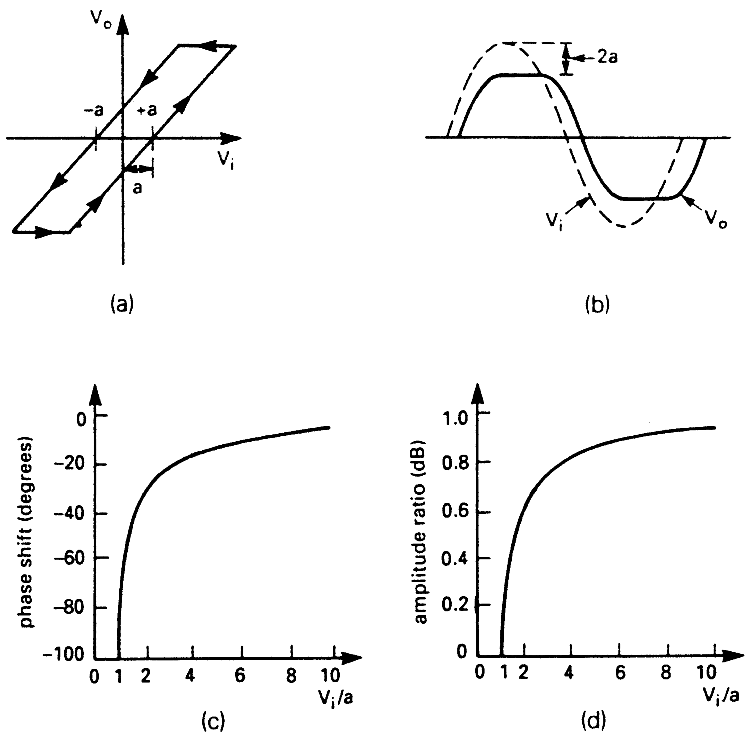


Figure 13.50 The effect of hysteresis: (a) relationship between input and output signals; (b) the effect of hysteresis on a sine wave input signal; (c) the relationship between phase shift and signal amplitude; (d) the relationship between 'gain' and signal amplitude

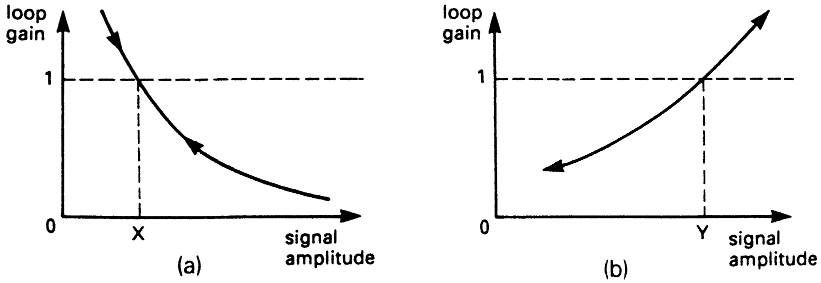


Figure 13.51 Possible relationships between gain and signal amplitude: (a) gain decreases with increasing amplitude; (b) gain increases with increasing amplitude

$$x = x_0 - \mathrm{T}K + \mathrm{T}V_0 + Kt + \mathrm{T}(K - V_0)e^{-t/\mathrm{T}} \qquad (13.20)$$

where x_0 and V_0 respectively represent the initial position and velocity.

Differentiating gives the velocity, V

$$V = K - (K - V_0)e^{-t/\mathrm{T}} \qquad (13.21)$$

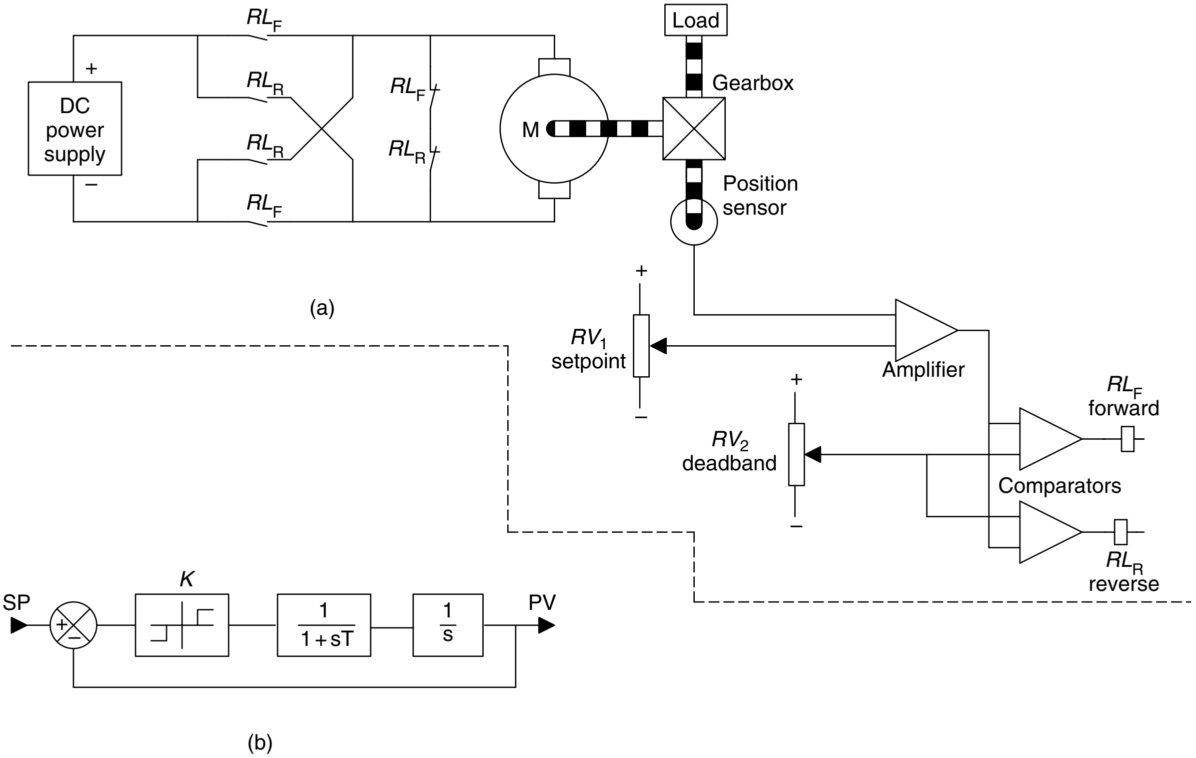
Equations 13.20 and 13.21 fully describe the behaviour of the system. These can be plotted graphically as *Figure 13.53* with velocity plotted against position for positive K for various times from $t = 0$. Each curve represents a different starting condition; curve D, for example, starts at $x_0 = -5$ and $v_0 = -2$

In each case, the curve ends towards $v = 2$ units/sec as t gets large. The family of curves have an identical shape, and the different starting conditions simply represent a horizontal shift of the curve.

A similar family of curves can be drawn for negative values of K. These are sketched on *Figure 13.54*. In this case, the velocity tends towards $V = -2$ units/sec.

Given these curves, we can plot the response of the system. Let us assume that the system is stationary at $x = -5$, and the setpoint is switched to $+5$. The subsequent behaviour is shown on *Figure 13.55*. The system starts by initially following the curve passing through $x = -5$, $V = 0$ for positive K, crossing $x = 0$ with a velocity 1.5 units/sec, reaching the setpoint at point X with a velocity of 1.76 units/sec. It cannot stop instantly however, so it overshoots.

At the instant the overshoot occurs K switches sign. The system now has a velocity of $+1.76$, with K negative, so it follows the corresponding curve of *Figure 13.54* from point X to point Y. It can be seen that an overshoot to $x = 7$ occurs. At point Y, another overshoot occurs and K switches


RL_F
RL_R
RL_R
RL_F
+
DC power supply
–
RL_F
RL_R
M
Load
Gearbox
Position sensor
+
RV_1 setpoint
–
Amplifier
+
RV_2 deadband
–
RL_F forward
Comparators
RL_R reverse
SP
+
–
K
1 / 1+sT
1 / s
PV

(a)

(b)

Figure 13.52 A non-linear position control system: (a) system diagram; (b) block representation

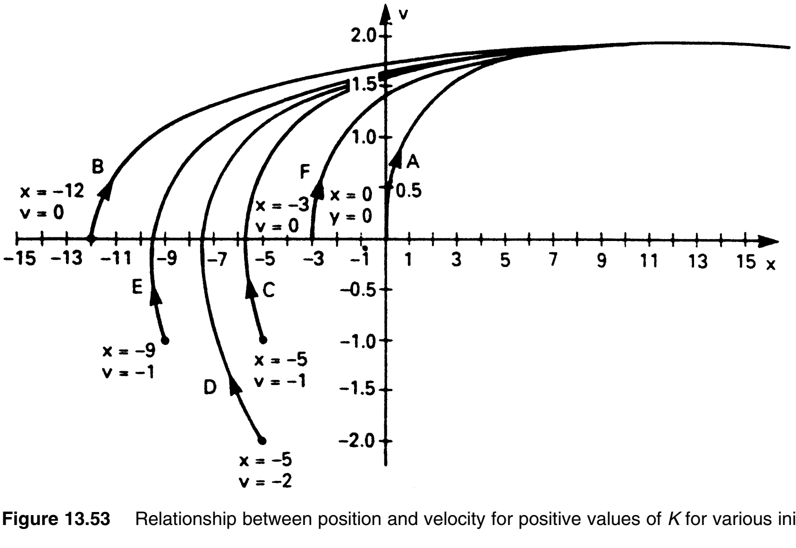


Figure 13.53 Relationship between position and velocity for positive values of *K* for various initial conditions

back positive. The system now follows the curve to Z with an undershoot of $x =$ ~~4.~~1. At Z another overshoot occurs and the system spirals inwards as shown. The predicted step response is shown on *Figure 13.56*.

In *Figure 13.57*(*a*), the deadband control (RV_2 on the earlier *Figure 13.52*) has been adjusted to energise RL_F for error voltages more negative than −1 unit and energise RL_R for error voltages above +1 unit. There is thus a deadband 2 units wide around the setpoint.

Figure 13.57(*b*) shows the effect of this deadband. We will assume initial values of $x_0 =$ ~~0,~~ $v_0 =$ ~~0~~ when we switch the setpoint to $x =$ ~~5.~~ The system accelerates to point *U* ($x =$ ~~4,~~ $v =$ ~~1.~~40) at which point RL_1 de-energises. The system loses speed ($K =$ ~~0)~~ until point *V*, where the position

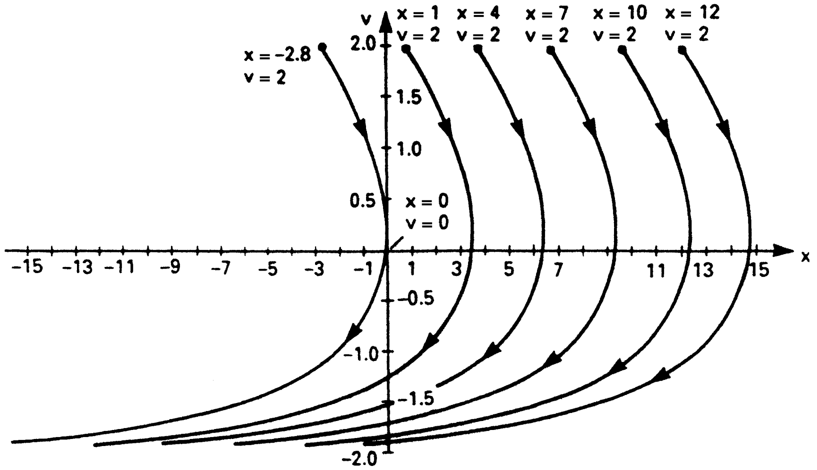


Figure 13.54 Relationship between position and velocity for negative values of *K* for various initial conditions

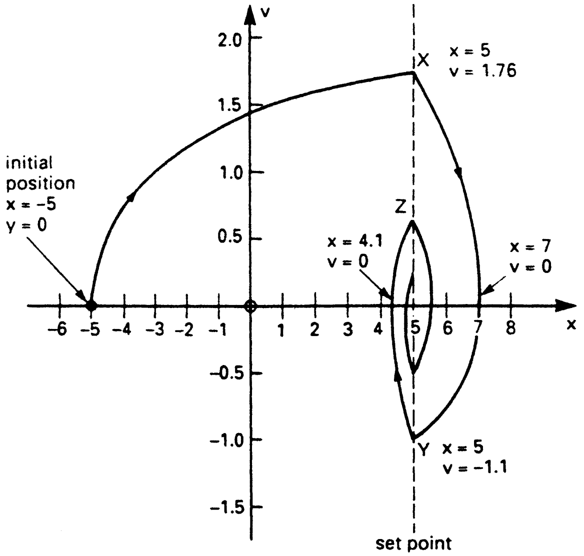


Figure 13.55 System behaviour following change of setpoint from $x = -5$ to $x = +5$

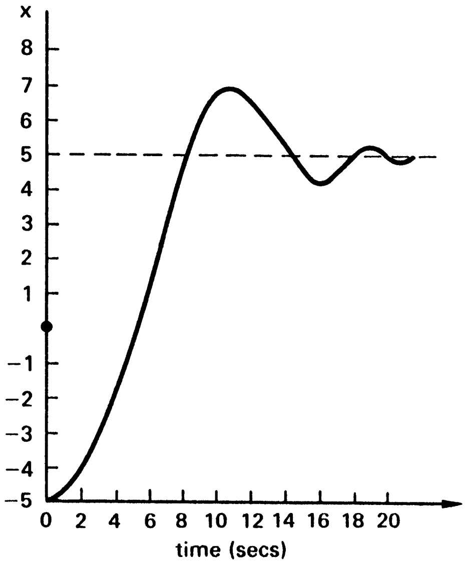


Figure 13.56 Predicted step response following change of setpoint

passes out of the deadband and RL_R energises. The system reverses, and re-enters the deadband at point *W*, where RL_F de-energises. An undershoot then occurs (*X* to *Y*) where the deadband is entered for the last time, coming to rest at point *Z* ($x = 4.75$, $v = 0$).

The position x and velocity $\mathrm{d}x/\mathrm{d}t$ completely describe the system and are known as *state variables*. A linear system can be represented as a set of first order differential equations relating the various state variables. For a second order system there are two state variables, for higher order systems there will be more.

For the system described by Equation 13.19, we can denote the state variables by x (position) and v (velocity). For a driving function *K*, we can represent the system by *Figure 13.58* which is called a *state space model*. This describes the position control system by the two first order differential equations.

$$T\frac{\mathrm{d}v}{\mathrm{d}t} = K - x$$

and

$$v = \Leftarrow \frac{\mathrm{d}x}{\mathrm{d}t}$$

Figure 13.56 and *Figure 13.57* plot velocity against position, and as such are plots relating state variables. For two state variables (from a second order system) the plot is known as a *phase plane*. For higher order systems, a multi-dimensional plot, called *state space*, is required. Plots such as *Figure 13.53* and *Figure 13.54* which show a family of possible curves are called *phase plane portraits*.

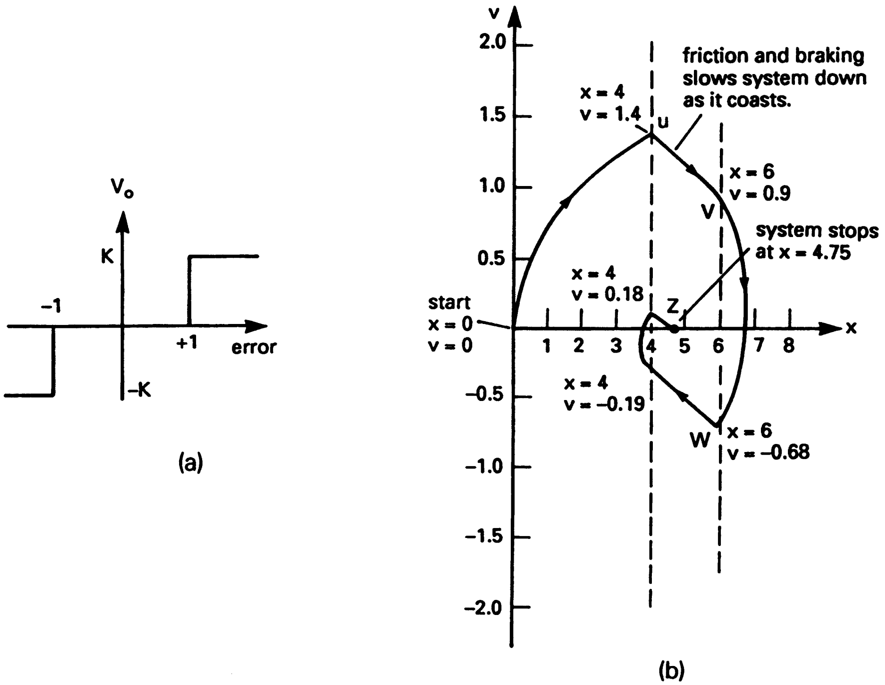


Figure 13.57 System with deadband and friction: (a) deadband response; (b) position/velocity curve for setpoint change from $x = 0$ to $x = 5$. Note system does not attain the setpoint

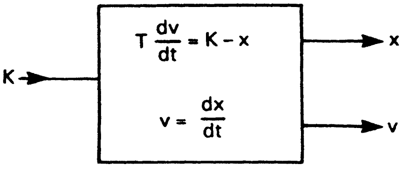


Figure 13.58 State variables for position control system

Similar phase planes can be drawn for other non linearities such as saturation, hysteresis etc. Various patterns emerge, which are summarised on *Figure 13.59*. The system behaviour can be deduced from the shape of the phase trajectory.

In a linear closed loop system stability is generally increased by adding derivative action. In a position control system this is equivalent to adding velocity (dx/dt) feedback. The behaviour of a non linear system can also be improved by velocity feedback. In *Figure 13.60*(*a*) velocity feedback has been added to our simple Bang/Bang position servo.

The switching point now occurs where

$$S_P - x - Lv = 0$$

or

$$v = \frac{1}{L}(S_P - x)$$

This is a straight line of slope $-1/L$, passing through $x = S_P$, $v = 0$ on the phase plane. Note that L has the units of time. The line is called the switching line, and advances the changeover as shown on *Figure 13.60*(*b*), thereby reducing the overshoot. Too much velocity feedback as *Figure 13.60*(*c*) simulates an overdamped system as the trajectory runs to the setpoint down the switching line.

13.23 Disturbances

13.23.1 Introduction

A closed loop control system has to deal with the malign effects of outside disturbances. A level control system, for example, has to handle varying throughput, or a gas fired furnace may have to cope with changes in gas supply pressure. Although disturbances can enter a plant at any point, it is usual to consider disturbances at two points; supply disturbances at the input to the plant and load/demand disturbances at the point of measurement as shown on *Figure 13.61*(*a*).

The closed loop block diagram can be modified to include disturbances as shown on *Figure 13.61*(*b*). A similar block diagram could be drawn for load disturbances or disturbances entering at any point by subdividing the plant block. By normal analysis we have

$$V = \frac{CPS_P}{1 + HCP} + \frac{PD}{1 + HCP} \qquad (13.22)$$

Equation 13.22 has two components; the first relates the plant output to the setpoint and is the normal closed loop transfer function $GH/(1 + GH)$. The product of controller and plant transfer function $C.P.$ is the forward gain G. The second term relates the performance of the plant to disturbance signals. In general, closed loop control reduces the effect of disturbances. If the plant was run open loop, the effect of the disturbances on the output would be simply

$$V = PD$$

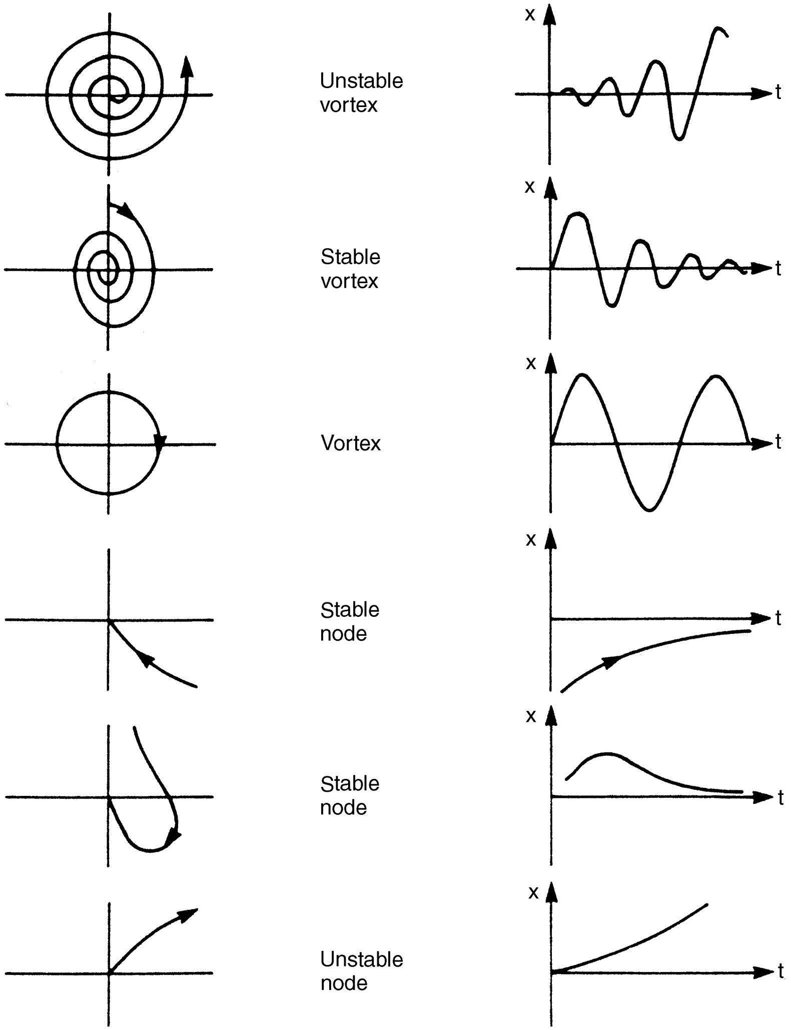


Figure 13.59 Various possible trajectories and their response

From Equation 13.22 with closed loop control, the effect of the disturbance is

$$V = \frac{PD}{1 + HCP}$$

i.e. it is reduced providing the magnitude of $(1 + HCP)$ is greater than unity. If the magnitude of $(1 + HCP)$ becomes less than unity over some range of frequencies, closed loop control will magnify the effect of disturbances in that frequency range. It is important, therefore, to have some knowledge of the frequency spectra of expected disturbances.

13.23.2 Cascade control

Closed loop control gives increased performance over open loop control, so it would seem logical to expect benefits from adding an inner control loop around plant items that are degrading overall performance. *Figure 13.62* shows a typical example, here the output of the outer loop controller becomes the setpoint for the inner controller. Any problems in the inner loop (disturbances, non linearities, phase lag etc.) will be handled by the inner controller, thereby improving the overall performance of the outer loop. This arrangement is known as *cascade control.*

To apply cascade control, there must obviously be some intermediate variable that can be measured (P_{Vi} on *Figure 13.62*) and some actuation point that can be used to control it.

Cascade control brings several benefits. The secondary controller will deal with disturbances before they can affect the outer loop. Phase shift within the inner loop is reduced, leading to increased stability and speed of response in the outer loop. Devices with inherent integral action (such as a motorised valve) introduce an inherent −90°⇐integrator phase lag. This can be removed by adding a valve positioner in cascade. Cascade control will also reduce the effect of non linearities (e.g. non linear gain, backlash etc.) in the inner loop.

There are a few precautions that need to be taken, however. The analysis so far ignores the fact that components saturate and stability problems can arise when the inner

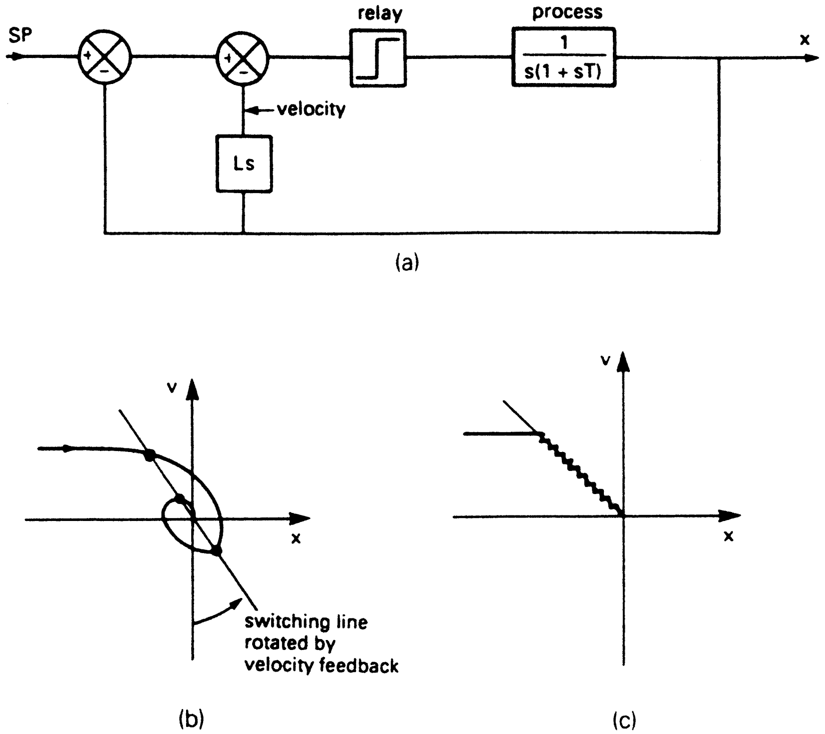


Figure 13.60 Addition of velocity feedback to a non-linear system: (a) block diagram of velocity feedback; (b) system behaviour on velocity/position curve; (c) overdamped system follows the switching line

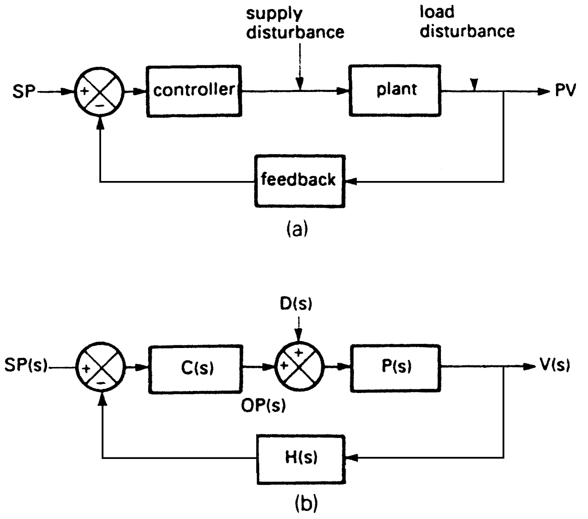


Figure 13.61 The effect of disturbances: (a) points of entry for disturbances; (b) block diagram of disturbances

loop saturates. This can be overcome by limiting the demands that the outer loop controller can place on the inner loop (i.e. ensuring the outer loop controller saturates first) or by providing a signal from the inner to the outer controller which inhibits the outer integral term when the inner loop is saturated.

The application of cascade control requires an intermediate variable and control action point, and should include, if possible, the plant item with the shortest time constant. In general, high gain proportional only control will often suffice for the inner loop, any offset is of little concern as it will be removed by the outer controller. For stability, the inner loop must always be faster than the outer loop.

Tuning a system with cascade control requires a methodical approach. The inner loop must be tuned first with the outer loop steady in manual control. Once the inner loop is tuned satisfactorily, the outer loop can be tuned as normal. A cascade system, once tuned, should be observed to ensure that the inner loop does not saturate, which can lead to instability or excessive overshoot on the outer loop. If saturation is observed, limits must be placed on the output of the outer loop controller, or a signal provided to prevent integral windup as described later in Section 13.27.6

13.23.3 Feedforward

Cascade control can reduce the effect of disturbances occurring early in the forward loop, but generally cannot deal with load/demand disturbances which occur close to, or affect directly the process variable as there is no intermediate variable or accessible control point.

Disturbances directly affecting the process variable must produce an error before the controller can react. Inevitably, therefore, the output signal will suffer, with the speed of recovery being determined by the loop response. Plants which are difficult to control tend to have low gains and long integral times for stability, and hence have a slow response. Such plants are prone to error from disturbances.

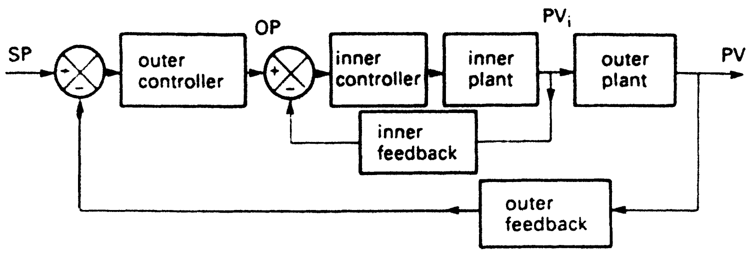


Figure 13.62 A system with cascade control

In general a closed loop system can be considered to behave as a second order system, with a natural frequency ω_n, and a damping factor. At frequencies above ω_n, the closed loop gain falls off rapidly (at 12 dB/octave). Disturbances occurring at a frequency much above $2\omega_n$ will be uncorrected. If the closed loop damping factor is less the unity (representing an underdamped system), the effect of disturbances with frequency components around ω_n can be magnified.

Figure 13.63(*a*) shows a system being affected by a disturbance. Cascade control cannot be applied because there is no intermediate variable between the point of entry and the process variable. If the disturbance can be measured, and its effect known, (even approximately), a correcting signal can be added to the controller output signal to compensate for the disturbance as shown on *Figure 13.63*(*b*). This is known as *feedforward* control.

This correcting signal, arriving by blocks H, F, and P_1 should ideally exactly cancel the original disturbance, both in the steady state and dynamically under changing conditions. The transfer functions of the transducer H and plant P_1 are fixed, with F a compensator block designed to match H and P_1.

In general, the compensator block transfer function will be

$$F = -\frac{1}{HP_1}$$

If the plant acts as a simple lag with time constant T (i.e. $1/(1+s\mathrm{T})$), the compensator will be a simple lead $(1+s\mathrm{T})$. In many cases a general purpose compensator $(1+s\mathrm{T_a})/(1+s\mathrm{T_b})$ is used.

The feedforward compensation does not have to match exactly the plant characteristics; even a rough model will give a significant improvement (although a perfect model will give perfect control). In most cases a simple compensator will suffice.

Cascade control can usually deal with supply disturbances and feedforward with load or demand disturbances. These neatly complement each other so it is very common to find a system where feedforward modifies the setpoint for the inner cascade loop.

13.24 Ratio control

13.24.1 Introduction

It is a common requirement for two flows to be kept in precise ratio to each other; gas/oil and air in combustion control, or reagants being fed to a chemical reactor are typical examples.

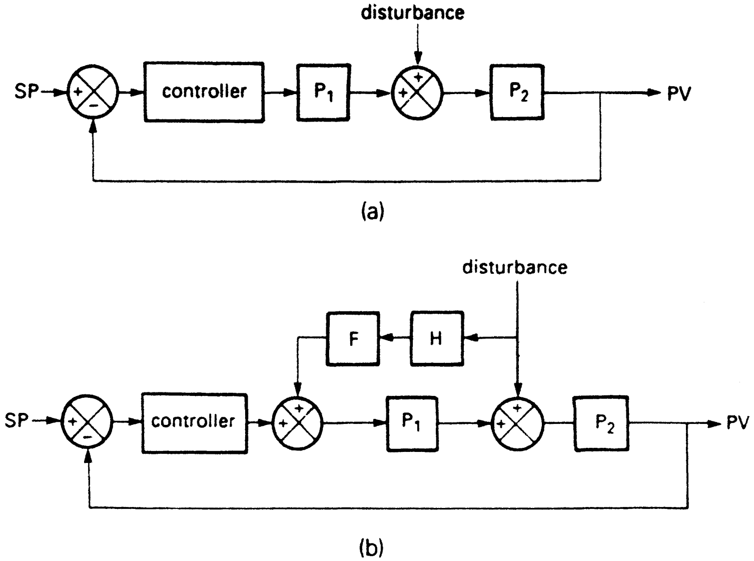


Figure 13.63 Effect of a disturbance reduced by feedforward: (a) a system to which cascade control cannot be applied being subject to a disturbance; (b) correcting signal derived by measuring the disturbance

13.24.2 Slave follow master

In simple ratio control, one flow is declared to be the master. This flow is set to meet higher level requirements such as plant throughput or furnace temperature. The second flow is a slave and is manipulated to maintain the set flow ratio.

The controlled variable here is ratio, not flow, so an intuitive solution might look similar to *Figure 13.64* where the actual ratio A/B is calculated by a divider module and used as the process variable for a controller which manipulated the slave control valve.

This scheme has a hidden problem. The slave loop includes the divider module and hence the term A. The loop gain varies directly with the flow A, leading to a sluggish response at low flows and possible instability at high flow. If the inverse ratio B/A is used as the controller variable the saturation becomes worse as the term $1/A$ now appears in the slave loop giving a loop gain which varies inversely with A, becoming very high at low flows. Any system based on *Figure 13.64* would be impossible to tune for anything other than constant flow rates.

Ratio control systems are often based on *Figure 13.65*. The master flow is multiplied by the ratio to produce the setpoint for the slave flow controller. The slave flow thus follows the master flow. Note that in the event of failure in the master loop (a jammed valve for example) the slave controller will still maintain the correct ratio.

The slave flow will tend to lag behind the master flow. On a gas/air burner, the air flow could be master and the gas loop the slave. Such a system would run lean on increasing heat and run rich on decreasing heat. To some extent this can be overcome by making the master loop slower acting than the slave loop, possibly by tuning.

In a ratio system, a choice has to be made for master and slave loops. The first consideration is usually safety. In a gas/air burner, for example, air master/gas slave (called *gas follow air*) is usually chosen as most failures in the air loop cause the gas to shut down. If there are no safety considerations, the slowest loop should be the master and the fastest loop the slave to overcome the lag described above. Since 'fuel' (in both combustion and chemical terms) is usually the smallest flow in a ratio system and consequently has smaller valves/actuators, the safety and speed requirement are often the same.

The ratio block is a simple multiplier. If the ratio is simply set by an operator this can be a simple potentiometer acting

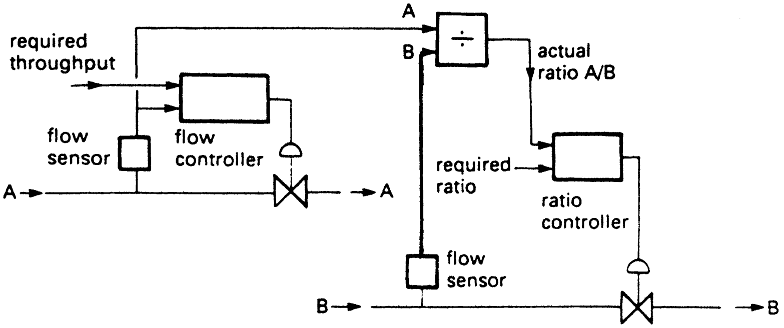


Figure 13.64 An intuitive, but incorrect, method of ratio control. The loop gain varies with throughput

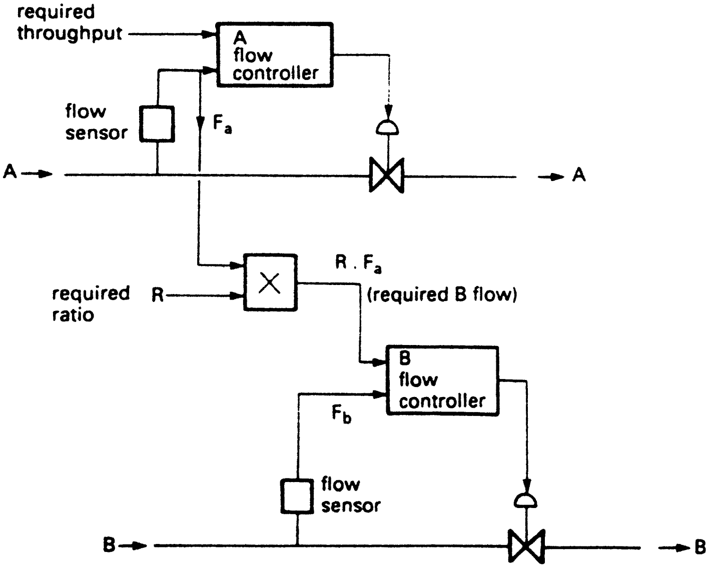


Figure 13.65 Master/slave ratio system with stable loop gains

as a voltage divider (for ratios less the unity) or an amplifier with variable gain (for ratios greater than unity). In digital control systems, of course, it is a simple multiply instruction. If the ratio is to be changed remotely (a trim control from an automatic sampler on a chemical blending system for example) a single quadrant analog multiplier is required.

Ratio blocks are generally easier to deal with in digital systems working in real engineering units. True ratios (an air/gas ratio of 10/1 for example) can then be used. In analog systems the range of the flow meters needs to be considered. Suppose we have a master flow with FSD of 12000 l/min, a slave flow of FSD 2000 l/min and a required ratio (master/slave) of 10/1. The required setting of R on *Figure 13.65* would be 0.6. In a well designed plant with correctly sized pipes, control valves and flow meters, analog ratios are usually close to unity. If not, the plant design should be examined.

Problems can arise with ratio systems if the slave loop saturates before the master. A typical scenario on a gas follow air burner control could go; the temperature loop calls for a large increase in heat (because of some outside influence). The air valve (master) opens fully, and the gas valve follows correctly but cannot match the requested flow. The resulting flame is lean and cold (flame temperature falls off rapidly with too lean a ratio) and the temperature does not rise. The system is now locked with the temperature loop demanding more heat and the air/gas loops saturated, delivering full flow but no temperature rise. The moral is; the master loop must saturate before the slave. If this is not achieved by pipe sizing the output of the master controller should be limited.

13.24.3 Lead lag control

Slave follow Master is simple, but one side effect is that the mixture runs lean for increasing throughput and rich for decreasing throughput because the master flow must always change first before the slave can follow. There is also a possible safety implication because a failure of the slave valve or controller could lead to a gross error in the actual ratio such as the fuel valve wide open and the air valve closed.

Better performance can be obtained with a system called *Lead-lag control* shown for an air/fuel burner on *Figure 13.66*. This uses cross linking and selectors to provide an air setpoint which is the highest of the external power demand signal or ratio'd fuel flow. The fuel setpoint is the lowest of the external power demand or ratio'd air flow.

This cross linking provides better ratio during changes, both air and fuel will change together. There is also higher security; a jammed open fuel valve will cause the air valve to open to maintain the correct ratio and prevent an explosive atmosphere of unburned fuel forming.

13.25 Transit delays

13.25.1 Introduction

Transit delays are a function of speed, time and distance. A typical example from the steel industry is the tempering process of *Figure 13.67* where red hot rolled steel travelling at 15 m/sec is quenched by passing beneath high pressure water sprays. The recovery temperature, some 50 m downstream, is the controlled variable which is measured by a pyrometer and used to adjust the water flow control valve. There is an obvious transit delay of 50/15 = ~~3.~~3 sec in the loop. A transit delay is a simple time shift which is independent of frequency.

Transit delays give an increasing phase shift with rising frequency which is de-stabilising. If conventional controllers are used significant detuning (low gain, large T_i) is necessary to maintain stability. The effect is shown on the

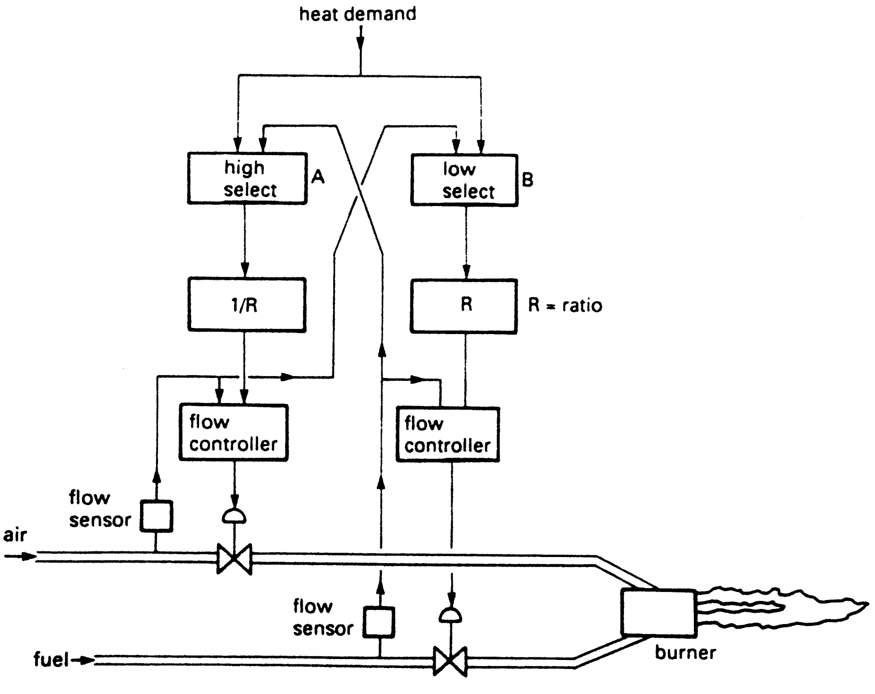


Figure 13.66 Lead/lag combustion control

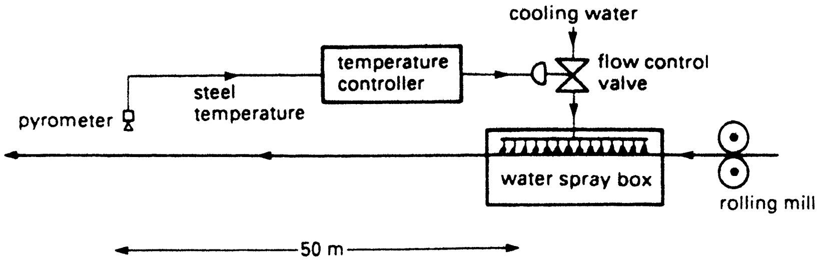


Figure 13.67 A tempering system dominated by a transit delay

Nichols charts of *Figure 13.68* for a simple system of two first order lags controlled by a PI controller. The de-stabilising effect of the increasing phase shift can clearly be seen. Derivative action, normally a stabilising influence, can also adversely affect a loop in which a transit delay is the dominant feature.

13.25.2 The Smith predictor

The effects of a transit delay can be reduced by the arrangement of *Figure 13.69* called a *Smith predictor*. The plant is considered to be an ideal plant followed by a transit delay. (This may not be true, but the position of the transit delay, before or after the plant, makes no difference to the plant behaviour.) The plant and its associated delay are modelled as accurately as possible in the controller.

The controller output, O_P, is applied to the plant and to the internal controller model. Signal *A* should thus be the same as the notional (and unmeasurable) plant signal *X*, and the signal *B* should be the same as the measurable controlled variable signal *Y*.

The PID controller however, is primarily controlling the model, not the plant, via summing junction 1. There are no delays in this loop, so the controller can be tuned for tight operation. With the model being the only loop, however, the plant is being operated in open loop control, and compensation will not be applied for model inaccuracies or outside disturbances.

Signal *Y* and *B* are therefore compared by a subtractor to give an error signal which encompasses errors from both disturbances and the model. These are added to the signal *A* from the plant model to give the feedback signal to the PID control block.

Discrepancies between the plant model and the real plant will be compensated for in the outer loop, so exact modelling is not necessary. The poorer the model, however, the less tight the control that can be applied in the PID block as the errors have to be compensated via the plant transit delay.

Smith predictors are usually implemented digitally, analog transit delays being difficult to construct. A digital delay line is simply a shift register in which values are shifted one place at each sample.

The Smith predictor is not a panacea for transit delays; it still takes the delay time from a setpoint change to a change in the process variable, and it still takes the delay time for a disturbance to be noted and corrected. The response to change, however, is considerably improved.

Systems with transit delays can benefit greatly from feedforward described previously in Section 13.23.3. Feedforward used in conjunction with a Smith predictor can be a very effective way of handling control systems with significant transit delays.

13.26 Stability

13.26.1 Introduction

At first sight it would appear that perfect control can be obtained by utilising a large proportional gain, short integral time and long derivative time. The system will then respond quickly to disturbances, alterations in load and set point changes.

Unfortunately life is not that simple, and in any real life system there are limits to the settings of gain T_i and T_d beyond which uncontrolled oscillations will occur. Like many engineering systems, the setting of the controller is a compromise between conflicting requirements.

13.26.2 Definitions and performance criteria

It is often convenient, (and not too inaccurate), to consider that a closed loop system behaves as a second order system, with a natural frequency ω_n and a damping factor β.

$$\frac{d^2x}{dt^2} + 2\beta\omega_n\frac{dx}{dt} + \omega_n^2 x = f(t) \Leftarrow$$

It is then possible to identify five possible performance conditions, shown for a set point change and a disturbance in *Figure 13.70*(*a*) and (*b*).

An unstable system exhibits oscillations of increasing amplitude. A marginally stable system will exhibit constant amplitude oscillations. An underdamped system will be somewhat oscillatory, but the amplitude of the oscillations decreases with time and the system is stable. (It is important to appreciate that oscillatory does not necessarily imply instability). The rate of decay is determined by the damping factor. An often used performance criteria is the '*quarter amplitude damping*' of *Figure 13.70*(*c*) which is an underdamped response with each cycle peak one quarter of the amplitude of the previous. For many applications this is an adequate, and easily achievable response.

An overdamped system exhibits no overshoot and a sluggish response. A critical system marks the boundary between underdamping and overdamping and defines the fastest response achievable without overshoot.

For a simple system the responses of *Figure 13.70*(*a*) and (*b*) can be related to the gain setting of a *P* only controller, overdamped corresponding to low gain with increasing gain causing the response to become underdamped and eventually unstable.

It is impossible for any system to respond instantly to disturbances and changes in set point. Before the adequacy of a control system can be assessed, a set of performance criteria is usually laid down by production staff. Those defined in *Figure 13.71* are commonly used.

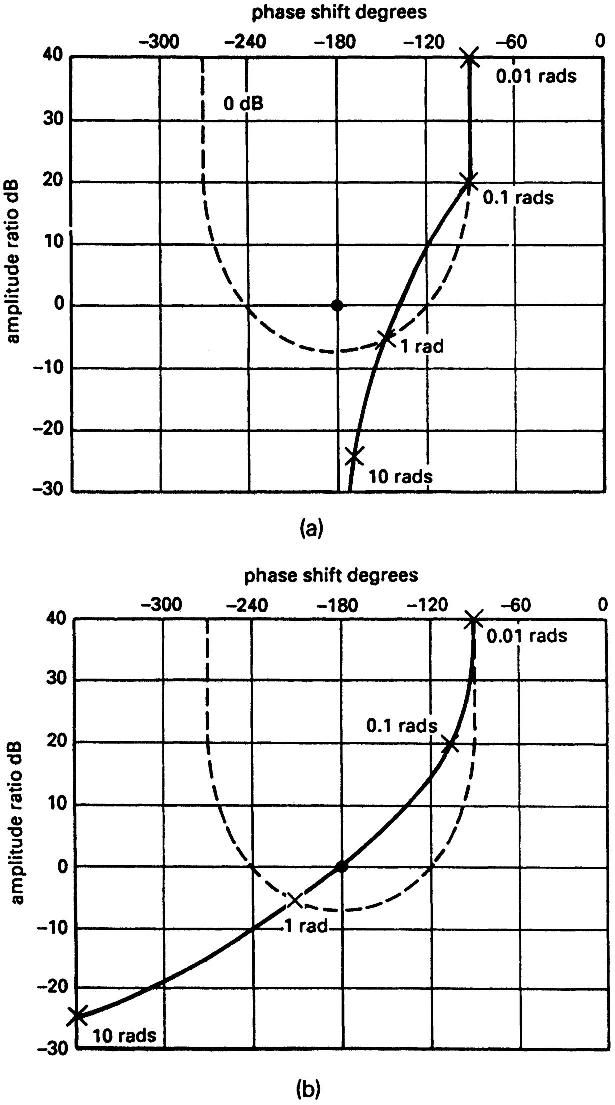


Figure 13.68 The effect of a transit delay on stability: (a) sketch of a Nichols chart for a system comprising a PI controller (K=~~5~~, T_i=~~5~~s) and two first order lags of time constants 5 secs and 2 secs. The system is unconditionally stable; (b) the same system with a one second transit delay. The transit delay introduces a phase shift which increases with rising frequency and makes the system unstable

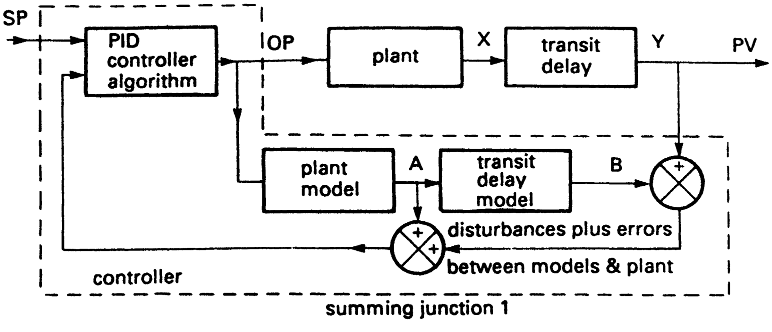

SP
PID controller algorithm
OP
plant
X
transit delay
Y
PV
plant model
A
transit delay model
B
+
+
+
disturbances plus errors
between models & plant
controller
summing junction 1

Figure 13.69 The Smith predictor used to reduce the effect of transit delays

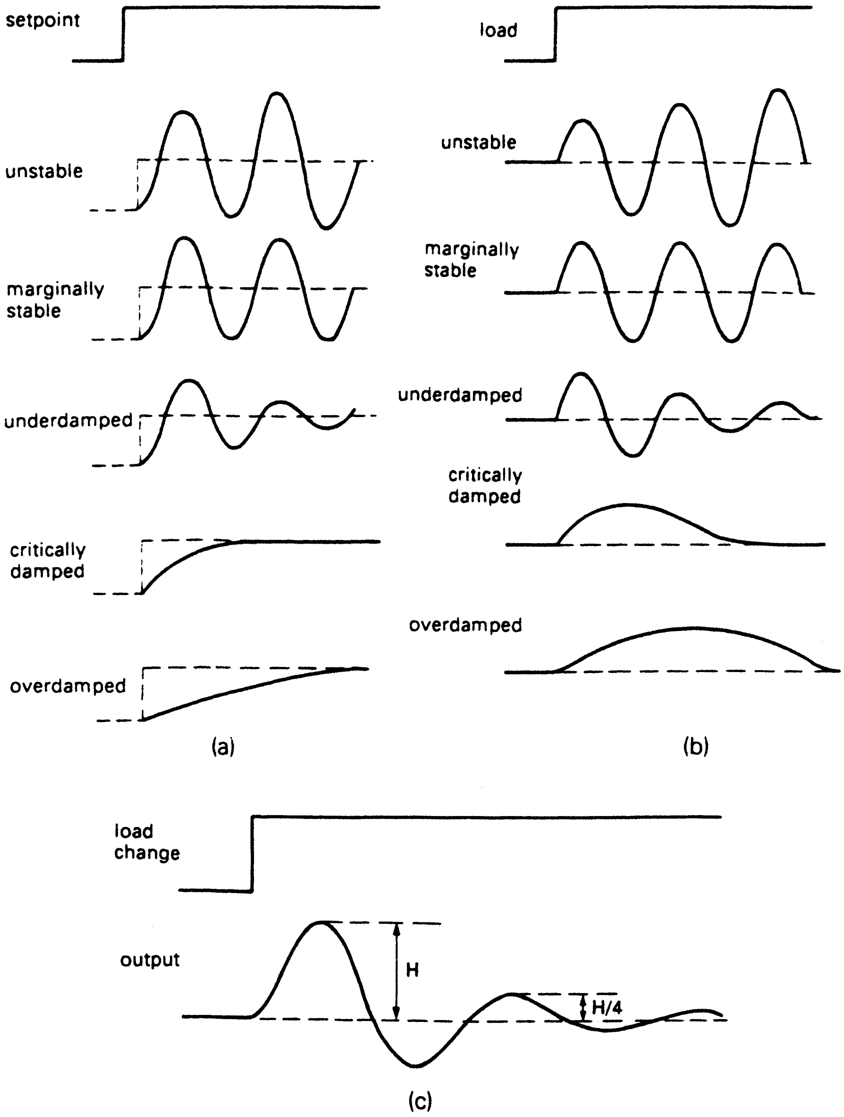

setpoint
unstable
marginally stable
underdamped
critically damped
overdamped
(a)
load
unstable
marginally stable
underdamped
critically damped
overdamped
(b)
load change
output
H
H/4
(c)

Figure 13.70 Various forms of system response: (a) step change in setpoint; (b) step change in load; (c) quarter amplitude damping

The '*rise time*' is the time taken for the output to go from 10% to 90% of its final value, and is a measure of the speed of response of the system. The time to achieve 50% of the final value is called the '*delay time*'. This is a function of, but not the same as, any transit delays in the system. The first overshoot is usually defined as a percentage of the corresponding set point change, and is indicative of the damping factor achieved by the controller.

As the time taken for the system to settle completely after a change in set point is theoretically infinite, a '*settling band*', '*tolerance limit*' or '*maximum error*' is usually defined. The settling time is the time taken for the system to enter, and remain within, the tolerance limit. Surprisingly an underdamped system may have a better settling time than a critically damped system if the first overshoot is just within the settling band. *Table 13.2* shows optimum damping factors for various settling bands. The settling time is defined in units of $1/\omega_n$.

Table 13.2

Settling band	*Optimum 'b'*	*Settling time*
20%	0.45	1.80
15%	0.55	2.00
10%	0.60	2.30
5%	0.70	2.80
2%	0.80	3.50

The shaded area is the integral of the error and this can also be used as an index of performance. Note that for a system with a standing offset (as occurs with a *P* only controller) the area under the curve will increase with time and not converge to a final value. Stable systems with integral action control have error areas that converge to a finite value. The area

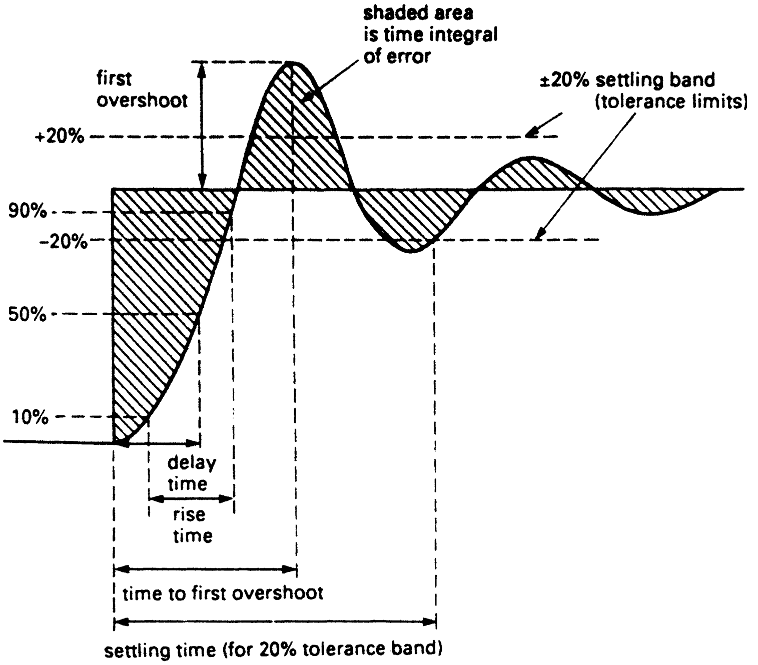


Figure 13.71 Common definitions of system response

between the curve and the set point is called the *integrated absolute error* (IAE) and is an accepted performance criterion.

An alternative criterion is the integral of the square of the instantaneous error. This weights large errors more than small errors, and is called *integrated squared error* (ISE). It is used for systems where large errors are detrimental, but small errors can be tolerated.

The performance criteria above were developed for a set point change. Similar criteria can be developed for disturbances and load changes.

13.26.3 Methods of stability analysis

The critical points for stability are open loop unity gain and a phase shift of −180°. It is therefore reasonable to give two figures of 'merit':

(a) The *Gain Margin* is the amount by which the open loop gain can be increased at the frequency at which the phase shift is −180°. It is simply the inverse of the gain at this critical frequency, for example if the gain at the critical frequency is 0.5, the gain margin is two.

(b) The *Phase Margin* is the additional phase shift that can be tolerated when the open loop gain is unity. With −140° phase shift at unity gain, there is a phase margin of 40°.

For a reasonable, slightly underdamped, closed loop response the gain margin should be of the order of 6–12 dB and the phase margin of the order of 40–65°.

Any closed loop control system can be represented by *Figure 13.72* where G is the combined block transfer function of the controller and plant and H the transfer function of the transducer and feed back components. The output will be given by:

$$P_V = \frac{G}{1 + GH} S_p$$

The system will be unstable if the denominator goes to zero or reverses in sign, i.e. $GH < -1$. This is not as simple a relationship as might be first thought, as we are dealing with the dynamics of the process. The response of the system (gain and phase shift) will vary with frequency; generally the gain will fall and the phase shift will rise with increasing frequency. A phase shift of 180° corresponds to multiplying a sine wave by −1, so if at some frequency the phase shift is 180° and the gain at that frequency is greater than unity the system will be unstable.

There are several methods of representing the gain/phase shift relationship, and inferring stability from the plot. *Figure 13.73* is called a Bode diagram and plots the gain (in dB) and phase shift on separate graphs. Log-Lin graph paper

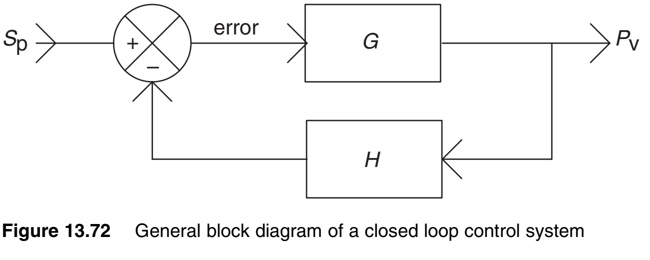


Figure 13.72 General block diagram of a closed loop control system

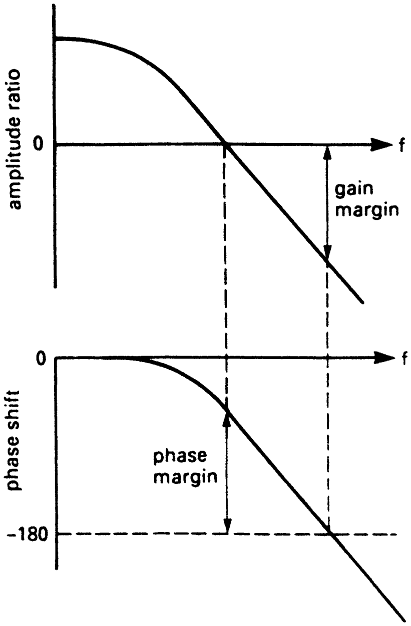


Figure 13.73 Gain and phase margins on the Bode diagram

(e.g. Chartwell 5542) is required. For stability, the gain curve must cross the 0 dB axis before the phase shift curve crosses the 180° line. From these two values, the gain margin and the phase margin can be read as shown.

Figure 13.74 is a Nichols chart and plots phase shift against gain (in dB). For stability, the 0 dB/ −180° intersection must be to the right of the curve for increasing frequency. Nichols charts are plotted on pre-printed graph paper (Chartwell 7514 for example) which allows the closed loop response to be read directly. If for example the curve is inside the closed loop 0 dB line damped oscillations will result. The gain and phase margins can again be read from the graph.

The final method is the Nyquist diagram of *Figure 13.75*. This plots gain again phase shift as a polar diagram (gain represented by distance from the origin). Chartwell graph paper 4001 is suitable. For stability the −180° point must be to the left of the graph for increasing frequency. Gain and phase margin can again be read from the graph.

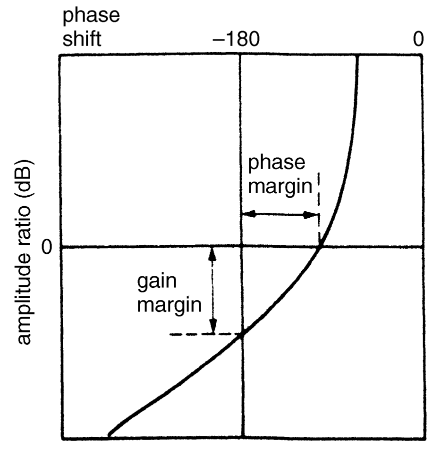


Figure 13.74 Gain and phase margins on a Nichols chart

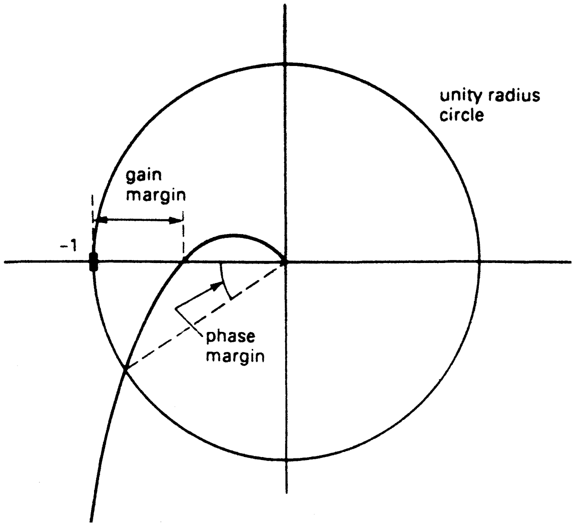


Figure 13.75 Gain and phase margins on a Nyquist diagram

13.27 Industrial controllers

13.27.1 Introduction

The commercial three term controller is the workhorse of process control and has evolved to an instrument of great versatility. This section describes some of the features of practical modern microprocessor based controllers.

13.27.2 A commercial controller

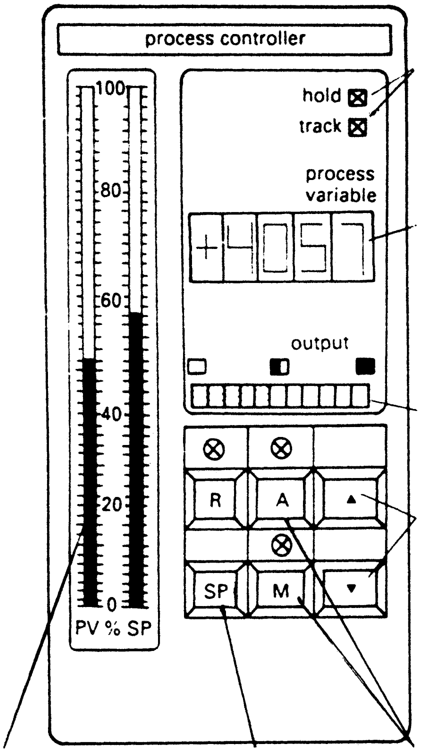
The description in this section is based on the 6360 controller manufactured by Eurotherm Process Automation Ltd of Worthing, Sussex.

The controller front panel is the 'interface' with the operator who may have little or no knowledge of process control. The front panel controls should therefore be simple to comprehend. *Figure 13.76* shows a typical layout.

The operator can select one of three operating modes—manual, automatic or remote—via the three push buttons labelled M, A, R. Indicators in each push-button show the current operating mode.

In manual mode, the operator has full control over the driven plant actuator. The actuator drive signal can be ramped up or down by holding in the M button and pressing the ▲ or ▼ buttons. The actuator position is shown digitally on the digital display, whilst the M button is depressed and continuously in analog form on the horizontal bargraph.

In automatic mode the unit behaves as a three term controller with a set point loaded by the operator. The unit is scaled into engineering units (i.e. real units such as °C, psi, litres/min) as part of the set up procedure so that the operator is working with real plant variables. The digital display shows the set point value when the SP button is depressed and the value can be changed with the ▲ and ▼ buttons.



status LEDs
hold: controller output frozen
track: controller output following external signal

digital display:
(i) PV default display (in engineering units)
(ii) output if M, A or R held pressed
(iii) setpoint if SP held pressed

controller output in bargraph form

raise lower PBs for
(i) output (in manual)
(ii) setpoint (in auto with SP pressed)

setpoint and process variable in bargraph form

setpoint used with ▲ ▼ PBs to adjust SP

mode select PBs: auto, manual, remote auto LEDs show selected mode

Figure 13.76 Front panel operator controls on a typical controller

The set point is also displayed in bargraph form on the right-hand side of the dual vertical bargraph.

Remote mode is similar to automatic mode except the set point is derived from an external signal. This mode is used for ratio or cascade loops (see Sections 13.23 and 13.24) and batch systems where the setpoint has to follow a predetermined pattern. As before the setpoint is displayed in bargraph form and the operator can view, but not change, the digital value by depressing the SP button.

The process variable itself is displayed digitally when no push button is depressed, and continually on the left-hand bargraph. In automatic or remote modes the height of the two left-hand bargraphs should be equal, a very useful quick visual check that all is under control.

Alarm limits, (defined during the controller set up), can be applied to the process variable or the error signal. If either move outside acceptable limits, the process variable bargraph flashes, and a digital output from the controller is given for use by an external annunciator audible alarm of data logger.

Figure 13.77 shows a simple block diagram representation of a controller.

Input analog signals enter at the left-hand side. Common industrial signal standards are 0–10 V, 1–5 V, 0–20 mA and 4–20 mA. These can be accommodated by two switchable ranges 0–10 V and 1–5 V plus suitable burden resistors for the current signals (a 250 ohm resistor, for example, converts 4–20 mA to 1–5 V).

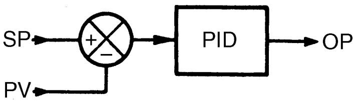


Figure 13.77 Block diagram of a typical controller

4–20 mA and 0–20 mA signals used on two wire loops require a DC power supply somewhere in the loop. A floating 30 V power supply is provided for this purpose.

Open circuit detection is provided on the main P_V input. This is essentially a pull up to a high voltage via a high value resistor. A comparator signals an open circuit input when the voltage rises. Short circuit detection can also be applied on the 1–5 V input (the input voltage falling below 1 V). Open circuit or short circuit P_V is usually required to bring up an alarm and trip the controller to manual, with the output signal driven high, held at last value, or driven low according to the nature of the plant being controlled. The open circuit trip mode is determined by switches as part of the set up procedure.

The P_V and remote S_P inputs are scaled to engineering units and linearised. Common linearisation routines are thermocouples, platinum resistance thermometers and square root (for flow transducers). A simple adjustable first order filter can also be applied to remove process or signal noise. The set point for the PID algorithm is selected from the internal set point or the remote set point by the from panel auto and remote push button contacts A, R.

The error signal is obtained by a subtractor (P_V and S_P both being to the same scale as a result of the scaling and engineering unit blocks). At this stage two alarm functions are applied. An absolute input alarm provides adjustable high and low alarm limits on the scaled and linearised P_V signals, and a deviation alarm (with adjustable limits) applied to the error signal. These alarm signals are brought out of the controller as digital outputs.

The basic PID algorithm is implemented digitally and includes a few variations to deal with some special circumstances. These modifications utilise the additional signals to the PID block (P_V, hold, track, output balance) and are described later.

The PID algorithm output is the actuator drive signal scaled 0–100%. The PID algorithm assumes that an increasing drive signal causes an increase in P_V. Some actuators, however, are reverse acting, with an increasing drive signal reducing P_V. A typical example is cooling water valves which are designed to fail open delivering full flow on loss of signal. Before the PID algorithm can be used with reverse acting actuators (or reverse acting transducers) its output signal must be reversed. A set up switch selects normal or inverted PID output. Note that reverse action does not alter the polarity of the controller output, merely the sign of the gain.

The output signal is selected from the manual raise/lower signal or the PID signal by the front panel manual/auto/remote pushbuttons M, A, R. At this stage limits are applied to the selected output drive. This limiting can be used to constrain actuators to a safe working range. The output limit allows the controller output to be limited just before the actuator's ends of travel, keeping the P_V under control at all times.

Two controller outputs are provided, 0–10 V and 4–20 mA for use with voltage and current driven actuators. The linearised P_V signal is also retransmitted as a 0–10 V signal for use with the separate external indicators and recorders.

13.27.3 Bumpless transfer

The output from the PID algorithm is a function of time and the values of the set point and the process variable. When the controller is operating in manual mode it is highly unlikely that the output of the PID block will be the same as the demanded manual output. In particular the integral term will probably cause the output from the PID block to eventually saturate at 0% or 100% output.

If no precautions are taken, therefore, switching from auto to manual, then back to auto again some time later will result in a large step change in controller output at the transition from manual to automatic operation.

To avoid this 'bump' in the plant operation, the controller output is fed back to the PID block, and used to maintain a PID output equal to the actual manual output. This balance is generally achieved by adjusting the contribution from the integral term.

Mode switching can now take place between automatic and manual modes without a step change in controller output. This is known as *manual/auto balancing*, *preload* or (more aptly) *bumpless transfer*.

A similar effect can occur on set point changes. With a straightforward PID algorithm, a setpoint change of ΔS_P will produce an immediate change in controller output of $K\cdot\Delta S_P$ where K is the controller gain. In some applications this step change in output is unacceptable. In *Figure 13.78* a term $K\cdot S_P$ is subtracted from the PID block output. The controller now responds to errors caused by changes in P_V in the normal way, but only reacts to changes in S_P via the integral and derivative terms. Changes in S_P thus result in a slow change in controller output. This is known as *setpoint change balance*, and is a switch selectable set up option.

This balance signal fed back from the output to the PID block is also used when the controller output is forced to follow an external signal. This is called *track mode*.

As before, the PID algorithm needs to be balanced to avoid a bump when transferring between track mode and automatic mode. The feedback output signal achieves this balance as described previously.

13.27.4 Integral windup and desaturation

Large changes in S_P or large disturbances to P_V can lead to saturation of the controller output or a plant actuator. Under these conditions the integral term in the PID algorithm can cause problems.

Figure 13.79 shows the probable response of a system with unrestricted integral action. At time A a step change in set point occurs. The output O_P rises first in a step (K × set point change) then rises at a rate determined by the integral time. At time B the controller saturates at 100% output, but the integral term keeps on rising.

At the time C P_V reaches, and passes, the required value, and as the error changes sign the integral term starts to decrease, but it takes until time D before the controller desaturates. Between times B and D the plant is uncontrolled, leading to an unnecessary overshoot and possibly even instability.

This effect is called *'integral windup'* and is easily avoided by disabling the integral term once the controller saturates either positive or negative. This is naturally a feature of all commercial controllers, but process control engineers should always by suspicious of 'home brew' control algorithms constructed (or written in software) by persons without control experience.

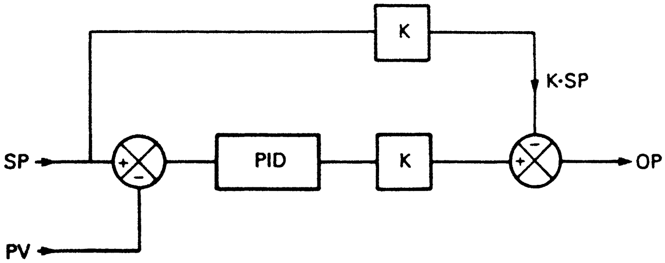


Figure 13.78 Set point change balance, the controller only follows set point changes on the integral term giving a ramped response to a change of set point

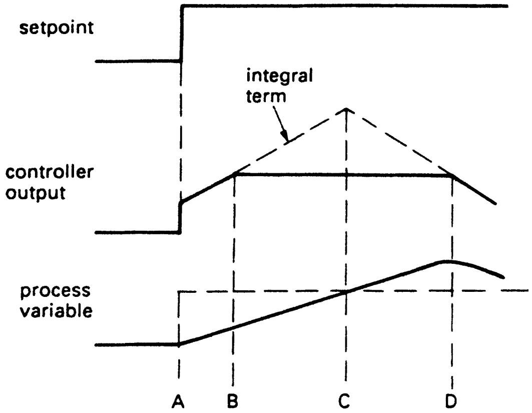


Figure 13.79 The effect of integral windup

In any commercial controller, the integral term would be disabled at point B on *Figure 13.80* to prevent integral windup. The obvious question now is at what point it is re-enabled again. Point C is obviously far too late (although much better than point D in the unprotected controller).

A common solution is to desaturate the integral term at the point where the rate of increase of the integral action equals the rate of decrease of the proportional and derivative terms. This occurs when the slope of the PID output is zero, i.e. when

$$e = -T_i\left(\frac{de}{dt} + T_d\frac{d^2e}{dt^2}\right) \tag{13.23}$$

with e being the error and T_i and T_d the controller constants.

Equation 13.23 brings the controller out of saturation at the earliest possible moment, but this can, in some cases, be too soon leading to an unnecessarily damped response. Some controllers allow adjustment of the desaturation point by adding an error limit circuit to delay the balance point to Equation 13.23 forcing the controller to remain in saturation for a longer time. The speed of desaturation and the degree of overshoot can thus be adjusted by the commissioning engineer.

13.27.5 Selectable derivative action

The term T_d (de/dt) in the three term controller algorithm can be rearranged as

$$T_d\left(\frac{dS_P}{dt} - \frac{dP_V}{dt}\right)$$

where S_P is the set point and P_V the process variable. The derivative term thus responds to changes in both the set point and the plant feedback signal.

This is not always desirable; in particular a step change in set point leads to an infinite spike controller output and a vicious 'kick' to the actuator. Commercial controller therefore include a selectable option for the derivative term to be based on true error (S_P–P_V) or purely on the value of P_V alone.

There is generally no noticeable difference in plant performance between these options; stability or the ability to deal with disturbances or load changes are unaffected, and derivative on P_V is normally the preferred choice. The only occasion when true derivative on error is advantageous is where the P_V is required to track a continually changing S_P.

13.27.6 Variations on the PID algorithm

The theoretical PID algorithm is described by the equation

$$O_P = K\left(e + \frac{1}{T_i}\int e\,dt + T_d\frac{de}{dt}\right)$$

where e is the error, K is the gain, T_i is the integral time and T_d the derivative time. Unfortunately different manufacturers use different terminology and even different algorithms.

Many manufacturers define the gain as the *proportional band*, denoted as *PB* or P_B. This is the inverse of the gain expressed as a percentage, i.e

$$P_B = \frac{100}{K}\%$$

A gain of two is thus the same as a proportional band of 50%, and decreasing the proportional band increases the gain.

The integral time is commonly expressed as '*Repeats per Minute*' or rpm. The relationship is given by:

$$\begin{aligned}\text{Repeats per min} &= 1/T_i \text{ (for } T_i \text{ in min)}\\ &= 60/T_i \text{ (for } T_i \text{ in sec)}\end{aligned}$$

The derivative time is often called the *rate* or *pre-act* term but these are all identical to T_d.

More surprisingly there are variations on the basic algorithm. Some manufacturers use a so called '*non interacting*' or '*parallel*' equation which can be expressed as:

$$O_P = K_e + \frac{1}{T_i}\int e\,dt + T_d\frac{de}{dt}$$

or

$$O_P = K_e + K_i\int e\,dt + K_d\frac{de}{dt}$$

In these the three terms are totally independent. In the second version K_i is called the integral gain and K_d the derivative gain. Note that increasing K_i has the same effect as decreasing T_i. It is tempting to think that the non interacting equations are simpler to use, but in practice the theoretical model is more intuitive. In particular, as the gain K is reduced in the non interacting equation, any integral action has more effect and contributes more phase shift. Increasing or decreasing the gain with a non interacting controller can thus cause instability.

There is yet a third form of PID algorithm known as the '*series*' equation. This can be expressed as:

$$O_P = K\left(1 + \frac{1}{T_i}\int e\,dt\right)\left(1 + T_d\frac{de}{dt}\right)$$

This algorithm is based on pneumatic and early electronic controllers, and some manufacturers have maintained it to give backward compatibility. This has the odd characteristic that the T_i and T_d controls interact with each other, with the maximum derivative action occurring when T_d and T_i are set equal. In addition the ratio between T_i and T_d interacts with the overall gain.

There are further variations on the way the derivative contribution is handled. We have already discussed the effect of derivative on process variable and derivative on error. Because the pure derivative term gives increasing gain with increasing frequency it amplifies any high frequency noise resulting in continual twitchy movements of the plant actuators. Many manufacturers therefore deliberately roll off the high frequency gain, either by filtering the signal applied to the derivative function or directly limiting the derivative action.

13.27.7 Incremental controllers

Diaphragm operated actuators can be arranged to fail open or shut by reversing the relative positions of the drive pressure and return spring. In some applications a valve will be required to hold its last position in the event of failure. One way to achieve this is with a motorised actuator, where a motor drives the valve via a screw thread.

Such an actuator inherently holds its last position but the position is now the integral of the controller output. An integrator introduces 90° phase lag and gain which falls off with increasing frequency. A motorised valve is therefore a destabilising influence when used with conventional controllers.

Incremental controllers are designed for use with motorised valves and similar integrating devices. They have the control algorithm

$$O_P = K\left(\frac{1}{T_i}e + \frac{de}{dt} + \frac{1}{T_d}\frac{d^2e}{dt^2}\right)$$

which is the time derivative of the normal control algorithm.

Incremental controllers are sometimes called *boundless controllers* or *velocity controllers* because the controller output specifies the actuator rate of change (i.e. velocity) rather than actual position.

Incremental controllers cannot suffer from integral windup per se, but it is often undesirable to keep driving a motorised valve once the end of travel is reached. End of travel limits are often incorporated in motorised valves to prevent jamming. The controller also has no real 'idea' of the valve true position, and hence cannot give valve position indication. If end of travel signals are available, a valve model can be incorporated into the controller to integrate the controller output to give a notional valve position. This model would be corrected whenever an end of travel limit is reached. Alternatively a position measuring device can be fitted to the valve for remote indication.

Pulse width modulated controllers are a variation on the incremental theme. Split phase motor drive valves require logic raise/lower signals, and normal proportional control can be simulated by using time proportional raise/lower outputs.

13.27.8 Scheduling controllers

Many loops have properties which change under the influence of some measurable outside variable. The gain of a flow control valve, (i.e. the change in flow for change in valve position) varies considerably over the stroke of a valve. The levitation effect of steam bubbles in a boiler drum causes the drum level control to have different characteristics under start-up, low load and high load conditions.

A scheduling controller has a built-in look up table of control parameters (gain, filtering, integral time etc.) and the appropriate values selected for the measured plant conditions.

13.27.9 Variable gain controllers

Process variable noise occurs in many loops; level and flow being possibly the worst offenders. This noise causes unnecessary actuator movement, leading to premature wear and inducing real changes in the plant state. Noise can, of course, be removed by first or second order filters, but these reduce the speed of the loop and the additional phase shift from the filters can often act to de-stabilise a loop.

A controller with gain K will pass a noise signal $K \cdot n(t)$ to the actuator where $n(t)$ is the noise signal. One obvious way to reduce the effect of the noise is to reduce the controller gain, but this degrades the loop performance. Usually the noise signal has a small amplitude compared with the signal range, if it has not the process will be practically uncontrollable. What is intuitively required is a low gain when the error is low, but a high gain when the error is high.

Figure 13.80(*a*) shows how such a scheme operates. The noise amplitude lies in the range AB, so this is made a low gain region. Outside this band the gain is much higher. The gain in the region AB should be low, but not zero, to keep the process variable at the set point. With a pure deadband (i.e. zero gain in region AB) the process variable would cycle between one side of the centre band and the other.

Figure 13.80(*b*) shows a possible implementation. A comparator switches between a low gain and high gain controller according to the magnitude of the error. Note that integral balancing is required between the two controllers to stop integral windup in the unselected controller.

Figure 13.80 has two gain regions. It is possible to construct a controller whose gain varies continuously with error. Such a controller has a response

$$O_P = Kf(e)\left(e + \frac{1}{T_i}\int e\,dt + T_d\frac{de}{dt}\right)$$

where $f(e)$ is a function of error.

A common function is

$$f(e) = abs\left(\frac{m + (1-m)e}{100}\right) \qquad (13.24)$$

where e is expressed as a percentage (0–100%) and m is a user set linearity adjustment ($0 < m < 1$). The *abs* operation (which always returns a positive sign) is necessary to prevent the controller action changing sign on negative error.

With $m = 1$, $f(e) = 1$ and Equation 13.24 behaves as a normal three term controller. With $m = 0$. the proportional part of Equation 13.24 follows a square law. Like *Figure 13.80*(*a*), this has low gain or small error (zero gain at zero error) but progressively increasing gain as the error increases.

Position control systems often need a fast response but cannot tolerate an overshoot. These often use Equation 13.24 with m at a low value approximating to the quadratic curve. This gives a high take off speed, but a low speed of approach.

13.27.10 Inverse plant model

The ideal control strategy, in theory, is one which mimics the plan behaviour. Given a totally accurate model of the plant, it should be possible to calculate what controller output is required to follow set point change, or compensate for a disturbance. The problem here is, of course, having an accurate plant model, but even a rough approximation should suffice as the controller output will converge to the correct value eventually.

One possible solution is shown in *Figure 13.81*. The process is represented by a block with transfer function $K \cdot f(s)$ where K is the d.c. (low frequency) gain. Following a change in set point, the signal A should mimic exactly the process

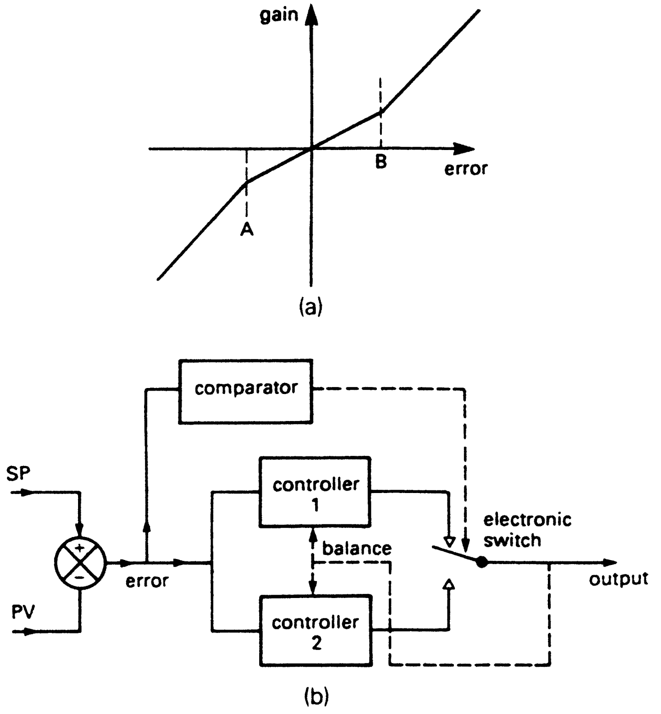


Figure 13.80 Variable gain controller: (a) system response; (b) block diagram

variable *B*, leading to a constant output from the controller exactly correct to bring the plant to the set point without overshoot. With a perfect model, the change at *A* should match the change at *B* as the set point is approached.

The inverse plant model is usually implemented with a sampled digital system. The problem with this simple, and apparently ideal, controller is that it will probably demand actuation signals which will drive the controller output, the actuator or parts of the plant, into saturation. It also requires an accurate plant model. A more gentle version of this technique aims to get a fraction, say 0.1 of the way from the current value to the desired value of the process variables at each sample time. This approximates to an exponential response.

13.28 Digital control algorithms

13.28.1 Introduction

So far we have assumed that controllers deal with purely analog signals. Increasingly controllers are digital, with the analog signal from the transducer being sampled by an ADC, the control algorithm being performed by software and the analog output being obtained from a DAC. ADCs and DACs are described in Chapter 14, Section 14.9. The system does not therefore continually control but takes 'snapshots' of the system state. Such an approach is called a *sampled system*.

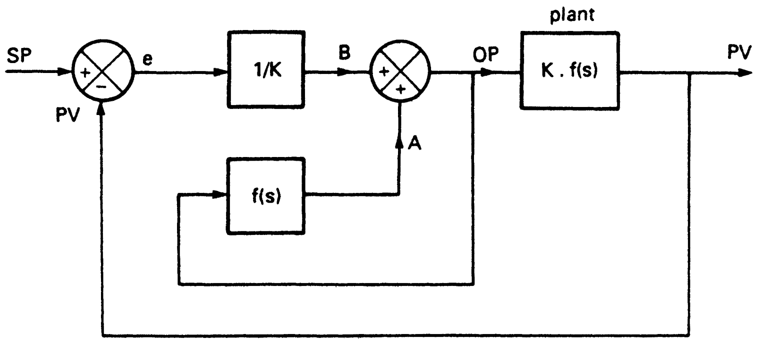


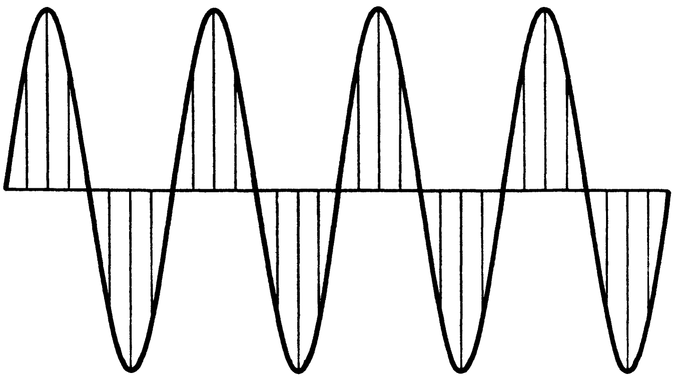
Figure 13.81 The inverse plant model

13.28.2 Shannon's sampling theorem

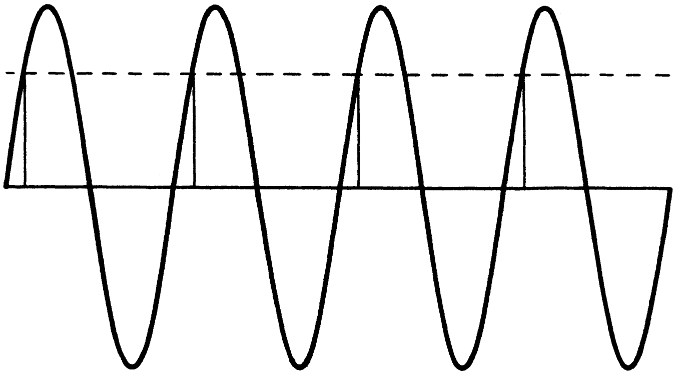
A sampled system only knows about the values of its samples. It cannot infer any other information about the signals it is dealing with. An obvious question, therefore, is what sample rate we should choose if our samples are to accurately represent the original analog signals.

In *Figure 13.82*(*a*) a sine wave is being sampled at a relatively fast rate. Intuitively one would assume this sampling rate is adequate. In *Figure 13.82*(*b*) the sample rate and the frequency are the same. This is obviously too slow as the samples imply a constant unchanging output.

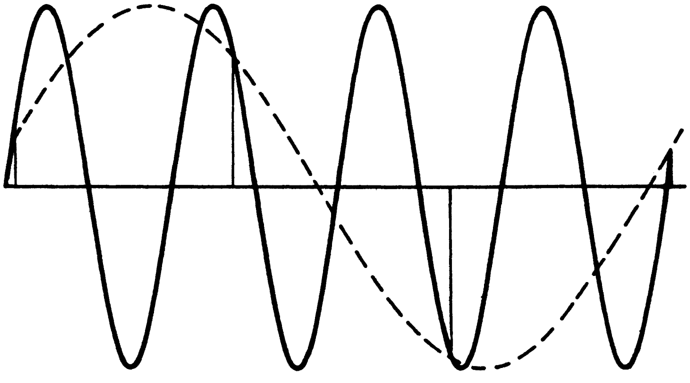
In *Figure 13.82*(*c*) the sample rate is lower than the frequency and the sample values are implying a sine wave of much lower frequency than the signal. This is called '*aliasing*'. A visual effect of aliasing can be seen on cinema screens where moving wheels often appear to go backwards. This effect occurs because the camera samples the world at about 50 times per second.



(a) Good sampling



(b) Sample rate too slow



(c) Aliasing

Figure 13.82 The effect of the sampling rate: (a) good sampling rate; (b) sampling frequency same as signal frequency, too slow; (c) sampling rate much too slow, aliasing is occurring

Any continuous signal will have a bandwidth of interest. The sampling frequency should be at least twice the bandwidth of interest. This is known as *Shannon's sampling theorem*. Any real life system will not, however, have a well defined bandwidth and sharp cut-off point. Noise and similar effects will cause any real signal to have a significant component at higher frequencies. Aliasing may occur with these high frequency components and cause apparent variations in the frequency band of interest. Before sampling, therefore, any signal should be passed through a low pass *anti-aliasing filter* to ensure only the bandwidth of interest is sampled.

Most industrial control signals have a bandwidth of a few Hz, so sampling within Shannon's limit is usually not a problem. Normally the critical bandwidth is not known precisely so a sample rate of about 5 to 10 times the envisaged bandwidth is used.

13.28.3 Control algorithms

To achieve three term control with sampled signals we must find the derivative and the integral of the error. As shown on *Figure 13.83* we are dealing with a set of sampled signals, y_n, y_{n-1}, y_{n-2} etc. where y_n is the most recent. If the sample time is Δt, the slope is then given by:

$$\text{slope} = \frac{Y_n - Y_{n-1}}{\Delta t}$$

Integration is equivalent to finding the area under a curve as shown for an analog and digital signal on *Figure 13.83*(*b*). The trapezoid integration of *Figure 13.83*(*c*) is commonly used where the area is given by:

$$\text{area} = \frac{\Delta t(Y_n + Y_{n-1})}{2}$$

Combining these gives a digital sampled PID algorithm

$$O_P = K\left(e_n + \frac{1}{T_i}\sum\frac{\Delta t(e_n + e_{n-1})}{2} + \frac{T_d}{\Delta t}(e_n - e_{n-1})\right)$$

where e_n is the error for sample n and Δt the sample time as before.

13.29 Auto-tuners

Tuning a controller is more of an art than an exact science and can be unbelievably time consuming. Time constants of tens of minutes are common in temperature loops, and lags of hours occur in some mixing and blending processes. Performing, say, the ultimate cycle test of Section 13.30.2 on such loops can take several days.

Self tuning controllers aim to take the tedium out of setting up a control loop. They are particularly advantageous if the process is slow (i.e. long time constants) or the loop characteristics are subject to change (e.g. a flow control loop where pressure/temperature changes in the fluid alter the behaviour of the flow control valve.)

Self tuning controllers give results which are generally as good, if not slightly better, than the manual methods of Section 13.30 (possibly because self tuning controllers have more patience than humans!). In the author's experience, however, the results from a self tuner should be viewed as recommendations or initial settings in the same way as the results from the manual methods described in the following sections. One early decision to be made when self tuners are used is whether they should be allowed to alter control parameters without human intervention. Many engineers (of whom the author is one) view self tuners as commissioning aids to be removed before a plant goes into production.

There are essentially two groups of self tuners. *Modelling self tuners* try to build a mathematical model of the plant (usually second order plus transit delay) then determine controller parameters to suit the model. These are sometimes called *explicit self tuners*.

Model identification is usually based on the principles of *Figure 13.84*. The controller applies a control action O_P to the plant and to an internal model. The plant returns a

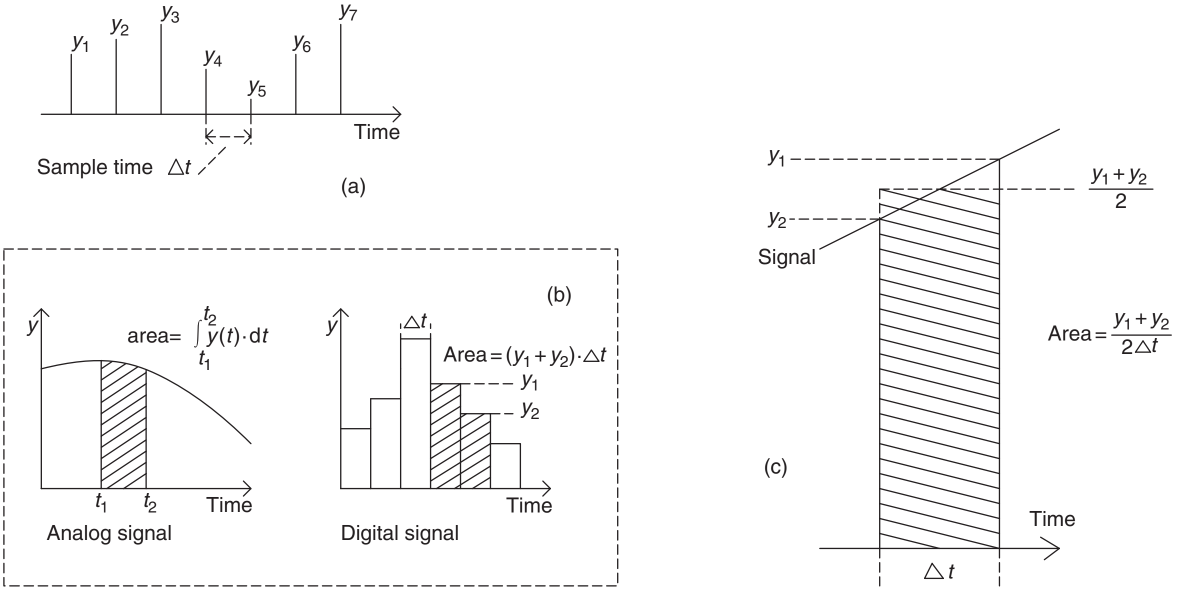


Figure 13.83 Control algorithm with sampled signals: (a) the sampled signals; (b) integration for an analog and digital signal; (c) trapezoid integration

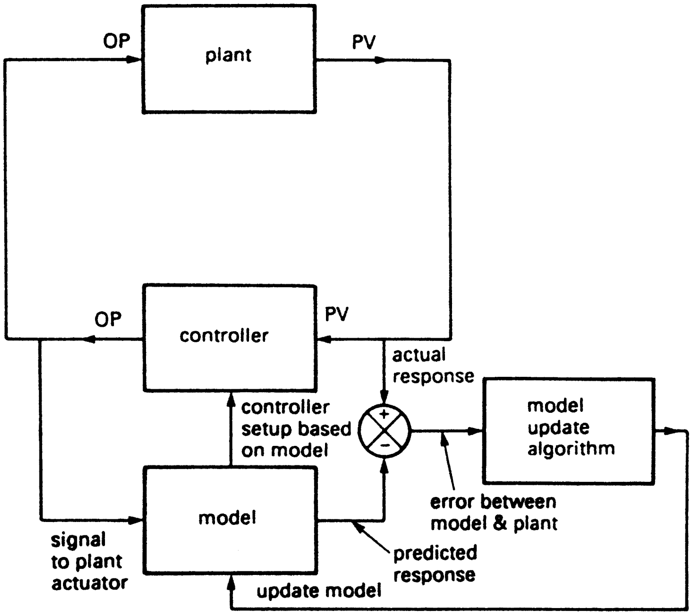


Figure 13.84 A modelling self tuning controller

process variable P_V and the model a prediction P_{Vm}. These are compared, and the model updated (often via the statistical least squares technique). On the basis of the new model, new control parameters are calculated, and the sequence repeated.

A model building self tuner requires actuation changes to update its model, so it follows that self tuners do not perform well in totally static conditions. In a totally stable unchanging loop, the model, and hence the control parameters, can easily drift off to ridiculous values. To prevent this, most self tuners are designed to 'kick' the plant from time to time, with the size and repetition rate of the kick being set by the control engineer. Less obviously, model building self tuners can be confused by outside disturbances which can cause changes in P_V that are not the result of the controller output.

The second group of self tuners (sometimes called *implicit tuners*) use automated versions of the manual tests described in Section 13.30, and as such do not attempt to model the plant. A typical technique will vary the controller gain until a damped oscillatory response is observed. The control parameters can then be inferred from the controller gain, the oscillation period and the oscillation decay rate.

The useful bang/bang test of Section 13.30.3 can be performed automatically by a controller which forces limit cycling in the steady state via a comparator. A limit block after the comparator restricts the effect on the plant.

Implicit self tuners, like their modelling brothers, do not perform well on a stable unchanging loop, and can be equally confused by outside disturbances.

13.30 Practical tuning methods

13.30.1 Introduction

Values must be set for the gain and integral/derivative times before a controller can be used. In theory, if a plant model is available, these values can be determined from Nichols charts or Nyquist diagrams. Usually, however, the plant characteristics are not known (except in the most general terms) and the controller has to be tuned by experimental methods.

It should be noted that all these methods require pushing the plant to the limit of stability. The safety implications of these tests must be clearly understood. Tuning can also be very time consuming. With large chemical plants tuning of one loop can take days.

Most of the tests aim to give a quarter cycle decay and assume the plant consists of a transit delay in series with a second order block (or two first order lags) plus possible integral action.

In conducting the tests, it is useful to have a two pen recorder connected to the P_V (process variable) and O_P (controller output) as shown on *Figure 13.85*. The range of the pens (e.g. 0–10 V or 1–5 V) should be the same.

In the tests below, the gain is expressed as proportional band (P_B) per cent. Time is used for integral and derivative

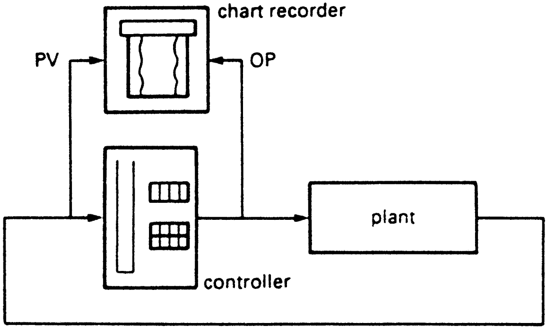


Figure 13.85 Suggested equipment setup for controller tuning

action. Conversion to gain or repeats per minute is straightforward.

13.30.2 Ultimate cycle methods

The basis of these methods is determining the controller gain which just supports continuous oscillation, i.e. point *A* and gain *K* on the Nichols chart and Nyquist diagram of *Figure 13.86*. The method is based on work by J. G. Ziegler and N. B. Nichols and is often called the Ziegler Nichols method.

The integral and derivative actions are disabled to give proportional only control, and the control output manually adjusted to bring P_V near the required value. Auto control is selected with a low gain.

Step disturbances are now introduced and the effect observed. One way of doing this is to go back into manual, shift O_P by, say 5%, then reselect automatic control. At each trial the gain is increased. The increasing gain will give a progressively underdamped response and eventually continuous oscillation will result. Care must be taken in these tests to allow all transients to die away before each new value of gain is tried.

If the value of gain is too high, the oscillations will increase. The value of gain which gives constant oscillations neither increasing or decreasing is called the ultimate gain, or P_u (expressed as proportional band). The period of the oscillations T_u should also be noted from the chart recorder (or with a watch).

The required controller settings are:

Proportional only control

P_B	$2P_u$ %

PI Control

P_B	$2.2P_u$ %
T_i	$0.8.T_u$

PID Control

P_B	$1.67 . P_u$ %
T_i	$T_u/2$
T_d	$T_u/8$

$T_i = 4T_d$ is a useful rule of thumb.

Other recommended settings for a PID controller are:

P_B	$2P_u$ %
T_i	T_u
T_d	$T_u/5$

and

P_B	$2P_u$ %
T_i	$0.34T_u$
T_d	$0.08T_u$

All of these values should be considered as starting points for further tests.

13.30.3 Bang/bang oscillation test

This is the fastest, but most vicious, test. It can, though, be misleading if the plant is non linear. Integral and derivative actions are disabled and the controller gain set as high as possible (ideally infinite) to turn the controller into a bang-bang controller. The controller output is set manually to bring the process value near the set point then the controller switch into automatic mode.

Violent oscillations will occur as shown on *Figure 13.87*. The period of the oscillations T_o is noted along with the peak to peak height of the process variable oscillations as a percentage H_o% of full scale.

The required controller settings are:

Proportional control

P_B	$2.H_o$%

PI Control

P_B	$3.H_o$%
T_i	$2.T_o$

PID Control

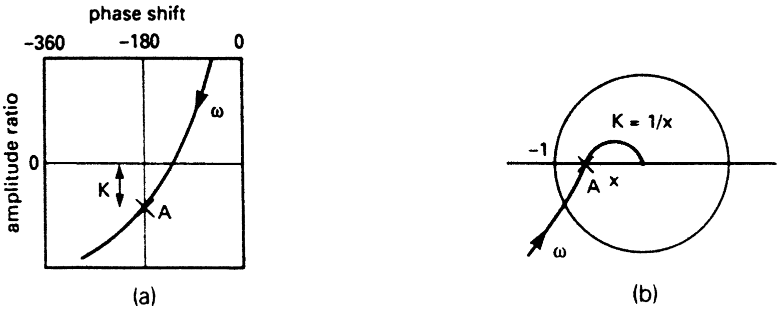
P_B	$2.H_o$%
T_i	T_o
T_d	$T_o/4$

13.30.4 Reaction curve test

This is an open loop test originally proposed by American engineers Cohen and Coon. It assumes the plant consists of a measurable transit delay and a dominant time constant. It cannot be applied to plants with integral action (e.g. level control systems).

A chart recorder must be connected to the plant as shown earlier on *Figure 13.85* to perform the test. The controller output is first adjusted manually to bring the plant near to the desired operating point. After the transients have died away a small manual step ΔO_P is applied which results in a small change ΔP_V as shown on *Figure 13.88*.

The process gain K_p is then simply $\Delta P_V/\Delta O_P$.



(a)

(b)

Figure 13.86 Basis of the ultimate cycle test. Point A determines the frequency at which continuous oscillations will occur when gain K is applied: (a) Nichols chart; (b) Nyquist diagram

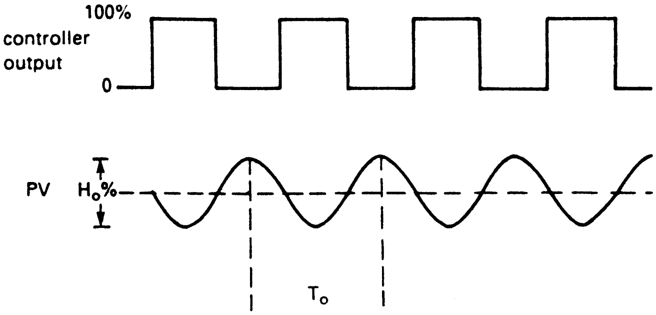


Figure 13.87 The bang-bang oscillation test

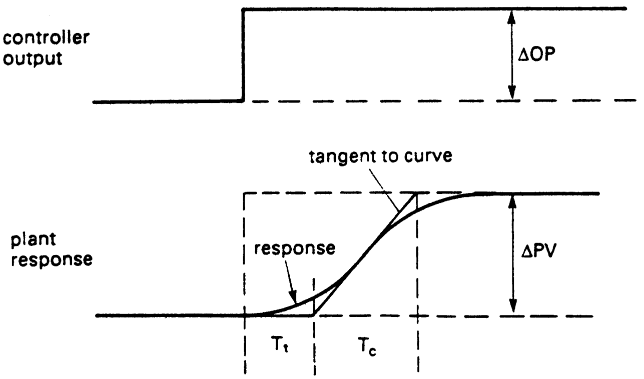


Figure 13.88 The reaction curve test

A tangent is drawn to the process variable curve at the steepest point from which an apparent transit delay T_t and time constant T_c can be read. The settings for the controller are then given by:

Proportional
P_B $100.K_p.T_t/T_c$ %

PI
P_B $110.K_p.T_t/T_c$ %
T_i $3.3\ T_t$

PID
P_B $80.K_p.T_t/T_c$ %
T_i $2.5\ T_t$
T_d $0.4\ T_t$

Because the test uses an open loop trial with a small change in the controller output it is the gentlest and least hazardous tuning method.

13.30.5 A model building tuning method

The closed loop tuning methods described so far require the plant to be pushed to, (and probably beyond), the edge of instability in order to set the controller. An interesting gentle tuning method was described by Yuwana and Seborg in the journal *AIChE* Vol 28 no 3 in 1982.

The method assumes the plant has gain K_M and behaves as a dominant first order lag T_M in series with a transit delay D_M. This assumption can give gross anomalies with plants with integral action such as level or position controls, but is commonly used for many manual and automatic/adaptive controller tuning methods. With the above warning noted, the suggested controller settings can be found from:

$$K = A(D_M/T_M)^{-B}/K_M$$

$$T_I = CT_M(D_M/T_M)^D$$

$$T_D = T_M E(D_M/T_M)^F$$

where A, B, C, D, E, F are constants defined:

Mode	*A*	*B*	*C*	*D*	*E*	*F*
P	0.490	1.084				
PI	0.859	0.997	1.484	0.680		
PID	1.357	0.947	1.176	0.738	0.381	0.99

These apparently random equations and constants come from experimental work described by Miller *et al* in *Control Engineering* Vol 14 no 12.

The method of finding the plant gain, time constant and transit delay is based on a single quick test with the plant operating under closed loop control. The test is performed on the plant operating under proportional only control, with a gain sufficient to produce a damped oscillation as *Figure 13.89* when a step change in set point from R_0 to R_1 is applied.

The subsequent process maximum C_{P1}, minimum C_{M1} and next maximum C_{P2} are noted along with the time D_{T2} between C_{P1} and C_{P2}. The controller proportional band used for the test, P_B, is also recorded, from which the controller gain $K_{PB} = 100/P_B$ is found.

Given the values from the test the method estimates the value of the plant steady state gain K_M, lag time constant T_M and the transit delay time D_M. The background mathematics is given at length in the original paper.

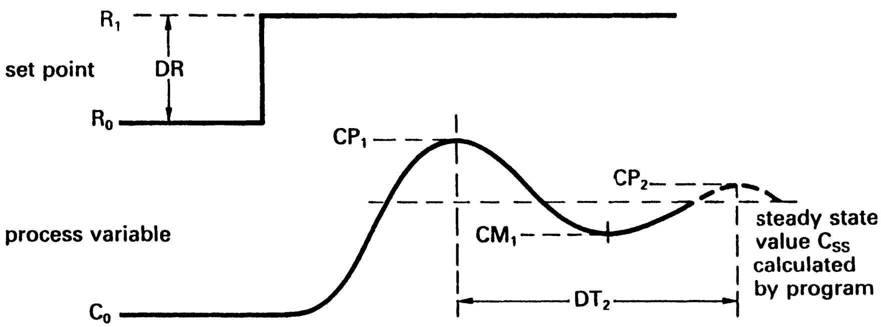


Figure 13.89 Test performed for model building tuning test. Note that R_0 and C_0 need not be the same and DR can be positive or negative

The equations above are not very practical for manual use on site, so the original paper was developed into a program for the Hewlett Packard HP-67 calculator by Jutan and Rodriguez and published in the magazine *Chemical Engineering September* 1984. The nomenclature used in the above equations is based on this article.

13.30.6 General comments

The above test procedures do not give guaranteed results and should be viewed as a method of putting the engineer in the right area. They should be viewed as the starting point for further trials. The important thing in these trials is only change one thing at once.

With values set as above the effect of changing the gain should be tried first. It is always useful to have the proportional gain as high as possible to give the largest initial control action to changes and disturbances. However, a large gain can give undesirable changes in the controller output if the process variable is noisy. The gain should be adjusted to give the desired overshoot and damping.

Integral action should be adjusted next to give best removal of offset error. During these trials it is best to disable any derivative action. Decreasing the integral time reduces the time taken to remove the offset error. It may be necessary to reduce the gain again as integral time is decreased. A useful rule of thumb is that the ratio of T_i/Gain is an 'index' of stability for a given system, i.e. a T_i of 12 sec and a gain of 2 will give a similar damping to a T_i of 24 sec and a gain of 4.

The derivative action should be adjusted last. Many systems do not benefit from derivative action, particularly those with a noisy process variable signal which causes large controller output swings. Where derivative action is required, $T_d = T_i/4$ is a good starting point. Many controllers allow the user to select derivative action on error or derivative action on process variable. The former is best for tracking systems, but gives large controller output swings for step changes in the set point. Derivative action on process variable is usually the best choice.

One final observation, based on experience rather than theory, is that a P_B of 200% (gain of 0.5), T_i of 20 sec and no derivative action is a good starting point for a majority of plants. Adjust the gain to give the required overshoot then adjust T_i to be as small as possible. Finally set T_d, if needed, to $T_i/4$.

References

1 SAUCEDO, R. and SCHIRING, E. E., *Introduction to Continuous and Digital Control Systems*, Macmillan, New York (1968)
2 FRANKLIN, G. F., POWELL, J. D. and EMAMI-NAEINI, A., *Feedback Control of Dynamic Systems*, Addison-Wesley, New York (1986)

Bibliography

The authors have found the following books useful for basic control engineering studies. This list is by no means exhaustive.

ANAND, D. K., *Introduction to Control Systems*, Pergamon Press, Oxford (1974)
CHEN, C. F. and HAAS, I. J., *Elements of Control Systems Analysis*, Prentice-Hall, Englewood Cliff, NJ (1968)
DISTEFANO,J.J.,STUBBERUD,A.R.andWILLIAMS,I.J., *Theory and Problems of Feedback and Control Systems*, Schaum's Outline Series, McGraw-Hill, New York (1990)
DORF, R. C., *Modern Control Systems*, Addison-Wesley, New York (1980)
DOUCE, J. L., *The Mathematics of Servomechanisms*, English Universities Press, London (1963)
ELGARD, O. I., *Control Systems Theory*, McGraw-Hill, New York (1967)
GOLTEN, J. and VERWER, A., *Control System Design and Simulation*, McGraw Hill (1991)
HEALEY, M., *Principles of Automatic Control*, Hodder and Stoughton, New York (1975)
JACOBS, O. L. R., *Introduction to Control Theory*, Oxford University Press, Oxford (1974)
LANGILL, A. W., *Automatic Control Systems Engineering*, Vols I and II, Prentice-Hall, Englewood Cliffs, NJ (1965)
MARSHALL, S. A., *Introduction to Control Theory*, Macmillan, New York (1978)
POWER, H. M. and SIMPSON, R. J., *Introduction to Dynamics and Control*, McGraw-Hill, New York (1978)
RAVEN, F. H., *Automatic Control Engineering*, McGraw-Hill, New York (1961)
SHINSKEY, F. G., *Process Control Systems*, McGraw Hill (1988)

14

Digital Control Systems

E A Parr MSc, CEng, MIEE, MInstMC
CoSteel Sheerness

Contents

14.1 Introduction 14/3
  14.1.1 Analog and digital circuits 14/3
  14.1.2 Types of digital circuits 14/3
  14.1.3 Logic gates 14/3

14.2 Logic families 14/5
  14.2.1 Introduction 14/5
  14.2.2 Speed 14/5
  14.2.3 Fan in/fan out 14/6
  14.2.4 Noise immunity 14/6
  14.2.5 Transistor transistor logic (TTL) 14/6
  14.2.6 Complementary metal oxide semiconductor (CMOS) logic 14/7
  14.2.7 Emitter coupled logic (ECL) 14/8
  14.2.8 Open collector and tri-state outputs 14/9
  14.2.9 Schmitt triggers 14/9
  14.2.10 Choosing a logic family 14/10

14.3 Combinational logic 14/11
  14.3.1 Introduction 14/11
  14.3.2 Truth tables 14/12
  14.3.3 Boolean algebra 14/13
  14.3.4 Karnaugh maps 14/14
  14.3.5 Conversion between P of S and S of P representations 14/16
  14.3.6 Formal minimisation, the Quine-McCluskey method 14/16
  14.3.7 Hazards, races and glitches 14/17
  14.3.8 Integrated circuits 14/18
  14.3.9 UCLAs, PALs and PLAs 14/18

14.4 Storage 14/19
  14.4.1 Introduction 14/19
  14.4.2 Cross coupled flip flops 14/19
  14.4.3 D type flip flop 14/21
  14.4.4 The JK flip flop 14/22
  14.4.5 Clocked storage 14/22

14.5 Timers and monostables 14/23

14.6 Arithmetic circuits 14/24
  14.6.1 Number systems, bases and binary 14/24
  14.6.2 Binary arithmetic 14/25
  14.6.3 Binary coded decimal (BCD) 14/27
  14.6.4 Unit distance codes 14/27

14.7 Counters and shift registers 14/27
  14.7.1 Ripple counters 14/27
  14.7.2 Synchronous counters 14/29
  14.7.3 Non binary counters 14/29
  14.7.4 Shift registers 14/30

14.8 Sequencing and event driven logic 14/30

14.9 Analog interfacing 14/33
  14.9.1 Digital to analog conversion (DAC) 14/33
  14.9.2 Analog to digital converters (ADCs) 14/33

14.10 Practical considerations 14/34

14.11 Data sheet notations 14/36

14.1 Introduction

14.1.1 Analog and digital circuits

Signals in process control are conventionally transmitted as a pneumatic pressure or electrically as a voltage or current. These signals are said to be continuously variable in that they can take any value between the two extreme limits. Such systems are called analog systems.

Digital circuits are concerned with signals that can only take certain values. Most digital circuits deal with electrical signals that can only have two values; 5 V or 0 V for example. Many circuits are inherently of this type, a light can be on or off, a valve open or shut, a motor running or stopped.

14.1.2 Types of digital circuits

Digital applications can, in general, be classified into three types. The simplest of these are called *combinational logic* (or *static logic*), and can be represented by *Figure 14.1*. Such systems have several digital inputs and one or more digital outputs. The output states are uniquely defined for every combination of input states, and the same input combination always gives the same output states.

A *sequencing logic* system is superficially similar to *Figure 14.1* but the output states depend not only on the inputs but also on what the system was doing last (i.e. its previous state). Sequencing systems therefore have memory and storage elements. A very simple example is the motor starter of *Figure 14.2*(*a*). The start input causes the motor to start running and keep running even when the start signal is removed. The stop input stops the motor. The action is summarised on *Figure 14.2*(*b*). Note that with neither signal present the motor could be running or stopped dependent on which signal occurred last; i.e. the output state is not defined solely by the present input states.

The final group of digital systems uses digital signals to represent, and manipulate, numbers. Such systems cover the range from simple counters and digital displays to complex arithmetic and computing circuits.

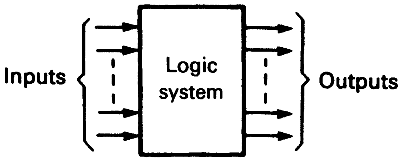


Figure 14.1 Representation of a combinational logic system. The output states are defined only by the input state

14.1.3 Logic gates

The simplest digital device is the electromagnetic relay, and it is useful to describe some of the fundamental ideas in terms of relay contacts. In *Figure 14.3*(*a*), the coil Z will energise when contact A *and* contact B *and* contact C are made. The series connection of contacts performs on AND function.

Similarly, in *Figure 14.3*(*b*) the coil Z will energise when contact A *or* contact B *or* contact C are made. The parallel connection of contacts performs an OR function.

In *Figure 14.3*(*c*), coil Z is energised when the push button is pressed. A normally closed contact of Z controls coil Y. When Z is energised, Y is de-energised and vice versa. The normally closed contact can be said to invert the state of its coil.

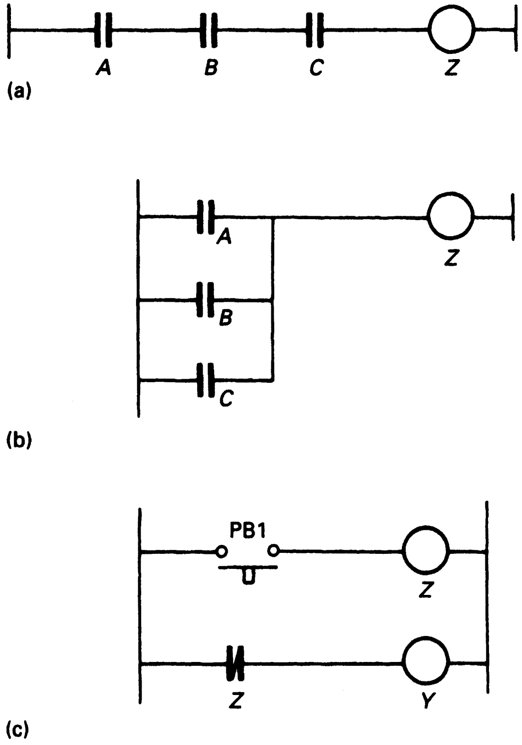


Figure 14.3 Simple relay logic: (a) AND combination, relay Z is energised when A & B & C are all energised; (b) OR combination, relay Z is energised if A or B or C is energised; (c) Inversion, relay Y is energised when PB1 is not made

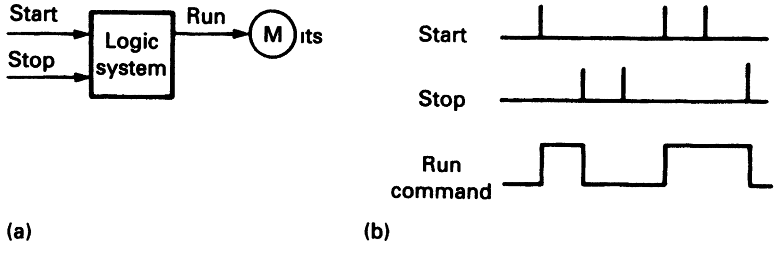


Figure 14.2 A simple sequencing system: (a) representation of a start/stop motor starter; (b) operation, the output depends not only on the input state, but also on the last operation

Combinational logic circuits are built round combinations of AND, OR and INVERT circuit. In *Figure 14.4*(*a*), for example, Z will be energised for:

(A not energised) AND (B energised OR C energised)

Such verbal descriptions are impossibly verbose for more simplex combinations. Circuit operations are more conveniently expressed as an equation. Normally closed contacts are represented by a bar over the top of the contact name (e.g. $\bar{A}$, verbalised as A bar). The circuit of *Figure 14.4*(*a*) can then be represented as:

$Z=\bar{A}$ AND (B OR C)⇐

Similarly the circuit of *Figure 14.4*(*b*) (commonly used for stairwell lighting) can be represented by:

$Z=(A$ AND $\bar{B})$ OR $(\bar{A}$ AND $B)$⇐

These are known as Boolean equations, a topic discussed further in Section 14.3.3.

Relays can perform all logic functions but are slow (typically 10 to 20 operations per second), bulky and power hungry. Electronic circuits performing similar functions are called logic gates. These work with signals that can only have two states. A signal in CMOS logic, for example, can be at 12 V or 0 V and could represent a limit switch made or open. The two logic states can be called high/low, on/off, true/false and so on. The usual convention, however, is to call the higher voltage ‘1’ and the lower voltage ‘0’. For a CMOS gate, therefore, 12 V is 1 and 0 V is 0.

Figure 14.5(*a*) shows the circuit of a simple AND gate. Neglecting diode drops, the output Z will be equal to the lower of the two input voltages. In other words, it will be a 1 if, and only if, both inputs are 1. This can be represented by *Figure 14.5*(*b*) (which is called a *truth table*). On circuit diagrams it is clearer to use logic symbols rather than the actual circuit diagram. The symbol for an AND gate is shown on *Figure 14.5*(*c*); the output Z being 1 when A AND B are both 1.

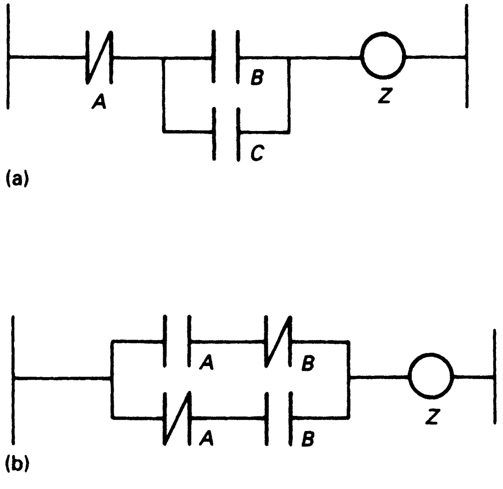


Figure 14.4 More complex relay logic: (a) Z is energised when (A is not energised) AND (B is energised OR C is energised); (b) Stairwell lighting circuit. A and B are the switches at the top and bottom of the stairs. Changing either switch will change the state of relay Z

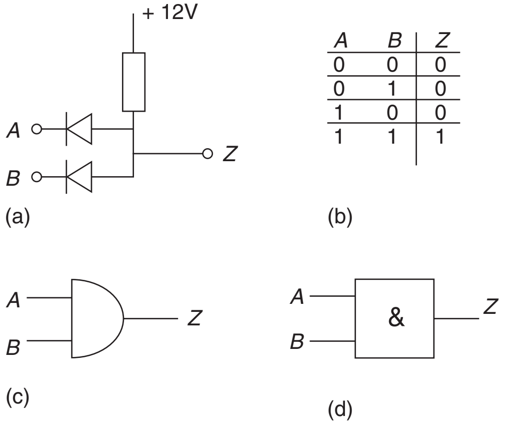


Figure 14.5 A simple diode based AND Gate: (a) circuit; (b) truth table; (c) logic symbol; (d) BS logic symbol

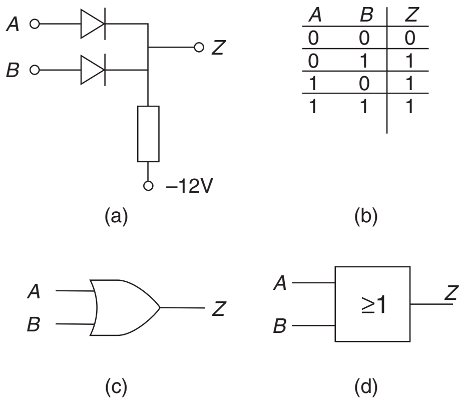


Figure 14.6 A simple diode based OR gate: (a) circuit; (b) truth table; (c) logic symbol; (d) BS logic symbol

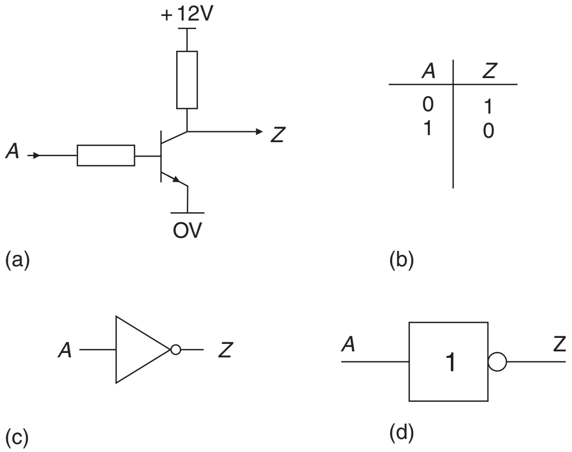
On *Figure 14.6*(*a*) the output Z will be equal to the higher of the two inputs (again neglecting diode drops). Z will therefore be 1 if either input is 1 giving the truth table of *Figure 14.6*(*b*). The logic symbol for an OR gate is shown on *Figure 14.6*(*c*).

The invert function is given by the simple saturating transistor of *Figure 14.7*(*a*). When A is 0, the transistor is turned off and the output Z is pulled to a 1 state by the collector load resistor. When A is 1, the transistor is saturated on taking Z to 0 V; a 0. The circuit behaves as the truth table of *Figure 14.7*(*b*) and has the logic symbol of *Figure 14.7*(*c*).

Combinational logic circuits can be drawn purely in terms of AND gates, OR gates and inverters. The stairwell lighting circuit of *Figure 14.4*(*b*) is drawn with logic symbols on *Figure 14.8*(*a*). This behaves as the truth table of *Figure 14.8*(*b*) which shows that Z is 1 if only one input is 1. This circuit is known as an Exclusive OR and is sufficiently common to merit its own logic symbol shown on *Figure 14.8*(*c*).

If an inverter is used after an AND gate as *Figure 14.9*(*a*), the truth table of *Figure 14.9*(*b*) is produced. This arrangement is called a NAND gate (for NOT-AND) and has the logic symbol of *Figure 14.9*(*c*). The NAND gate is probably the commonest logic gate.

Adding an inverter to an OR gate as *Figure 14.10*(*a*) gives the truth table of *Figure 14.10*(*b*). This is known as a NOR gate (for NOT-OR) and is given the logic symbol of *Figure 14.10*(*c*).



A	Z
0	1
1	0

Figure 14.7 A transistor inverter: (a) circuit; (b) truth table; (c) logic symbol; (d) BS logic symbol

Note that the logic symbols for NAND/NOR gates are similar to those of the AND/OR gates with the addition of a small circle on the output. The circle denotes an inversion operation.

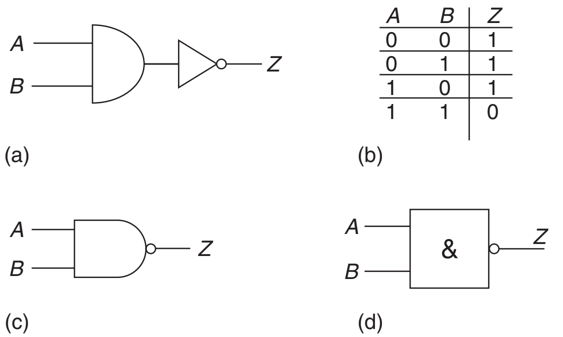
14.2 Logic families

14.2.1 Introduction

Most logic circuits are constructed from integrated circuits, and have high operating speed and well defined levels. Two logic families (TTL and CMOS) are widely used in industrial applications and a third family (ECL) may be encountered where very high speed is required. Before these are described, we must first examine how the various factors of a logic gates performance are specified.

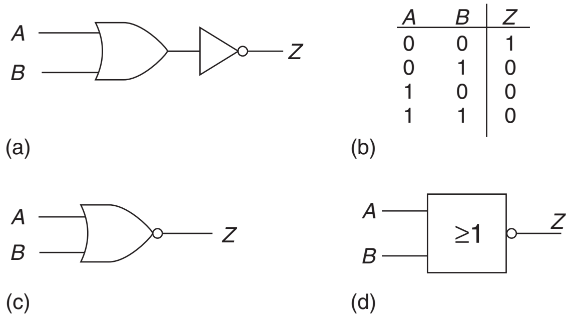
14.2.2 Speed

A logic gate does not respond instantly to a change at its input. For infinitely fast input signals the output will be delayed and the edges slowed as shown on *Figure 14.11.*



A	B	Z
0	0	1
0	1	1
1	0	1
1	1	0

Figure 14.9 A NAND gate: (a) circuit; (b) truth table; (c) logic symbol; (d) BS logic symbol



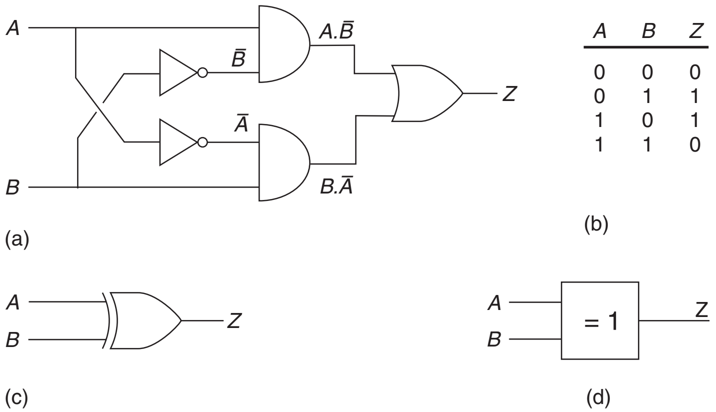
A	B	Z
0	0	1
0	1	0
1	0	0
1	1	0

Figure 14.10 A NOR gate: (a) circuit; (b) truth table; (c) logic symbol; (d) BS logic symbol

The delay is called the *propagation delay* and is defined from the mid point of the input signal to the mid point of the output signal. Typical values are around 5 nS for TTL.

The edge speeds are defined by the *rise time* (for the 0 to 1 edge) and the *fall time* (for the 1 to 0 edge). These are measured between the 10% and 90% points of the output signal. Typical values are 2 nS for TTL.

Propagation delays and rise/fall times determine the maximum speed at which a logic family can operate. TTL can operate in excess of 10 MHz, basic CMOS around 5 MHz and ECL at over 500 MHz (although considerable care needs to be taken with board layout at speeds over 10 MHz).



A	B	Z
0	0	0
0	1	1
1	0	1
1	1	0

Figure 14.8 An Exclusive OR (XOR) gate: (a) circuit; (b) truth table; (c) logic symbol; (d) BS logic symbol

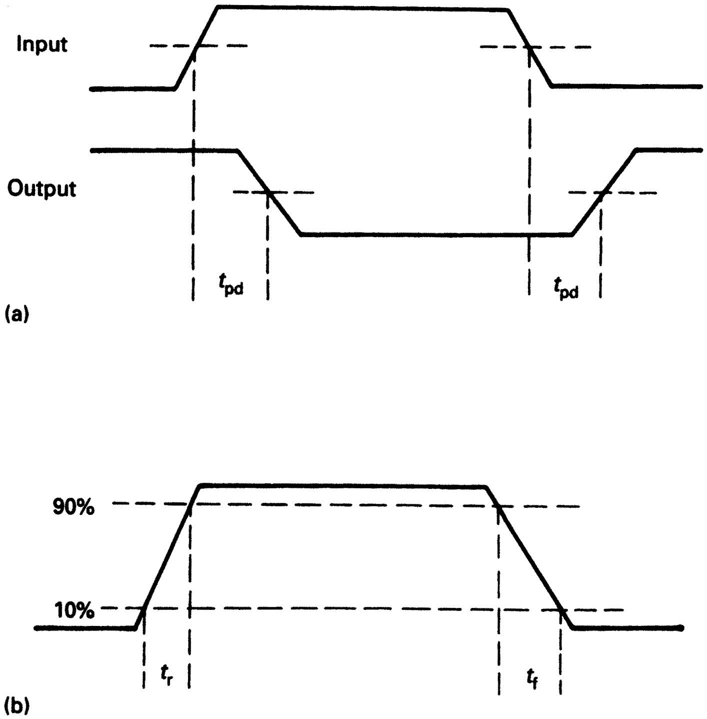


Figure 14.11 Speed definitions: (a) propagation delay (t_{pd}); (b) rise and fall times (t_r and t_f)

Power consumption is related to speed, as increased speed is obtained by reducing RC time constants formed by stray capacitance, and by using non saturating transistors. CMOS, for example, has a power consumption of about 0.01 mW per gate compared with ECL's figure of 60 mW/gate.

14.2.3 Fan in/fan out

The output of a logic gate can only drive a certain load and remain within specification for speed and voltage levels. There is therefore a maximum number of gate inputs a given gate output can drive. A simple gate input is called a *standard load*, and is said to have a *fan in* of one. A gate output's drive capability is called its *fan out*, and is defined in unit loads. A TTL gate output, for example, can drive ten standard gate inputs and correspondingly has a fan out of ten.

Some inputs appear as a greater load than a standard gate. These are defined as a fan in of an equivalent number of standard gate inputs. An input with a fan in of three, for example, looks like three gate inputs. Obviously the sum of all the fan in loads connected to a gate output must not exceed the gate's fan out.

14.2.4 Noise immunity

Electrical interference may cause 1 signals to appear as 0 signals and vice versa. The ability of a gate to reject noise is called its *noise immunity*. Defining noise immunity is more complex than it might at first appear, but the method usually adopted is that shown on *Figure 14.12(a)*. The voltages given are those for a TTL gate which has a nominal 1 V of 4.5 V and a nominal 0 Vof 0 V.

Next we define how far an output 1 can fall to (2.4 V) and a 0 rise to (0.4 V). These are respectively termed V_{OH} and V_{OL}. Finally we define how low a gate's input 1 can fall and an input 0 rise without allowing its output to go between V_{OH} and V_{OL}. These voltages are called V_{IH} (2.0 V) and V_{IL}(0.8 V). The noise immunity is then the smaller of $(V_{OH} - V_{IH})$ or $(V_{IL} - V_{OL})$. For TTL the figure is 0.4 V. This is a worse case value, a more typical noise immunity is about 1.2 V.

A figure sometimes quoted is the *AC noise margin*. This is defined as the largest pulse that will not propagate down a chain of gates similar to *Figure 14.12(b)*. This gives a more favourable result than *Figure 14.12(a)*, but is a more realistic test.

14.2.5 Transistor transistor logic (TTL)

TTL is NAND based logic, with the circuit of a 2 input NAND gate being shown on *Figure 14.13*. The rather odd looking dual emitter transistor can be considered as two transistors in parallel or three diodes as shown.

If both inputs are high, Q_2 is turned on by current from R_1 supplying base current to Q_3. The output is therefore nominally 0 V. With either input low, Q_1 is turned on, Q_2 turned off and Q_4 pulls the output high to a nominal 4.5 V.

The output transistors Q_3, Q_4 are called a *totem pole output* and play a significant part in increasing the operating speed. When the output is a 0, Q_3 acts as a saturated transistor. When the output is a 1, Q_4 acts as an emitter follower. Both states have low output impedances which reduce RC time constants from stray capacitance.

There are at least six versions of TTL with differing speeds and power consumption. Schottky versions (with S as part of the suffix) use Schottky diodes within the gate to reduce hole

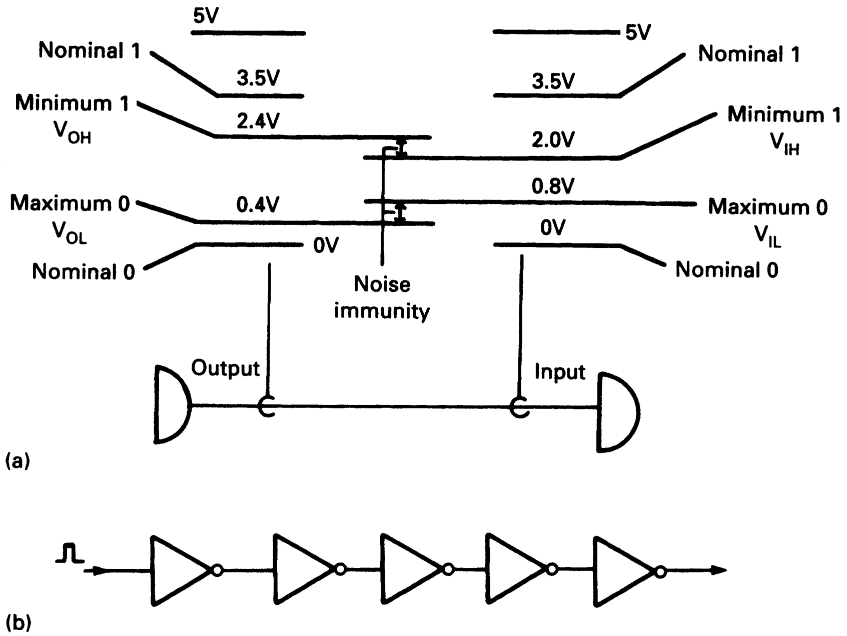


Figure 14.12 Definitions of noise immunity: (a) D.c. noise margin. The voltages shown are for standard TTL; (b) A.c. noise immunity. The test sees what is the smallest pulse amplitude that will propagate through the chain

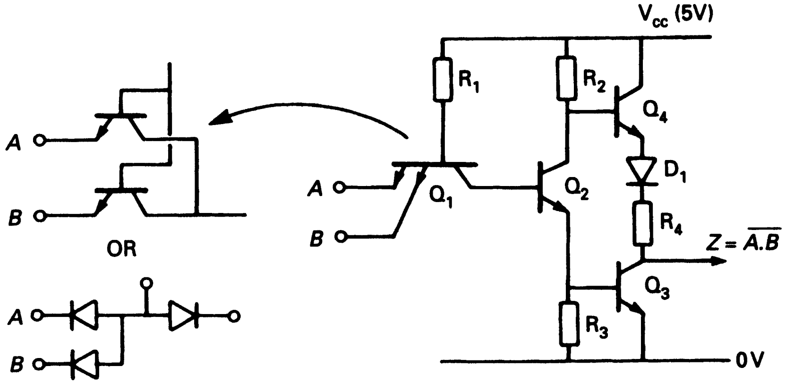


Figure 14.13 Transistor transistor logic (TTL) circuit diagram the multi-emitter transistor can be considered to act as two transistors in parallel or three diodes

storage delays. All TTL are members of the so called 74 series (originally conceived by Texas Instruments) and have the same pin arrangements on the ICs. They can also be intermixed although care must be taken because of the different input loadings and output capability (an LS gate input, for example, looks like half a normal gate input). All run on a 5 V supply and use nominal logic levels of 4.5 V and 0 V.

14.2.6 Complementary metal oxide semiconductor (CMOS) logic

CMOS is virtually the ideal logic family. It can operate on a wide range of power supplies (from 3 to 15 V), uses little power (approximately 0.01 mW at low speeds), has high noise immunity (about 4 V on a 12 V supply) and very large fan out (typically in excess of 50). It is not as fast as TTL or ECL but its maximum operating speed of 5 MHz is adequate for most industrial purposes. (Too high a maximum speed can actually be a disadvantage as it makes a system more noise prone.)

CMOS is built around the two types of field effect transistors shown on *Figure 14.14*. From a logic point of view these can be considered as a voltage operated switch. These switches can be used to manufacture logic gates.

Figure 14.15(a) shows how an inverter can be implemented. With *A* low, Q_1 is turned on and Q_2 off. With A high Q_2 is turned on and *Z* is low.

Similarly a NAND gate can be constructed as *Figure 14.15(b)*. If *A* or *B* is low, *Z* will be high because one of the parallel pair Q_1, Q_2 will be on, and one of the series pair Q_3, Q_4 will be off. The output *Z* will be low only when both *A* and *B* are high when Q_1, Q_2 are both off and Q_3, Q_4 are both on.

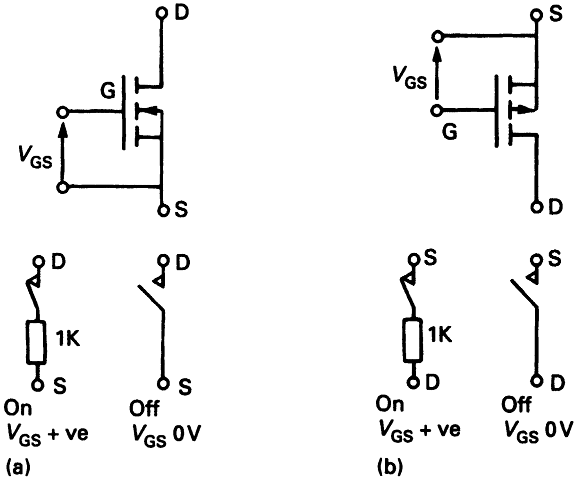

D
G
V_{GS}
S
D
1K
S
On
V_{GS} + ve
D
S
Off
V_{GS} 0V
(a)
S
V_{GS}
G
D
S
1K
D
On
V_{GS} + ve
S
D
Off
V_{GS} 0V
(b)

Figure 14.14 Metal oxide semiconductor (MOS) transistors: (a) *n* channel; (b) *p* channel

Figure 14.15(c) shows a CMOS NOR gate. If *A* or *B* is high, one of Q_3 or Q_4 will be on taking the output *Z* low (with one of Q_1, Q_2 off). When both *A* and *B* are low, Q_1, Q_2 will both be on and Q_3, Q_4 off taking the output high.

The high input impedance of FETs can present handling problems, and early devices could be irreparably damaged by static electricity from, say, nylon clothing or leakage currents from unearthed soldering irons. Modern CMOS now includes protection diodes and can be treated like any other component.

Another effect of the high input impedance is the tendency for unused inputs to charge to an unpredictable voltage. *All* CMOS inputs must go somewhere; even unused inputs on unused gates on multigate packages must go to a supply rail thereby forcing a 1 or 0 state.

CMOS is generally sold in the so called 4000 series which is a rationalisation of the original RCA COSMOS and Motorola McMOS ranges. 'B' suffix CMOS denotes buffered signals and improved protection; needless to say the B devices are better suited for industrial systems.

14.2.7 Emitter coupled logic (ECL)

ECL is the fastest commercially available logic family, and with care it can operate at 500 MHz. At these speeds,

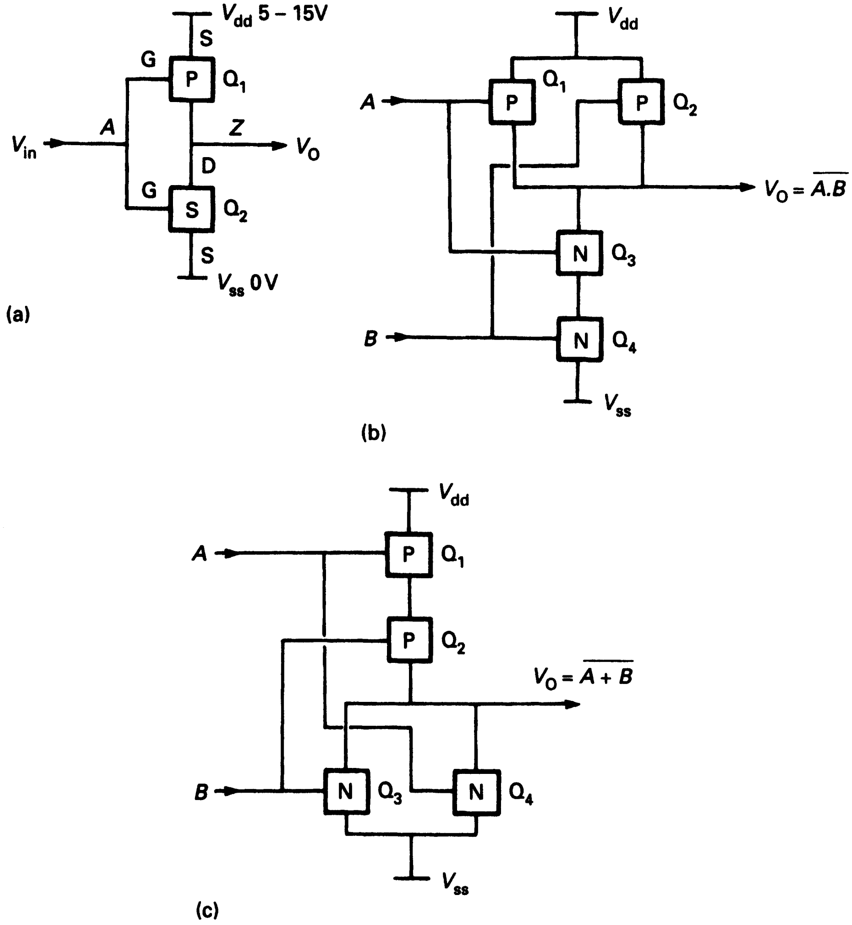

V_{dd} 5 – 15V
S
G
P
Q_1
V_{in}
A
Z
V_o
D
G
S
Q_2
S
V_{ss} 0V
(a)
V_{dd}
Q_1
A
P
P
Q_2
$V_O = \overline{A.B}$
N
Q_3
B
N
Q_4
V_{ss}
(b)
V_{dd}
A
P
Q_1
P
Q_2
$V_O = \overline{A + B}$
B
N
Q_3
N
Q_4
V_{ss}
(c)

Figure 14.15 Complementary metal oxide semiconductor (CMOS) logic gates: (a) NAND gate; (b) NOR gate

however, extreme care needs to be taken with the circuit board layout to avoid crosstalk and power supply induced noise. ECL obtains its speed from the use of non saturating transistors and high power levels (around 60 mW per gate compared with CMOS figure of 0.01 mW). The logic levels in ECL are −0.8 V and −1.6 V (giving a rather poor noise immunity of 0.25 V). ECL is very fast, but its odd voltage levels, strict wiring and power supply requirements and poor noise immunity preclude its use in industrial applications except where very high speed is needed.

14.2.8 Open collector and tri-state outputs

The TTL NAND gate of *Figure 14.16*(*a*) has a single output transistor rather than the usual totem pole output. The output is connected to Vcc by an external pull up resistor. This is known as an open collector output. In *Figure 14.16*(*b*) the outputs of several open collector gates are connected in parallel with a single pull-up resistor. A half circle on the gate output symbol is often, but by no means universally, used to show an open collector output. The output *Z* will be high if, and only if, all the parallel gates have high outputs. Using positive logic conventions the paralleled outputs provide a positive AND function on the outputs of the input gates.

Another description of the operation is the output *Z* will be low if any of the outputs are low. This will occur if (*A* and *B* are high) OR (*C* and *D* are high) OR (*E* and *F*) are high. The linking of collectors can be considered to perform a negative OR function on the outputs of the input gates and can give a possibly complex function for the cost of a single pull up resistor. Its main disadvantage is a poor rising edge (caused by the RC time constant of the pull up resistor and stray capacitance) and a slight degradation of noise immunity.

Open collector gates are a possible solution for applications where many devices communicate via a bus system, the backplane of a computer is a typical application. A more common approach for bus systems though is the tri-state gate. Strictly speaking tri-state is a registered trade mark of National Semiconductors. The term tri-state is a bit of a misnomer as the gate does not have three logic levels but rather three logic states: high, low and disconnected. A tri-state gate has normal inputs plus a separate control input which enables the gate or puts the output into a high impedance (disconnected) state. *Figure 14.17*(*a*) shows the symbol for a tri-state two input NAND gate and *Figure 14.17*(*b*) shows three tri-state buffers which are used to route data from *A*, *B* or *C* to output *Z* as selected by control inputs *L*, *M* or *N*. It should be noted that this is fundamentally different from the wire AND open collector circuit of *Figure 14.16*.

14.2.9 Schmitt triggers

Many logic elements require fast edges to operate correctly. Edges can be degraded for a variety of reasons; stray capacitance or a non digital device for example. The Schmitt trigger always gives fast edges on its output signals regardless of the edge speed of the input signals.

The transfer function of a conventional gate is shown on *Figure 14.18*(*a*). The transfer function of a Schmitt trigger incorporates hysteresis and is shown on *Figure 14.18*(*b*). *Figure 14.18*(*d*) shows a slow changing input and the resulting output, which always has fast clean edges.

A Schmitt trigger has the conventional logic symbol with an added hysteresis loop similar to *Figure 14.18*(*c*). Schmitt triggers are usually available in hex inverter or quad two input gate ICs. The 74123, for example, is a popular quad two input Schmitt trigger NAND gate.

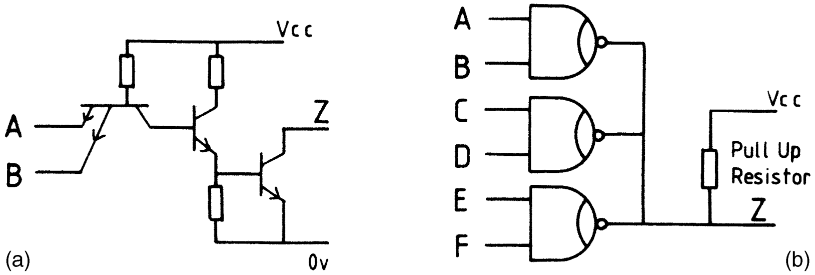


Figure 14.16 An open collector TTL NAND gate: (a) circuit diagram; (b) logic function using open collector gates. The linking of the collectors gives a positive AND or negative OR function depending on the interpretation of the logic states

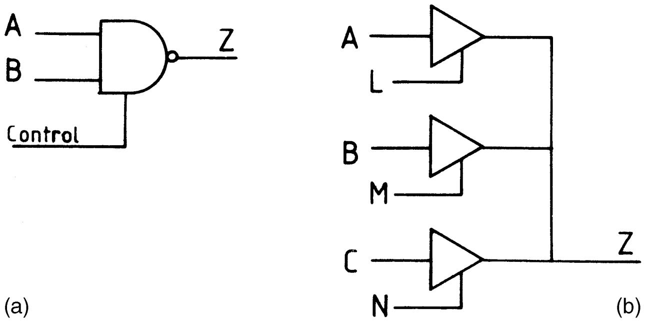


Figure 14.17 Tri-state gates: (a) tri-state NAND; (b) tri-state data selection

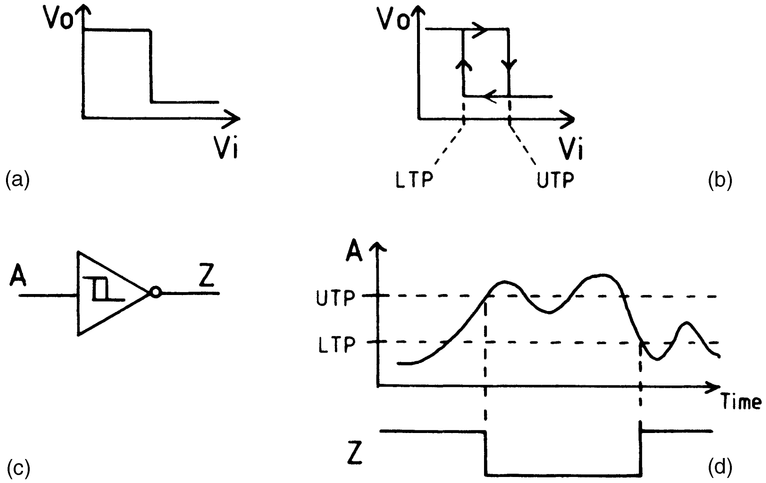


Figure 14.18 The Schmitt trigger: (a) operation of a conventional inverter; (b) operation of an inverter with hysteresis; (c) logic symbol; (d) use of a Schmitt trigger to convert a slowly changing signal to a crisp digital signal

Comparison of *Figures 14.18*(*a*) and *14.18*(*b*) shows that a Schmitt trigger has better noise immunity than a conventional gate. They are therefore commonly used for interfacing to slow and possibly noisy signals from the outside world.

14.2.10 Choosing a logic family

Until the latter part of the 1980s the designer really had to choose between TTL (with the low powered Shottky (LS) family being the popular choice) and CMOS. The latter was slower and had a much smaller range of devices, but had the advantages of very low power consumption, better noise immunity and a wide supply tolerance. Since then, though, there has been a tendency for the families to merge.

The trend started with the 74C series of CMOS which provided CMOS devices with the same pinning as TTL, but with CMOS B series electrical characteristics (slower than TTL, but with 3 to 15 V supply). These were useful, but the major impact was the introduction of the 74HC, 74AC, 74HCT and 74HCT families. These use improved technologies, were as fast as TTL, and (as their name implies) they follow the 74 series pinning. Taking them in turn:

74AC is the high speed member of the family, capable of operating at speeds of 125 MHz. The voltage supply range is 3 to 6 V (essentially TTL with a wider tolerance), and the transfer characteristic is the standard CMOS near ideal symmetrical curve of *Figure 14.19*(*a*).

74HC is a near replacement for LS TTL with an operating speed of 30 MHz. Other characteristics are similar to 74AC.

As mentioned earlier, the output levels of TTL, shown on *Figure 14.19*(*b*), are approximately 0.5 V in the 0 state, and 4.5 V in the 1 state. Standard forms of TTL can, just, connect to 74AC or 74HC devices, but the resultant noise immunity is poor. Two further forms of CMOS were developed with a transfer function whose input side mimicked a TTL device. These are known as 74ACT (high speed version) and 74HCT (practically a direct replacement for LS TTL). These should NOT be viewed as a family to be used for a complete project, as to do so would give the poorer TTL noise immunity. They are, though, exceedingly useful when a circuit has to mix TTL and CMOS devices.

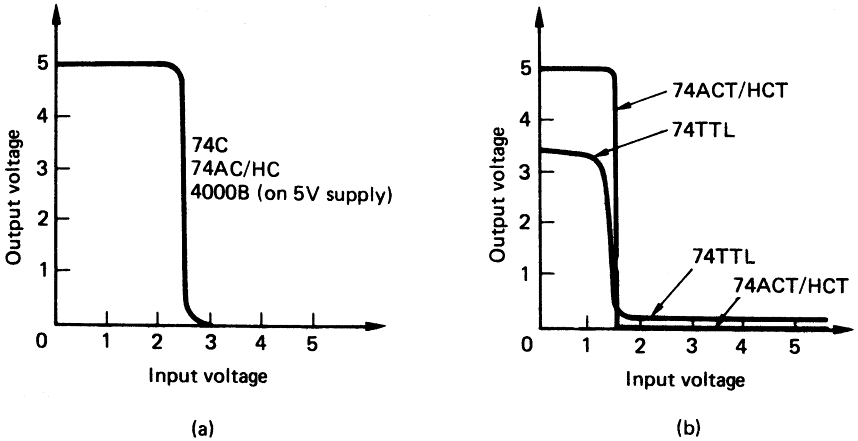


Figure 14.19 Transfer functions for CMOS and TTL: (a) CMOS; (b) TTL and CMOS ACT/HCT

In CMOS, therefore, there is now:

Family	*Supply*	*Speed* (MHz)	*Comments*
4000B	3 to 15 V	2	Useful for battery circuits, slow, seems unlikely to develop further
74C	3 to 15 V	2	As 4000B with TTL pinning
74AC	2 to 6 V	125	Very fast. TTL pinning. Power rises with speed so normal CLOS low power may not be relevant. Care needed with layout
74HC	2 to 6 V	30	Designed as direct replacement for LS TTL (but with CMOS signal levels)
74ACT	2 to 6 V	125	As 74AC with TTL input levels
74HCT	2 to 6 V	30	As 74HC with TTL input levels

An interesting development is the view that the corner supply pins on TTL are not the best arrangement for power supply and ground noise, and there seems to be a move toward centre pinning on some high speed CMOS circuits.

There are four TTL families in common current use; LS, ALS, F and AS. It is worth listing these in tabular form:

Family	*Speed* (MHz)
74LS	25
74ALS	35
74F	100
74AS	105

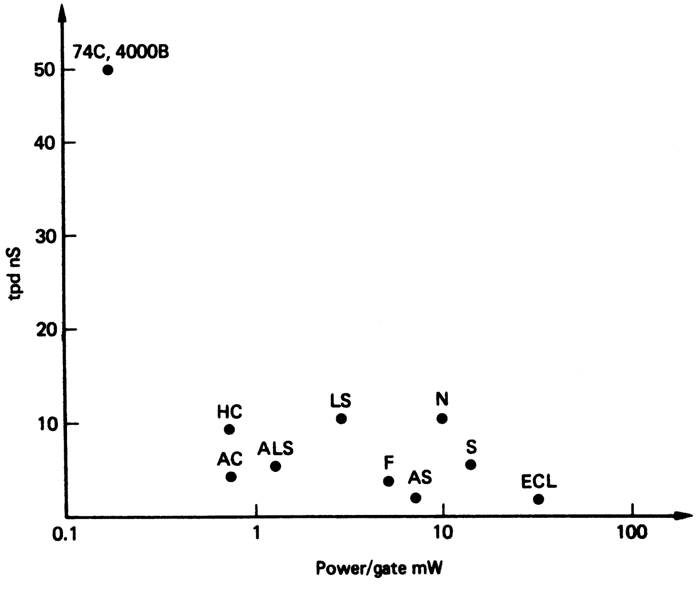


Figure 14.20 Comparison of logic families operating at about 1 MHz

All require a 5 V +/−0.25 V supply. The suffix in the above table appears as part of the device identification; a 74LS06, for example, is a low power Schottky gate.

Choosing a device is quite straightforward. First where there is little choice; for out and out speed use ECL (but remember the precautions needed to avoid noise). For high speed, use 74AC (but again take care with the layout).

Battery circuits are best designed with 4000B or 74C devices, the supply is less critical, and both will run on a 9 V battery until it is flat without the need of a regulator circuit. Being slower they are also less prone to noise.

For ‘cooking’ logic, 74HC seems best suited with a reasonable speed, lower power and better noise immunity than LS TTL. The only problem is an incomplete coverage of the TTL family at present (the useful 7490/92 counters are missing for example), so the odd LS or ALS TTL circuit may be needed, with 74HCT devices being used as interfaces between TTL and CMOS.

Figure 14.20 shows a comparison between these families.

14.3 Combinational logic

14.3.1 Introduction

Combinational logic is based around the block diagram of *Figure 14.21*(*a*). Such systems have several inputs and one, or more, outputs. The output states are uniquely defined for each and every combination of inputs and the ‘block’ does not contain any device such as storage, timers or counters. We therefore have n inputs I_1 to I_n and Z outputs Q_1 to Q_Z. In systems with multiple outputs it is usually easier to consider each separately as *Figure 14.21*(*b*), allowing us to consider the circuit as Z blocks, each different but represented by *Figure 14.21*(*c*).

The number of possible input states depends on the number of inputs:

For two inputs there are four input combinations
For three inputs there are eight input combinations
For four inputs there are sixteen input combinations

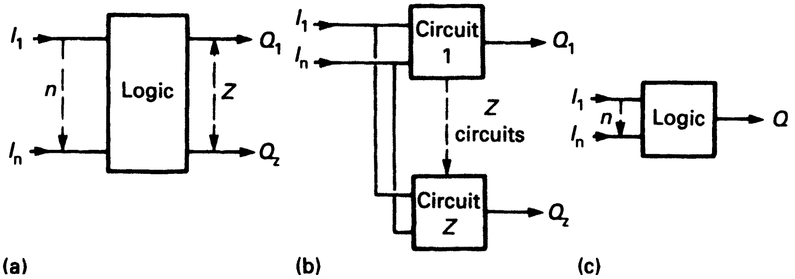


Figure 14.21 Combinational logic block diagrams: (a) the generalised problem with *n* inputs and *Z* outputs; (b) problem split into *Z* independent circuits; (c) one of the *Z* circuits with a single output

and so on. Not all of these may be needed. There are frequently only a certain number of input combinations that may occur because of physical restrictions elsewhere in the system.

The design of combinational logic systems first involves examining all the input states that can occur and defining the output states that must occur for each and every input state. A logic design to achieve this is then constructed from the gates described in Section 14.1.3. In many systems the design can be done in an intuitive manner, but the rest of this section describes more formal design procedures.

Few real life systems need pure combinational logic, most need storage and similar dynamic functions. Such systems can be analysed and designed considering them as smaller subsystems linked together. The design of dynamic systems is discussed in Section 14.8.

14.3.2 Truth tables

A truth table is a useful way of representing a combinational logic circuit, and can be used to design the circuit needed to achieve a desired function.

Suppose we have three contacts monitoring some event (overpressure in a chemical reactor for example) and we wish to construct a majority vote circuit. If the three switches are called *A*, *B*, *C* and the majority vote *Z* this would have the truth table:

A	B	C	Z	
0	0	0	0	
0	0	1	0	
0	1	0	0	
0	1	1	1	<-
1	0	0	0	
1	0	1	1	<-
1	1	0	1	<-
1	1	1	1	<-

It can be seen that *Z* is 1 for:

	$\bar{A}$ and *B* and *C*
or	*A* and $\bar{B}$ and *C*
or	*A* and *B* and $\bar{C}$
or	*A* and *B* and *C*

The desired logic function can then be constructed directly from the truth table as *Figure 14.22*. In general, the circuit derives from a truth table will consist as a set of AND gates whose outputs are OR'd together. This form of circuit is known as a *Sum of Products* (see Section 14.4.3),

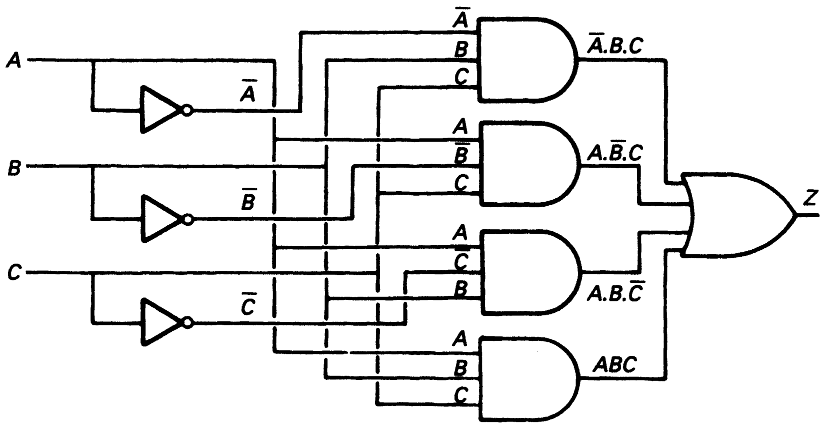


Figure 14.22 Non minimal implementation of majority vote logic direct from the truth table

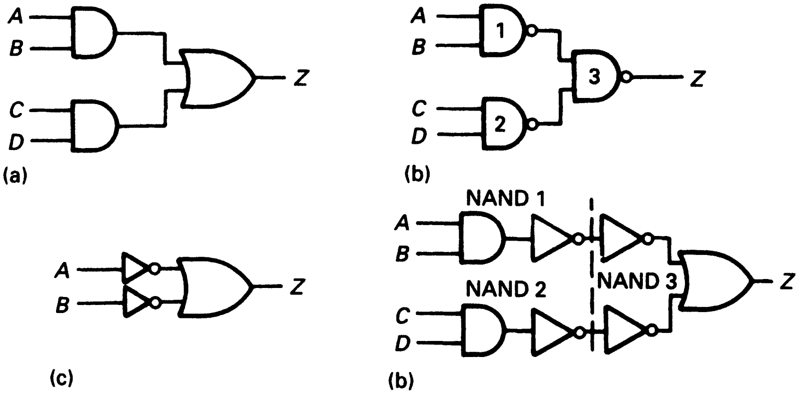


Figure 14.23 Sum of products (AND/OR) logic implementation based solely on NAND gates: (a) required function; (b) NAND based circuit; (c) representation of a NAND gate; (d) circuit b redrawn in the style of representation c. The inverters cancel giving the required function

and one of the reasons for the popularity of NAND gates is that an *s* of *p* expression can be formed purely with NAND gates.

A truth table design always gives a design which works and is logically correct, but does not always give a circuit which uses the minimum combination of gates. To do this we need one of the other techniques described below.

Consider the expression

$Z = (A \ \& \ B)$ OR $(C \ \& \ D)$

This has the simple circuit of *Figure 14.23*(*a*), which obviously fulfils the logic function. Consider, however, the totally NAND based circuit of *Figure 14.23*(*b*). Straightforward, if laborious, testing of all possible sixteen input states will show that it behaves identically to *Figure 14.23*(*a*). In some mysterious way, the right-hand NAND gate is behaving as an OR gate.

This rather surprising fact is a result of De Morgan's theorem, described in the next section. Intuitively, however, we can see the reason by drawing up the truth table for the OR gate preceded by inverters as *Figure 14.23*(*c*):

A	B	Z
0	0	1
0	1	1
1	0	1
1	1	0

This is the same as a NAND gate, so a NAND gate can, with legitimacy, be drawn as *Figure 14.23*(*c*).

The circuit of *Figure 14.23*(*b*) could now be drawn as *Figure 14.23*(*d*) with the ingoing NANDs drawn as ANDs followed by an inverter, and the outgoing NAND by the arrangement of *Figure 14.23*(*c*). Obviously the intermediate inverters cancel, leaving the equivalent circuit of *Figure 14.23*(*a*).

14.3.3 Boolean algebra

In the nineteenth century a Cambridge mathematician and clergyman George Boole, devised an algebra to express and manipulate logical expressions. His algebra can be used to represent, design and minimise combinational logic circuits.

The AND function is represented by a dot (.), so

$Z = A.B$

means Z is 1 when A is 1 AND B is 1. Often the dot is omitted (e.g. $Z = AB$)

The OR function is represented by an addition sign (+), so

$Z = A + B$

means Z is 1 when A is 1 OR B is 1.

The invert function is represented by a bar –, so

$Z = \bar{A}$

means Z takes the opposite state to A. Some sources use the ' to denote inversion so $\bar{A}$ and A' both mean the inverse of A.

Boolean algebra allows complex expressions to be written in a concise manner and can also be used to simplify expressions. To achieve this, a series of rules are used. The first eleven of these are self obvious (or can be visualised by considering the equivalent relay circuits).

(a) $A.1 = A$
(b) $A.0 = 0$
(c) $A + 1 = 1$
(d) $A + 0 = A$
(e) $A.A = A$ (e and f are known as the *Idempotent laws*)
(f) $A + A = A$
(g) $\bar{\bar{A}} = A$ (known as the *Involution law*)
(h) $A.\bar{A} = 0$ (h and i are known as the *Complementary laws*)
(i) $A + \bar{A} = 1$
(j) $A + B = B + A$ (j and k are known as the *Commutative laws*)
(k) $A.B = B.A$

The next two laws, called the *Associative laws*, allow us to group brackets around variables with the same operator

(l) $(A + B) + C = A + (B + C) = A + B + C$
(m) $(A.B).C = A.(B.C) = A.B.C$

The next two laws are called the *Absorption laws*, and tell us what happens if the same variable appears with AND and OR operators

(n) $A + A.B = A$
(o) $A.(A + B) = A$

The next laws, called the *Distributive laws*, tell us how to factorise Boolean equations

(p) $A + B.C = (A + B).(A + C)$
(q) $A.(B + C) = A.B + A.C$

In general, Boolean expressions can be expressed in two forms. The first form, called *product of sums,* or P of S, brackets OR terms and ANDs the results for example:

$$Z=(A+\bar{\boldsymbol{B}}).(B+C+D).(\bar{\boldsymbol{A}}+\bar{D})\Leftarrow$$

The second form, called *sum of product*s, or S of P, groups AND terms and ORs the results, for example:

$$Z=(A.B.\bar{\boldsymbol{D}})+(\bar{\boldsymbol{B}}.C)+(A.\bar{\boldsymbol{D}})\Leftarrow$$

Truth tables, described in Section 14.4.2, inherently give an S of P result.

The complementary function of a Boolean expression yields the inverse of the expression (i.e. where the expression yields 1, the complement yields 0). The expressions $(A+B)$ and $(\bar{\boldsymbol{A}}.\bar{\boldsymbol{B}})$ for example, can be shown to be complementary by simply constructing their truth tables.

The last two laws, known as *De Morgan's theorem*, show how to form the complement of a given expression (and gives one way to interchange S of P and P of S forms).

(r) $\overline{(A+B)}=\bar{\boldsymbol{A}}.\bar{\boldsymbol{B}}$

(s) $\overline{(A.B)}=\bar{\boldsymbol{A}}+\bar{\boldsymbol{B}}$

In its formal representation, De Morgan's theorem appears rather daunting. It can be more easily expressed:

To form the complement of an expression

(1) Replace each '+' in the original expression with '.' and vice versa.
(2) Complement each term in the original expression.

For example, to complement the expression $\bar{A}+B.C$:

Step 1, replace '+' by '.' and '.' by '+' giving:

$$\bar{A}.(B+C)\Leftarrow$$

Step 2, complement each term

$$A.(\bar{\boldsymbol{B}}+\bar{\boldsymbol{C}})\Leftarrow$$

which is the result.

Boolean Algebra can be used to minimise logical expressions, but the method is rarely obvious, and it is easy to make errors with double bars and swapping of '.'s and '+'s. Minimisation by Boolean algebra makes good examination questions, but is rarely used in practice. An easier way to achieve minimisation is to use the graphical Karnaugh map, described below.

14.3.4 Karnaugh maps

A Karnaugh map is an alternative way of presenting a truth table. The map is drawn in two dimensions; two, three and four variable maps are shown on *Figure 14.24*.

Each square within the map represents one line on the truth table. For example:

square X represents $A=1$, $B=0$ which can be written $A.\bar{\boldsymbol{B}}$
square Y represents $A=0$, $B=1$, $C=1$ which can be written $\bar{\boldsymbol{A}}.B.C$
square Z represents $A=1$, $B=0$, $C=1$, $D=0$ which can be written $A.\bar{\boldsymbol{B}}.C.\bar{\boldsymbol{D}}$

The essential feature of a Karnaugh map is the way in which the axes are labelled. It will be seen that only one variable changes for a move between any adjacent horizontal or vertical squares

The use of this feature is not immediately apparent, but consider *Figure 14.25*. This contains four terms giving a 1 output. These are:

$$A.\bar{\boldsymbol{B}}.C.\bar{D},\quad A.B.C.\bar{\boldsymbol{D}},\quad A.\bar{\boldsymbol{B}}.C.D,\quad A.B.C.D$$

so we could write (quite correctly)

$$Z=A.\bar{\boldsymbol{B}}.C.\bar{\boldsymbol{D}}+A.B.C.\bar{\boldsymbol{D}}+A.\bar{\boldsymbol{B}}.C.D+A.B.C.D$$

Examination of the map, however, shows that the D variable and B variable can change state without affecting the output. The circled squares, in fact, represent AC, so the above expression can be simplified to

$$Z=AC$$

Groups of two adjacent cells on a three variable map represent some combination of TWO of the three variables. On *Figure 14.26*(*a*), groupings for $A.B$ and $C.\bar{\boldsymbol{B}}$ are shown. This map represents

$$Z=\bar{\boldsymbol{A}}.B+C.\bar{\boldsymbol{B}}$$

Two adjacent cells on a four variable map represent some combination of three of the four variables. On *Figure 14.26*(*b*), groupings for $\bar{\boldsymbol{A}}.B.C$, $B.\bar{\boldsymbol{C}}.\bar{\boldsymbol{D}}$, $A.\bar{\boldsymbol{B}}.\bar{\boldsymbol{D}}$ and $\bar{\boldsymbol{B}}.\bar{\boldsymbol{C}}.D$ are shown. This map thus represents

$$Z=\bar{\boldsymbol{A}}.B.C+B.\bar{\boldsymbol{C}}.\bar{\boldsymbol{D}}+A.\bar{\boldsymbol{B}}.\bar{\boldsymbol{D}}+\bar{\boldsymbol{B}}.\bar{\boldsymbol{C}}.D$$

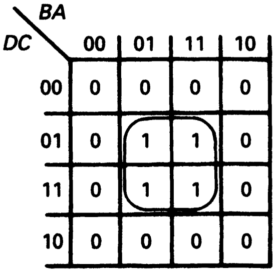
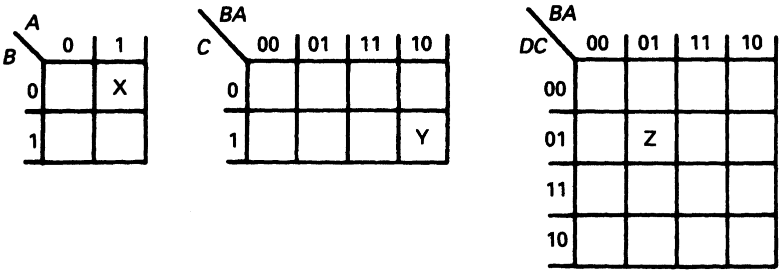
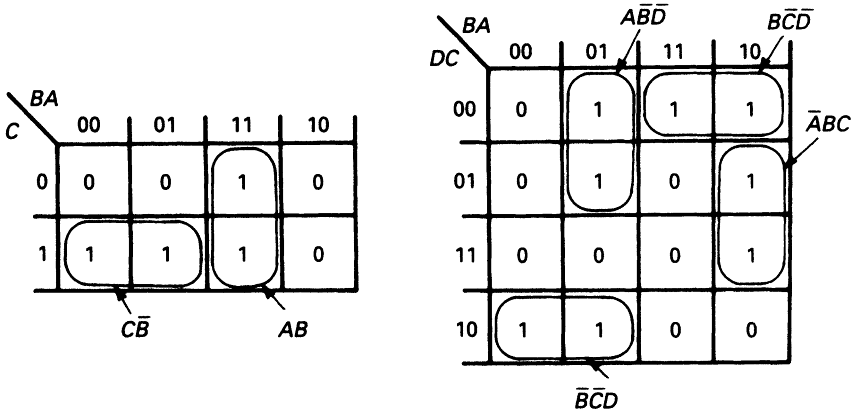


Figure 14.25 Minimisation of $Z=A.\bar{\boldsymbol{B}}.C.\bar{\boldsymbol{D}}+A.B.C.\bar{\boldsymbol{D}}+A.\bar{\boldsymbol{B}}.C.D+A.B.C.D$ to $Z=A.C$ using a Karnaugh map



(a) (b) (c)

Figure 14.24 Karnaugh maps: (a) two variable; (b) three variable; (c) four variable



(a)

BA C	00	01	11	10
0	0	0	1	0
1	1	1	1	0

(b)

BA DC	00	01	11	10
00	0	1	1	1
01	0	1	0	1
11	0	0	0	1
10	1	1	0	0

Figure 14.26 Grouping of two adjacent cells: (a) on a three variable map; (b) on a four variable map

Groups of four adjacent cells on a three variable map represent a single variable. The group on *Figure 14.27*(*a*) represents the variable A, hence

$Z = A$

Groups of four adjacent cells on a four variable map represent some combination of two of the four variables. The groups on *Figure 14.27*(*b*) represent $\bar{B}\bar{D}$ and BD. The map represents

$Z = \bar{B}.\bar{D} + B.D$

A group of eight adjacent cells on a four variable map represent a single variable. The group on *Figure 14.28* represents C and $\bar{B}$, so

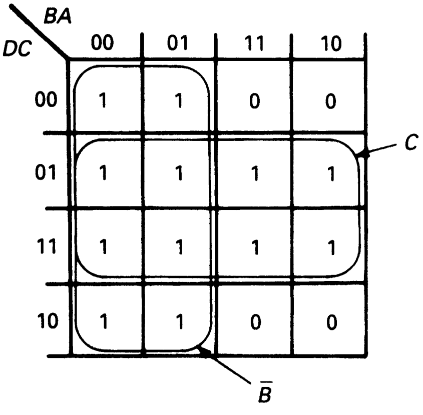
$Z = C + \bar{B}$

It is important to realise that top and bottom edges are considered adjacent as are right and left sides. Grouping can therefore be made around the tops and sides as *Figure 14.29* which represents

$Z = \bar{A}.C + A\bar{C}$

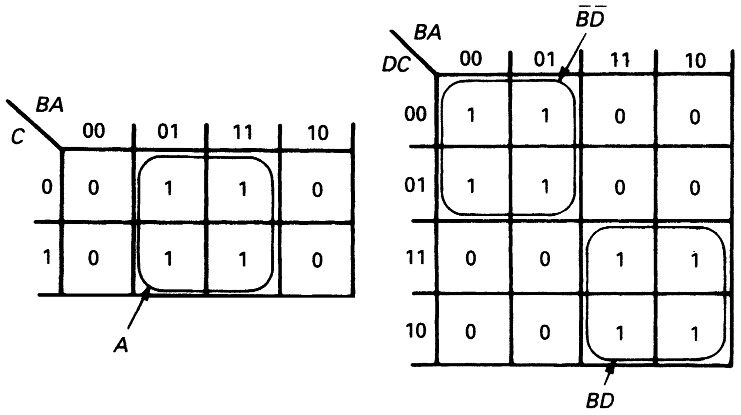
The rules for minimisation using Karnaugh maps are simple and straightforward:

(1) Plot the Boolean expression or truth table onto the Karnaugh map
(2) Form new groups of 1s on the map. Groups must be rectangular and contain 1, 2, 4 or 8 cells. Groups should



BA DC	00	01	11	10
00	1	1	0	0
01	1	1	1	1
11	1	1	1	1
10	1	1	0	0

Figure 14.28 Grouping of eight adjacent cells



(a)

BA C	00	01	11	10
0	0	1	1	0
1	0	1	1	0

(b)

BA DC	00	01	11	10
00	1	1	0	0
01	1	1	0	0
11	0	0	1	1
10	0	0	1	1

Figure 14.27 Grouping of four adjacent cells: (a) on a three variable map; (b) on a four variable map

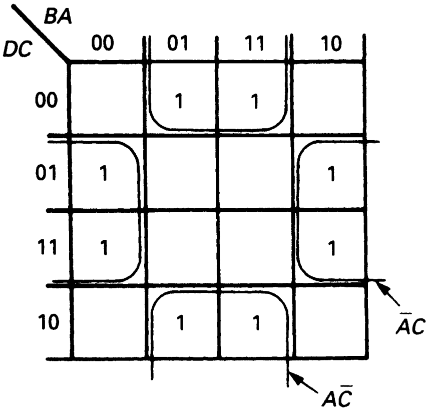


Figure 14.29 Top and bottom sides are adjacent

be as large as possible and there should be as few groups as possible. Do not forget overlaps and possible round the edge groupings.

(3) From the map, read off the expression for each group. The minimal expression is then obtained in S of P form, and can be directly implemented in AND/OR gates or NAND gates

Figure 14.30(*a*) shows a majority vote circuit (2 out of 3) plotted onto a Karnaugh map and grouped as *Figure 14.30*(*b*). It will be seen that this has three terms giving the simple NAND based circuit of *Figure 14.30*(*c*).

14.3.5 Conversion between P of S and S of P representations

It is occasionally required to translate an S of P expression into a P of S expression and vice versa

These are most conveniently handled in the form

$Z = \Leftarrow\Sigma(N_1, N_2, N_3 \ldots)\Leftarrow$for S of P

and

$Z = \Leftarrow A(N_1, N_2, N_3 \ldots)\Leftarrow$for P of S

where N_n is the numerical equivalent of the binary pattern at the corresponding gate input. If, for example, a gate input is C, B, $\bar{A}$, N will be 6 corresponding to 110.

The first step is to note the largest number, which determines how many bits we are dealing with (three bits for seven or less, four bits for fifteen or less and so on.) Call the maximum number corresponding to this number of bits N_{max} (seven, fifteen, thirty-one etc.).

Note the *unused* numbers in the expression to be converted. For each unused number N_{un} there will be a number $(N_{max} - N_{un})$ in the expression in the other form. For example, to convert the S of P expression

$Z = \Leftarrow\Sigma(1, 4, 5, 6)\Leftarrow$

to P of S form we first note N_{max} is seven (three bits). The terms in the P of S representation will be given by

Unused S of P	0 2 3 7 N_{un}
P of S	7 5 4 0 $(N_{max} - N_{un})$

Giving a P of S representation of

$Z = \Leftarrow A(0, 4, 5, 7)$ which is the equivalent to the original S of P expression.

The method for reverse conversion is identical.

14.3.6 Formal minimisation, the Quine-McCluskey method

The Karnaugh map is an excellent way of minimising combinational logic, but is essentially limited to five inputs and relies on human intuition. More formal methods are needed for more complex functions. The most common of these is Quine-McCluskey. The method can deal with any number of inputs, but is lengthy and error prone for direct human implementation. It is, however, ideally suited for computer implementation.

The start point is an S of P expression in the form

$Z = \Leftarrow\Sigma(N_1, \varsigma N_2, \varsigma N_3 \ldots)\Leftarrow$

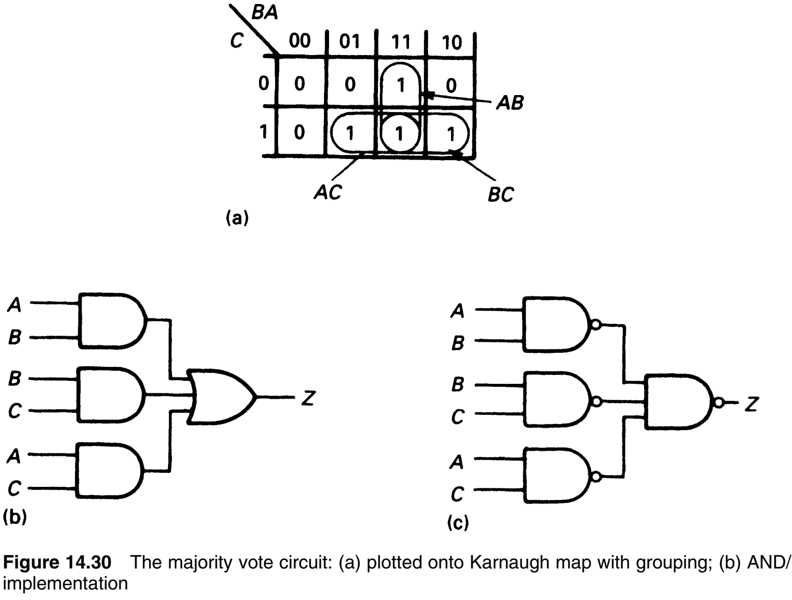


Figure 14.30 The majority vote circuit: (a) plotted onto Karnaugh map with grouping; (b) AND/OR implementation; (c) equivalent NAND based implementation

From this the minterms are grouped according to whether they have one, two, three etc. 1s in them. For example, with

$$Z = \Sigma(2, 3, 4, 5, 6, 7, 9, 11, 12, 13)$$

we would group them
Minterms with one 1

0010 (2)
0100 (4)

Minterms with two 1s

0011 (3)
0101 (5)
0110 (6)
1001 (9)
1100 (12)

Minterms with three 1s

0111 (7)
1011 (11)
1101 (13)

Each term in each group is compared with each term in the group immediately below. If one *and only one* digit difference is found, a new entry in the lower group is formed with X replacing the single differing digit. Comparing 0010 with 0011 gives a new entry of 001X in the lower table. For the above groups this gives:

One 1	
0010	a, b
0100	c, d, e

Two 1s			Created
0011	a	:f, g	a 001X n
0101	c,	:h, I	b 0X10 o
0110	b, d	:j	c 010X p, s
1001		:k,l	d 01X0 q
1100	e	:m	e X100 r .

Three 1s		Created
0111	f,h,j	f 0X11 o
1011	g,j	g X011 #
1101	i,l,m	h 01X1 q
		i X101 r
		j 011X n,p
		k 10X1 #
		l 1X01 #
		m 110X s
		n 0X1X #
		o 0X1X duplicate
		p 01XX #
		q 01XX duplicate
		r X10X #
		s X10X duplicate

The letters a, b, c etc. show the comparisons made and the groups created. Any group which does not create a new group is called a *prime implicant*, denoted by # above. These are X011, 10X1, 1X01, 0X1X, 01XX, X10X (representing *DCBA*, *A* being the least significant as usual) with X denoting don't care. X011 is thus $\bar{C}.B.A$. An S of P circuit based on these prime implicants will work, but is not necessarily minimal.

Next a chart is drawn of these prime implicants, as shown on *Figure 14.31* where each prime implicant is represented by a ♦. For four bit numbers, each minterm with full four bits will have one ♦, with one X there will be two ♦s and with two crosses four ♦s. X011, for example, represents minterms 3 and 11. The first stage is to identify columns with only a single ♦. The corresponding prime implicants MUST be in the final expression. These are noted down and all the corresponding ♦s marked for each row as these are now covered.

For each column, if there is a marked ♦ in the column, all ♦s in the column can now be marked, as this minterm has been included. If a prime implicant has all its ♦s marked it is redundant and can be deleted (e.g. 01XX).

There will probably be one or more ♦s left unmarked. Choose from the remaining prime implicants to give the best grouping. Give preference to minterms with the largest number of Xs, and remember that once a single ♦ is marked in a column, all the ♦s in the column can be marked. When all ♦s have been marked a solution has been reached.

Following this procedure for *Figure 14.31* gives:

$$Z = A\bar{C}D + B\bar{D} + \bar{B}C$$

a result which could, in all honesty, have been arrived at much faster with a Karnaugh map and common sense. The procedure is, however the basis of computer minimisation of logic circuits as used in PLA and PAL configuration programs.

14.3.7 Hazards, races and glitches

Gate propagation delays discussed in Section 14.2.2 can cause unwanted random pulses to appear in logic circuits. These unwanted pulses are known variously as *hazards*, *races* or *glitches*.

The logical output of *Figure 14.32*(*a*) should always be zero since

$$Z = A.\bar{A} = 0$$

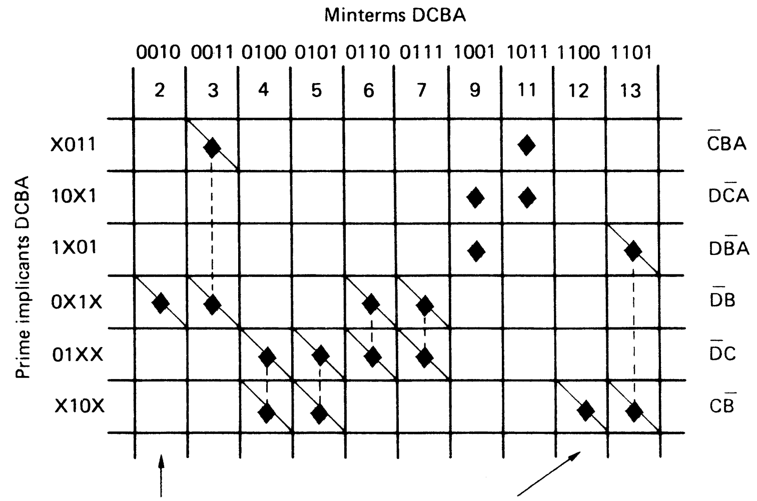
In practice, however, $\bar{A}$ will be delayed by the propagation delay of an inverter giving the possible waveforms of *Figure 14.32*(*b*). As *A* changes a small pulse may appear at the output.

Glitches are not always immediately obvious. A similar problem can occur with the NAND based AND/OR circuit of *Figure 14.33*(*a*). This implements the relationship

$$Z = A.B + \bar{A}.C$$

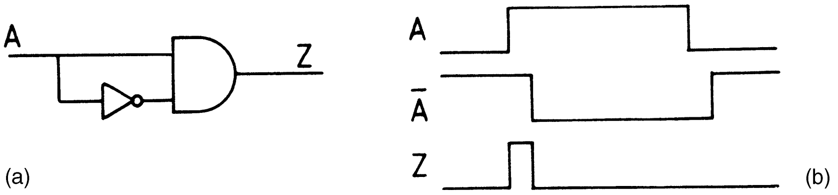
As before $\bar{A}$ must be obtained from some form of inverter as *Figure 14.33*(*b*). The circuit is logically correct but if $B = C = 1$ then the circuit is behaving in a similar way to *Figure 14.32*(*a*). If *B* and *C* are both 1 and *A* changes state a small pulse will probably appear at the output.

Plotting *Figure 14.33* onto a Karnaugh map as *Figure 14.34*(*a*) shows a way to identify and eliminate glitches. There are two groups on the map; *AB* and $\bar{A}C$. Moving between $AB = 11$ and $AB = 01$ we move between groups. This corresponds to *A* changing from 1 to 0 or 0 to 1. A potential glitch has adjacent 1s not covered by the same group.



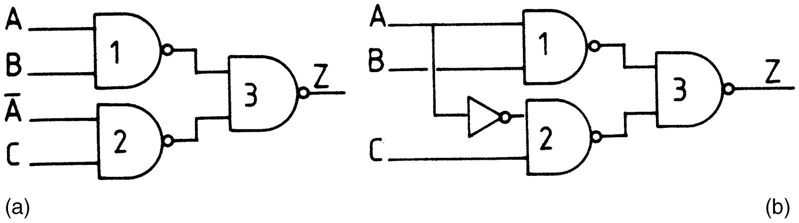
Start with columns with only one ◆. These are essential ($\bar{D}B$ and $C\bar{B}$). With these ◆s marked, $\bar{D}C$ is seen to be redundant, and $D\bar{C}A$ will mark remainder

Figure 14.31 Prime implicant chart after essential implicants and subsequent minterms have been marked



(a) (b)

Figure 14.32 An obvious glitch producing circuit: (a) logic diagram, the output should always be '0'; (b) actual circuit behaviour



(a) (b)

Figure 14.33 Non obvious glitch producing circuit: (a) logic diagram; (b) redrawn to show source of the glitch

To remove the risk of a glitch we add an additional group as *Figure 14.34(b)*. There are now no adjacent 1s not in the same group. The resulting circuit is shown on *Figure 14.34(c)*. Note that the group *BC* is logically redundant and is included solely to prevent glitches when *A* changes state with $B=C=1$. Glitch free circuits are often non minimal.

Glitches may not always be important. In general, if the output of a glitch prone circuit is not feeding directly (or indirectly) a counter, storage device or timer the glitches will probably have no effect. Glitches can also be ignored by using clocked synchronous systems. Different logic families have different propensities for generating and ignoring) glitches. The important factor is the relationship between edge speeds and propagation delays. CMOS, with edge speed similar to or longer than the propagation delay, has a useful tendency to ignore glitches. ECL, with very fast edge speeds, is very prone to glitches.

14.3.8 Integrated circuits

Many complex functions are available in IC form, and a circuit designer should aim to minimise cost and the number of IC packages rather than the number of gates. A minimisation exercise, whether by Boolean algebra or Karnaugh map, should always be preceded by a search of an IC catalogue for a suitable off the peg device.

14.3.9 UCLAs, PALs and PLAs

An integrated circuit consists of a small slice of silicon into which is etched the various individual unconnected components required to make the required circuit. These are then connected by a thin metallised layer to form the required device function. An UCLA (for *uncommitted logic array*, also known as an ULA) consists initially of a large

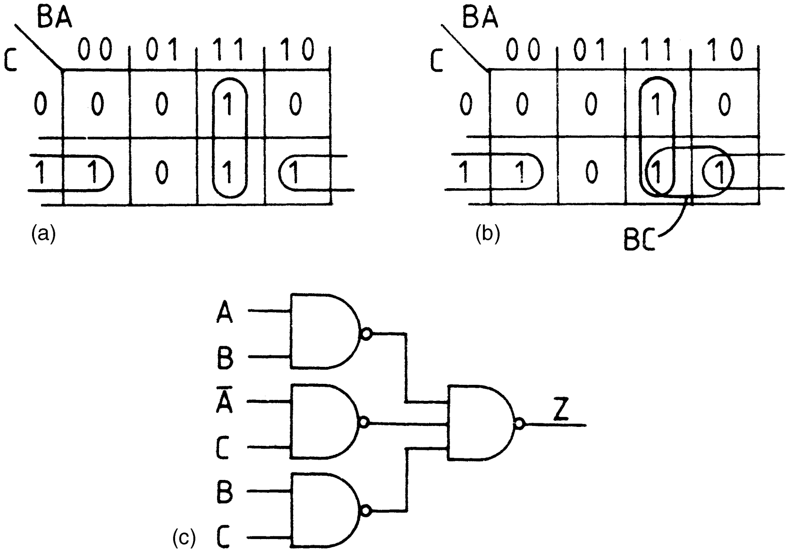


Figure 14.34 Glitch free design using a Karnaugh map: (a) original minimal grouping; (b) *BC* term added to remove the glitch. The final grouping is non minimal; (c) the resulting glitch free, but non minimal, logic

number of assorted gates, storage and memories but without the metallised interconnection layer. The user specifies the required circuit which is then formed by the design of the metallised layer. The basic IC silicon slice (which is the expensive part) is thus common to many users and the relatively cheap metallisation layer is specific to one user's application. UCLAs therefore allow designers to have their own ICs at a reasonable price. They are, though, only cost effective for reasonable volume production runs.

An alternative approach, suitable for smaller volumes, is *programmable logic*. These are essentially a combination of true/complement inputs with an AND/OR output as shown on *Figure 14.35*. Each connection point is originally linked, but can be blown open by the designer (using a programming terminal) to leave the desired function. The original devices were based on bipolar construction, and literally used small metallic fuses. Once blown, they could not be re-used. Later MOS devices can be erased by UV light in a similar way to EEPROMs.

The simplest devices use a programmable AND combination (selected from the true/complement inputs) with a fixed AND/OR logic. These are known as *Programmable Array Logic*, or PALs. The more versatile (but more complex) arrangement of *Figure 14.36* uses programmable AND plus programmable OR connections. These are known as *Programmable Logic Arrays* or PLAs, (this distinction is not quite true, the terms PLA and PAL are used interchangeably by some manufacturers).

Figures 14.35 and *14.36* are essentially combinational logic in sum of product (S of P) form (see Section 14.3.2). Sequential programmable logic is also available, and is typically of the form of the *Figure 14.37* based around an AND/OR/D-type circuit. These are known as *registered* or *sequential* PALs. They are very useful for building logic networks built around state transition diagrams (see Section 14.8)

There are some disadvantages. Early devices had a voracious power appetite, several hundred mA for some. The later MOS devices are better, but their use should be questioned on battery driven devices. There is also a one-off investment needed in a programming terminal, and programming languages such as ABEL, CUPL and PALASM. These work out the required link blowing from a designer specified logic function defined in combinational or state transition form. They do not, however, check out for glitches and even seem to encourage them by aiming for truly minimal logic. Some care is needed by the designer, but the languages do allow redundant combinations to be specified to give glitch free circuits.

With bipolar devices, it should also be remembered that bipolar devices cannot be reprogrammed if an error is made. Mistakes with bipolar programmable logic are not cheap, and even with MOS versions, erasure with UV light is not instantaneous.

Programmable logic is very popular where standard boards (with fixed connections to the outside world) can be used in different applications. Typical examples are vending and ticket machines, interface devices or testing of a logic circuit before building the final version.

14.4 Storage

14.4.1 Introduction

Most logic systems require some form of memory. A typical relay circuit is the motor starter circuit of *Figure 14.38* which 'remembers' which of the two operator push buttons was pressed last. The memory is achieved by the latching contact A1.

14.4.2 Cross coupled flip flops

The logical equivalent of *Figure 14.38* is the cross coupled NOR gate circuit of *Figure 14.39*(*a*). Assume both inputs are 0, and output Q is at a 1 state. The output of gate a will be 0, and the two 0 inputs to gate b will maintain Q in its 1 state. The circuit is therefore stable. If the reset input is now taken to a 1, Q will go to a 0. and $\bar{Q}$ to a 1. Similar

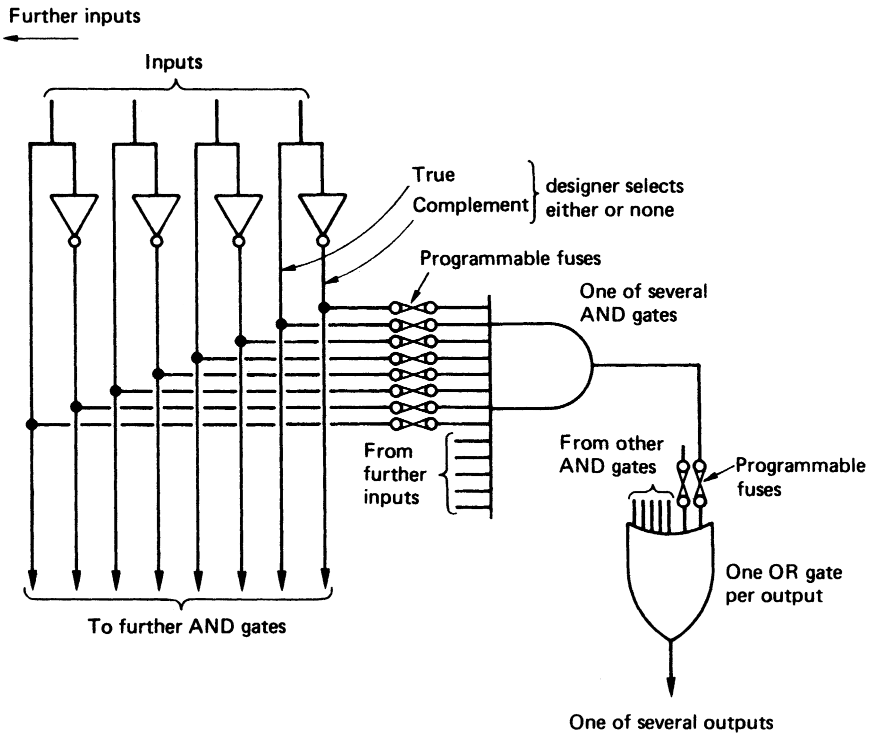

Further inputs
Inputs
True
Complement
designer selects either or none
Programmable fuses
One of several AND gates
From further inputs
From other AND gates
Programmable fuses
One OR gate per output
To further AND gates
One of several outputs

Figure 14.35 The basis of programmable logic

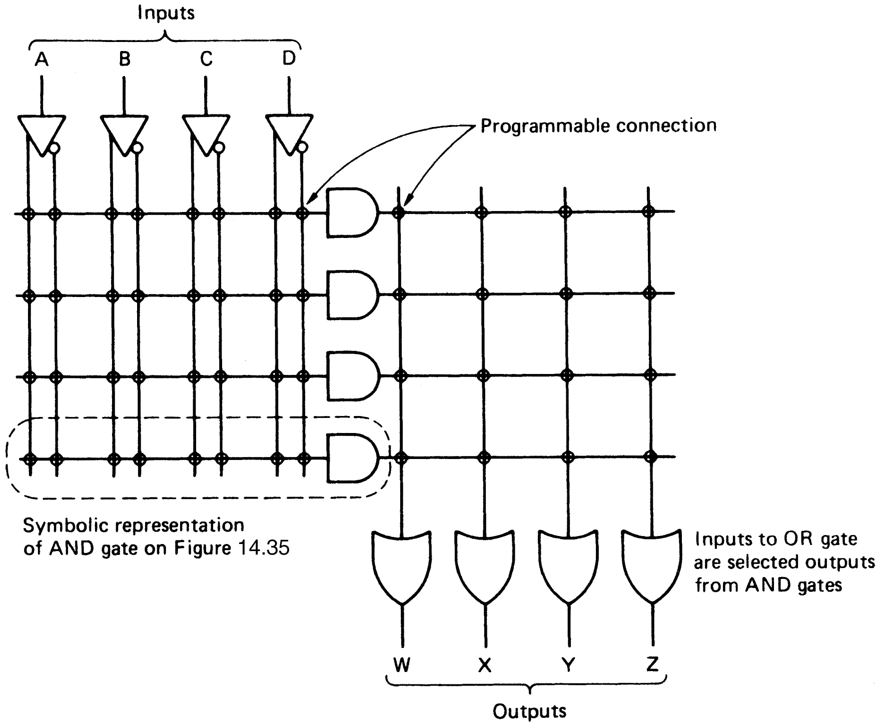

Inputs
A
B
C
D
Programmable connection
Symbolic representation of AND gate on Figure 14.35
Inputs to OR gate are selected outputs from AND gates
W
X
Y
Z
Outputs

Figure 14.36 A programmable logic array with AND/OR inputs both programmable

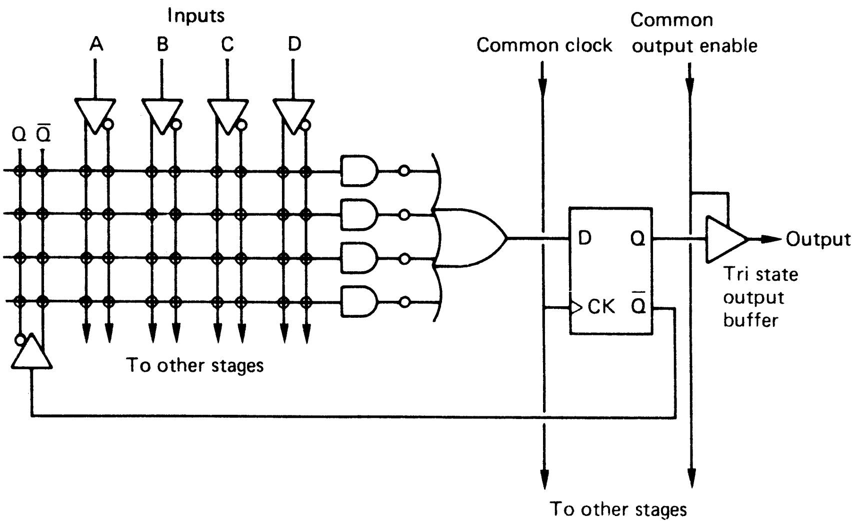


Figure 14.37 Sequential programmable logic with tri-state outputs. A typical device would have eight inputs and eight D type flip flops

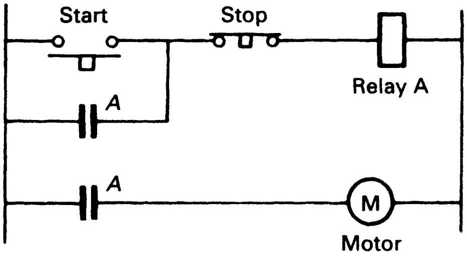


Figure 14.38 A simple relay storage circuit used to start a motor. The circuit remembers which button (Start or Stop) was last pressed

analysis to that above will show that the circuit is stable in this state, even when the reset input goes back to a 0.

The set input can be used now to switch the Q output to 1 and the $\bar{Q}$ back to a 0. The set and reset inputs cause the output to change state, with the outputs indicating which input was last at a 1 state as summarised by *Figure 14.39*(*b*). If both inputs are 1 together, both outputs go to a 0, but this condition is normally disallowed.

The cross coupled NOR gate circuit is called an *RS Flip Flop*, and is shown on logic diagrams by the symbol of *Figure 14.39*(*c*).

It is also possible to construct a cross coupled flip flop from NAND gates as *Figure 14.40*(*a*). Analysis will show that this behaves similar to *Figure 14.39*, but the circuit remembers which input last went to a 0 as shown on *Figure 14.40*(*b*). The logic symbol for a NAND based RS flip flop is shown on *Figure 14.40*(*c*); the small circles on the input showing that the flip flop responds to 0 inputs.

14.4.3 D type flip flop

The D type flip flop shown on *Figure 14.41*(a) has a single data input (D), a clock input and the usual Q and $\bar{Q}$ outputs. Superficially this is similar to the latch memory above, but the clock operates in a more subtle way. The operation of a typical D type flip flop is shown on *Figure 14.41*(*b*). The clock samples the *D* input when the clock input goes from a 0 to 1, but the output changes state when clock goes from 1 to 0. The significance of this is explained below in Section 14.4.6.

There are several ways in which a D type flip flop can be implemented. A common circuit uses the master/slave arrangement of *Figure 14.41*(*c*). When the clock input is 1, the *D* input sets, or resets, the master flip flop. When the

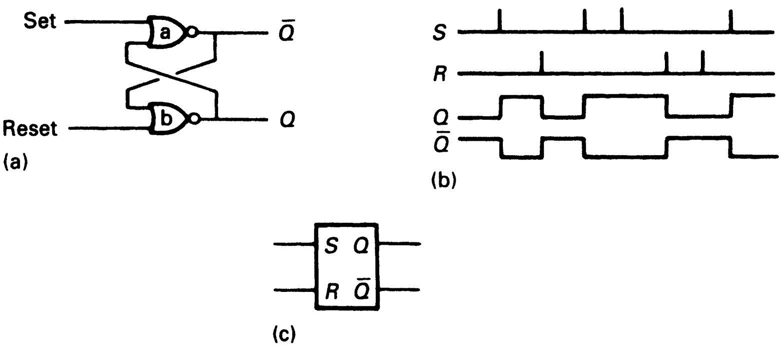


Figure 14.39 A NOR based RS flip flop. This circuit remembers which input was last a '1': (a) logic diagram; (b) operation; (c) logic symbol

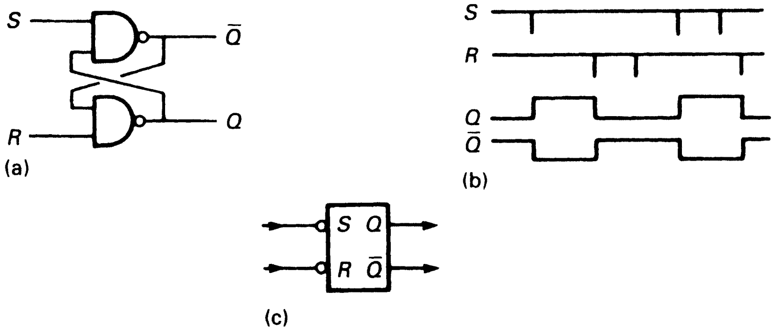

S
R
$\bar{Q}$
Q
(a)
S
R
Q
$\bar{Q}$
(b)
S Q
R $\bar{Q}$
(c)

Figure 14.40 A NAND based RS flip flop. This circuit remembers which input was last a ‘0’: (a) logic diagram; (b) operation; (c) logic symbol

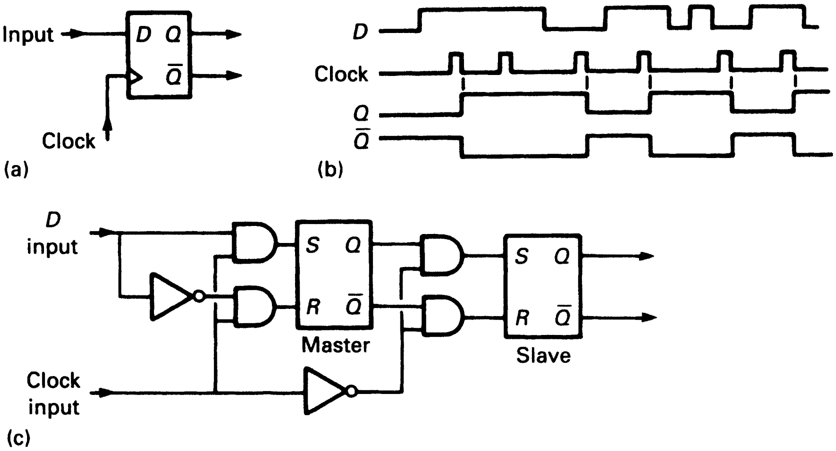

Input
D Q
$\bar{Q}$
Clock
(a)
D
Clock
Q
$\bar{Q}$
(b)
D input
S Q
R $\bar{Q}$
Master
S Q
R $\bar{Q}$
Slave
Clock input
(c)

Figure 14.41 The D type flip flop: (a) logic symbol; (b) operation; (c) logic diagram for a master/slave D type

clock input is 0 the state of the master flip flop is transferred to the slave flip flop (and the outputs take up the state of *D* when the clock input was 1). Note that the master flip flop is isolated from the *D* input whilst the clock is 0.

Although it would be feasible to construct a master/slave flip flop from discrete gates, integrated circuit *D* types (such as the TTL 7474 or the CMOS 4013) are readily available.

14.4.4 The JK flip flop

In Section 14.4.2 the NOR based RS flip flop was described, and it was stated that the input state $R=S=1$ was normally disallowed. The JK flip flop, shown on *Figure 14.42*(*a*) is a clocked RS flip flop with additional logic to cover this previously disallowed state. The clock input acts as described above for the D type flip flop, i.e. sampling the inputs on one edge, and causing the outputs to change on the other.

The outputs after a clock pulse for $J=1$, $K=0$; $J=0$, $K=1$; $J=0$, $K=0$ are as would be expected for a clocked RS flip flop. If $J=K=1$, the outputs toggle; that is the states of the Q and $\bar{Q}$ interchange. This action is summarised on *Figure 14.42*(*b*).

The toggle state is the basis for counters, described in Section 14.7.

14.4.5 Clocked storage

The D type and JK flip flops described above are examples of clocked storage. The advantages, and implications of this are probably not immediately obvious.

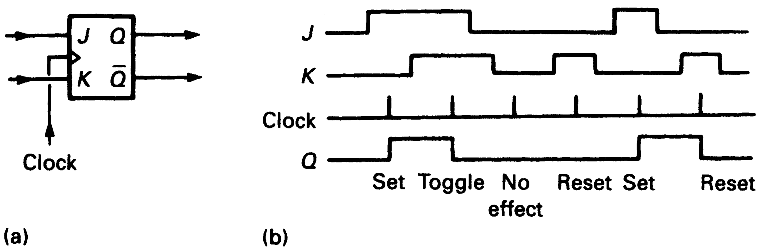

J Q
K $\bar{Q}$
Clock
(a)
J
K
Clock
Q
Set
Toggle
No effect
Reset
Set
Reset
(b)

Figure 14.42 The JK flip flop: (a) logic symbol; (b) operation

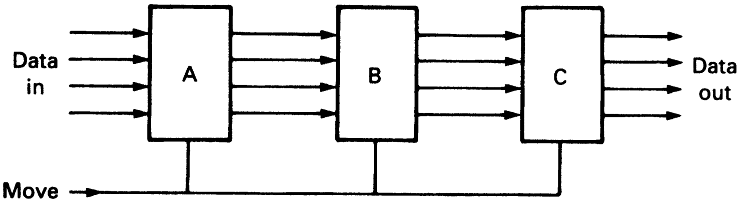


Figure 14.43 Clocked storage. Data moves one position for each pulse on the move line

In all bar the simplest systems, data is often required to be moved around from one storage position to another. In *Figure 14.43*, for example, data is to be moved through stores A, B, C in an orderly manner. If simple flip flops were used along with a signal enable as shown, the data would shoot straight through all the stages. If clocked storage is used, the data will sequence from A to B to C, moving one position for each clock pulse.

14.5 Timers and monostables

Control systems often need some form of timer. Timing functions in logic circuits are provided by devices called monostables or delays. There are many types of delay, although all can be considered as *Figure 14.44(a)*, to consist of an input, Q and $\bar{Q}$ outputs and an RC network which determines the delay period.

The commonest timer, often called the one shot or monostable, gives an output pulse, of known duration, for an input edge. The user can select which edge (0–>1 or 1–>0) triggers the circuit. On *Figure 14.44(b)* a 0–>1 edge is used. Monostables are the basis of all other delay circuits and are widely available (74121, 74122 in TTL, 4047, 4098 in CMOS). Pure delays are shown on *Figure 14.44(c–e)*, and these can be constructed by adding gates to monostable outputs. *Figure 14.44(f/g)* shows the circuit for a delay off.

A variation of the monostable is the retriggerable monostable. In most monostables circuits the timing logic ignores further input edges once started. In a retriggerable monostable each edge sets the timing circuit back to the start again. The action of a retriggerable and normal monostable are compared on *Figure 14.45*.

The time delay of any timer is of the order of RC seconds where R is the value of the timing resistor in ohms and C is the value of the timing capacitor in farads. For delays of more than a few seconds very large values of R and C are

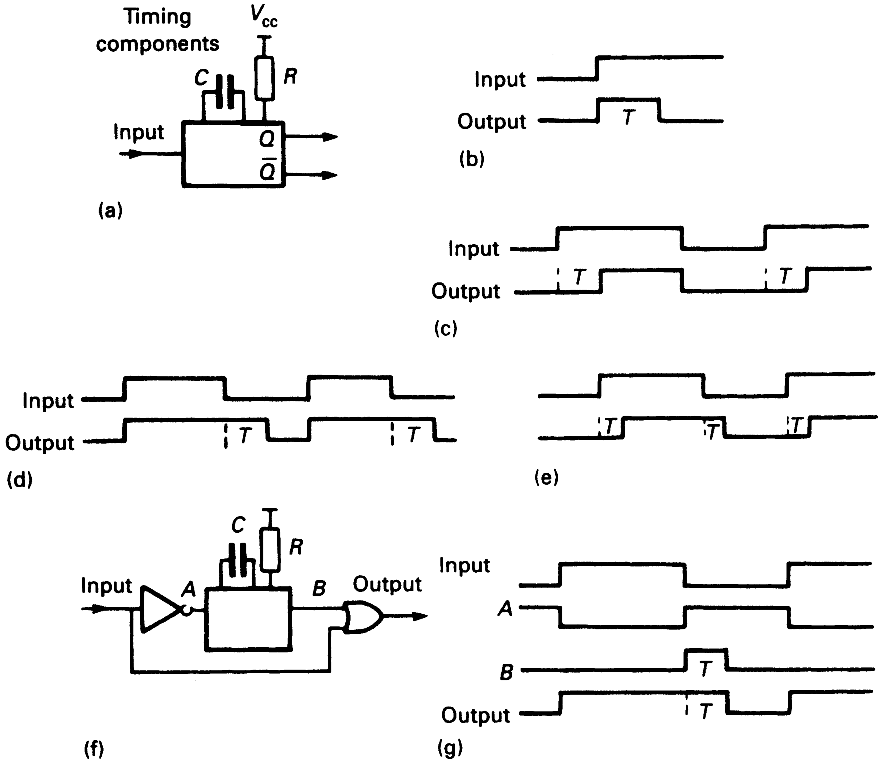


Figure 14.44 Various forms of timers and monostables: (a) basic form of a timer. The timer duration is determined by the values of R and C and is usually of the order of RC seconds; (b) one shot timer, often called a monostable; (c) delay on timer; (d) delay off timer; (e) delay on and off timer; (f) delay off timer built using a simple monostable; (g) timing waveforms for circuit f

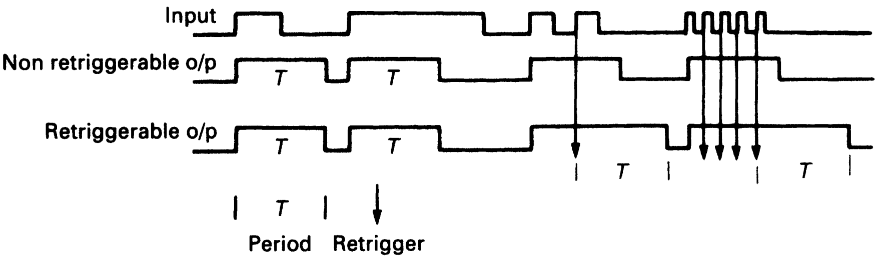


Figure 14.45 A re-triggerable monostable. Each input 0 to 1 edge re-starts the timing function

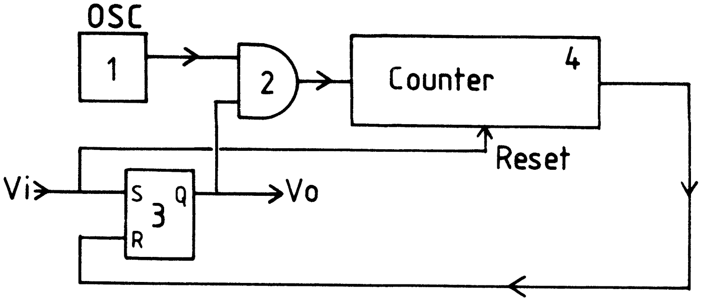


Figure 14.46 Implementation of a long period timer

required. High value resistors are prone to changes in value from leakage and large value capacitors must be electrolytics with problems from leakage, size and long term drift. For periods of more than a few seconds it is usually better to produce a time delay with an oscillator and counter as *Figure 14.46*. The oscillator produces a free running pulse chain which is normally blocked by gate 2. A start pulse sets flip flop 3 and resets the counter. With flip flop 3 set, pulses are passed to the counter which counts up. When the counter reaches a pre-determined count it resets the flip flop. The Q output of the flip flop thus goes high for a time

$T = N \times P$ seconds

where N is the count preset and P the oscillator period. Integrated circuits based on this principle, such as the ZN1034, are available giving very long delays (up to days) with reasonable value components and little problems from drift.

14.6 Arithmetic circuits

14.6.1 Number systems, bases and binary

In previous sections, logic signals have been assumed to represent events such as printer ready, or low oil level. Digital signals can also be used to represent, and manipulate numbers.

We are so used to the decimal number system that it is hard to envisage any other way of counting. Normal every day arithmetic is based on multiples of ten. For example, the number 9156 means:

9 thousands $= 9 \times 10 \times 10 \times 10$
plus 1 hundred $= 1 \times 10 \times 10$
plus 5 tens $= 5 \times 10$
plus 6 units $= 6$

Each position in a decimal number represents a power of ten. Our day to day calculations are done to a base of ten because we have ten fingers. Counting can be done to any base, but of special interest are bases 8 (called *octal*), 16 (called *hex* for *hexadecimal*) and two (called *binary*).

Octal uses only the digits 0–7, the octal number 317, for example, means

```
        3 x 8 x 8      =  decimal 192
plus    1 x 8          =  decimal   8
plus    7              =  decimal   7
              TOTAL       decimal 207
```

Hex uses the letters A–F to represent decimal ten to fifteen, so hex C52, for example, means

```
        12 x 16 x 16   =  decimal 3072
plus    5 x 16         =  decimal   90
plus    2              =  decimal    2
              TOTAL       decimal 3164
```

Binary needs only two symbols, 0 and 1. Each position in a binary number represents a power of two and is called a *bit*, for BInary digiT, most significant to the left as usual, so 101101 is evaluated:

```
        1 x 2 x 2 x 2 x 2 x 2    = 32
plus    0 x 2 x 2 x 2 x 2        =  0
plus    1 x 2 x 2 x 2            =  8
plus    1 x 2 x 2                =  4
plus    0 x 2                    =  0
plus    1                        =  1
              TOTAL         decimal 45
```

Fractions can also be represented in binary, although this is not commonly encountered. Taking fractions as powers of two we get 1/2 (0.1 in binary), 1/4 (0.01 in binary), 1/8 (0.01) and so on. The binary number 110.101 is thus 6 plus 0.5 plus 0.125 giving 6.625.

Conversion from decimal to binary is achieved by successive division by two noting the remainders. Reading the remainders from the top (LSB) to bottom (MSB) gives the binary equivalent. For example, decimal 23

23			
11	r	1	(LSB)
5	r	1	
2	r	1	
1	r	0	
0	r	1	(MSB)

Decimal 23 is binary 10111.

Octal and hex give a simple way of representing binary numbers. To convert a binary number to octal, the binary number is written in groups of three (from the LSB) and the octal equivalent written underneath, for example 11010110

grouped in threes	11	010	110
Octal	3	2	6

Hex conversion is similar, but groupings of four are used. Taking again the binary number 11010110;

grouped in fours	1101	0110
Hex	D	6

The octal number 326 and the hex number D6 are both representations of the binary number 11010110.

14.6.2 Binary arithmetic

Consider the decimal sum:

$$\begin{array}{rr} & 345 \\ + & \underline{272} \\ & 617 \end{array}$$

This is evaluated in three stages:

5 + 2 = 7 no carry
4 + 7 = 11 one down (as result)
plus carry
3 + 2 + carry = 6

At each stage we consider three 'inputs'; two digits and a possible carry from the previous stage. Each stage has two outputs, a sum digit and a possible carry to the next, more significant state. A single digit adder can therefore be considered as *Figure 14.47*(*a*). Several single digit adders can be cascaded, as *Figure 14.47*(*b*), to give an adder of any required number of digits. Note the carry out of the most significant stage becomes the most significant digit.

Binary addition is similar, except that there are only two possible values for each digit. If *Figure 14.47*(*a*) is a binary adder, there are eight possible input combinations:

Inputs			Outputs	
D1	D2	Carry	Sum	Carry
0	0	0	0	0
0	1	0	1	0
1	0	0	1	0
1	1	0	0	1
0	0	1	1	1
0	1	1	0	1
1	0	1	0	1
1	1	1	1	1

An example of binary arithmetic is

$$\begin{array}{rl} 1\;0\;1\;1\;0\;1\;0 & \\ \underline{0\;1\;0\;1\;0\;1\;1} & \\ 1\;0\;0\;0\;0\;1\;0\;1 & \text{Sum (result)} \\ 1\;1\;1\;1\;0\;1\;0\;\; & \text{Carry} \end{array}$$

The implementation of the adder truth table is a simple problem of combinational logic; one possible solution is shown on *Figure 14.47*(*c*). In practice, of course, adders such as the TTL 7483 are readily available in IC form.

Negative numbers are generally represented in a form called *two's complement*. The most significant digit represents the sign, being 1 for negative numbers and 0 for positive numbers. The value part of the number is complemented and 1 added. For example:

+12 in two's complement is 01100 (the MSB 0 indicating a positive number).

To get to two's complement for −12 we complement 1100 giving 0011, set the MSB to 1 giving 10011 then add 1 giving 10100 which is the two's complement representation of −12.

Similarly

+43	0101011	
Complement	1010100	
Add 1	1010101	which is -43.

In each case, addition of the positive and negative number will give the result zero, e.g.

$$\begin{array}{lr} +43 & 0\;1\;0\;1\;0\;1\;1 \\ -43 & \underline{1\;0\;1\;0\;1\;0\;1} \\ & \cancel{1}\;0\;0\;0\;0\;0\;0\;0 \end{array}$$

The top carry is lost, giving the correct result of zero.

Two's complement representation allows subtraction to be done by adding a negative number, for example 12 − 3

$$\begin{array}{rl} 0\;1\;1\;0\;0 & +12 \\ \underline{1\;1\;1\;0\;1} & -3 \\ \cancel{1}\;0\;1\;0\;0\;1 & \end{array}$$

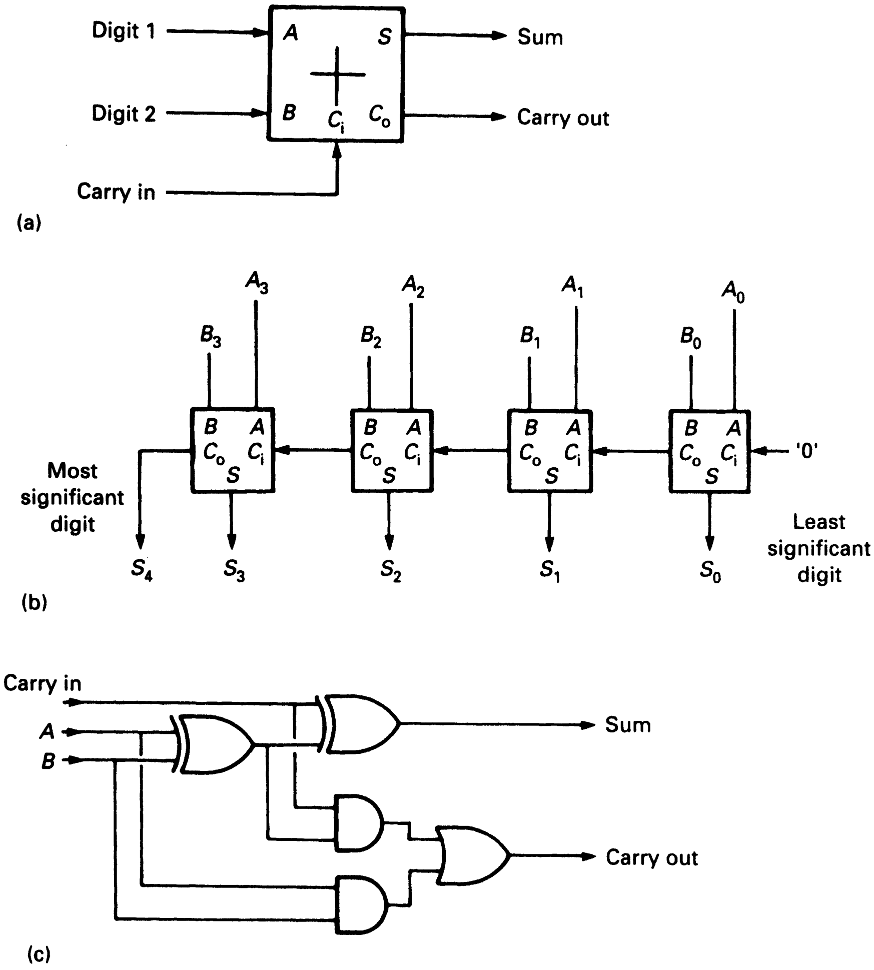

Digit 1
Digit 2
Carry in
A
B
S
C_i
C_o
Sum
Carry out
(a)
A_3
A_2
A_1
A_0
B_3
B_2
B_1
B_0
B A
C_o C_i
S
'0'
Most significant digit
Least significant digit
S_4
S_3
S_2
S_1
S_0
(b)
Carry in
A
B
Sum
Carry out
(c)

Figure 14.47 Adder circuits: (a) representation of a one digit adder. This block diagram will be the same regardless of the number base used; (b) construction of a four digit adder from four identical one digit adders; (c) one bit (i.e. one digit) binary adder logic diagram

The top bit is lost giving the correct result of +9.

Multiplication and division are rarely required in simple logic systems and are generally best implemented with some form of microprocessor assembly in conjunction with specialist mathematical co-processors such as the AMD9511. If a hardware solution is required it can be based on the addition of weighted partial sums. Consider the decimal multiplication

```
  456    Multiplicand
  123    Multiplier
  ---
45600    Partial sums
 9120
 1368
-----
56088    Result
```

The multiplicand is multiplied by each digit of the multiplier in turn and the part results added with appropriate weighting to give the result. Binary multiplication is similar but simpler in that only four multiplication results need to be considered:

$1 \times 1 = 0,\ 1 \times 0 = 0,\ 0 \times 1 = 0,\ 0 \times 0 = 0$

A typical binary multiplication is therefore

```
   1011   Multiplicand (decimal 11)
   1001   Multiplier (decimal 9)
   ----
1011000   (1011 x 1)
 000000   (1011 x 0)
  00000
   1011
-------
1100011   Result (decimal 99)
```

Note that multiplying two four bit numbers can give an eight bit result. If two binary numbers A & B are to be multiplied, therefore, partial sums are obtained by multiplying (i.e. gating) A by each bit of B to form as many partial sums as there are bits in B. These partial sums are then weighted and added to give the result as above.

The fastest multiplication can be obtained by using Read Only Memories (ROMs) programmed with an entire multiplication table. The multiplicand and the multiplier then act as the ROM address and the result is simply read. Two 1 k bit ROMs can form a four by four multiplier with no additional logic. The 74284 & 74285 are integrated circuits designed specifically for this purpose.

Division is even rarer, but can also be performed using ROMs.

14.6.3 Binary coded decimal (BCD)

A single decimal digit can take any value between 0 and 9. Four binary digits are therefore needed to represent one decimal digit. In BCD, each decimal digit is represented by four bits. For example:

9	4	0	7	6
1001	0100	0000	0111	0110

BCD is not as efficient as pure binary. 12 bits in pure binary can represent 0–4095, compared with 0–999 in BCD. BCD, however, has advantages where decimal numbers are to be read from decade switches or sent to digital displays.

14.6.4 Unit distance codes

Figure 14.48 shows a possible application of binary coding. The position of a shaft is to be measured to 1 part in 16 by means of an optical grating moving in front of four photocells. The photocell outputs give a binary representation of the shaft angular position.

Consider what may happen as the shaft goes from position 7 (0111) to position 8 (1000). It is unlikely that all the cells will switch together, so we could get

```
     0111 > 0000 > 1000
or   0111 > 1111 > 1000
```

or any other lengthy sequence of four bits. These possible incorrect intermediate states can be avoided by using a code in which only one bit changes between adjacent positions. Such codes are called *unit distance codes*.

The commonest unit distance code is the *Gray code*, shown in four bit form below.

It will be noted that the code is reflected about the centre. Sometimes the term '*reflected code*' is used for unit distance codes. A unit distance code can be constructed to any even base by taking an equal number of combinations above and below the centre point of a Gray code. A decimal version (called the *XS3 cyclic BCD code*) is also shown. In this code zero is 0010, one is 0110, two is 0111 and so on to nine which is 1010.

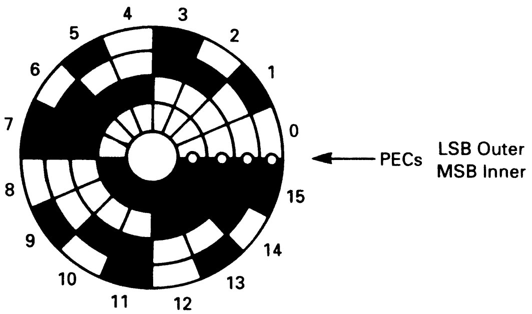


Figure 14.48 Encoding an angular position into a binary signal with a shaft encoder

Conversion between binary and Gray code is straightforward, and is achieved with XOR gates as shown on *Figure 14.49*(*a*) and (*b*).

```
Decimal    Gray
  0        0000
  1        0001
  2        0011
  3        0010  )
  4        0110  )
  5        0111  )
  6        0101  )
  7        0100  ) decimal
-----------------------symmetrical
  8        1100  ) cyclic
  9        1101  )
 10        1111  )
 11        1110  )
 12        1010  )
 13        1011
 14        1001
 15        1000
```

14.7 Counters and shift registers

14.7.1 Ripple counters

Counters are used for two basic purposes. The first, and obvious, use is the counting, or totalising, of external events. The second use of counters is the division of a frequency to give a new, lower frequency.

The 'building block' of all counters is the toggle flip flop which changes state each time its clock input is pulsed. Usually the toggling occurs on the negative edge as shown on *Figure 14.50*(*a*). A toggle flip flop can be constructed from JK or D type flip flops as shown on *Figure 14.50*(*b*, *c*).

If the Q output of a toggle flip flop is connected to the clock input of the next stage as shown on *Figure 14.51*(*a*), a simple binary counter can be constructed to any desired length. *Figure 14.51* is a 3 bit counter with A the LSB and C the MSB. This counts:

Pulse	C	B	A
0	0	0	0
1	0	0	1
2	0	1	0
3	0	1	1
4	1	0	0
5	1	0	1
6	1	1	0
7	1	1	1

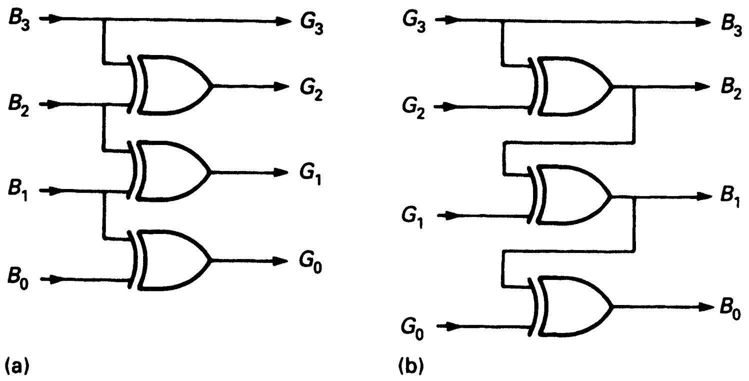


Figure 14.49 Conversion between binary and Gray codes: (a) binary to Gray; (b) Gray to binary

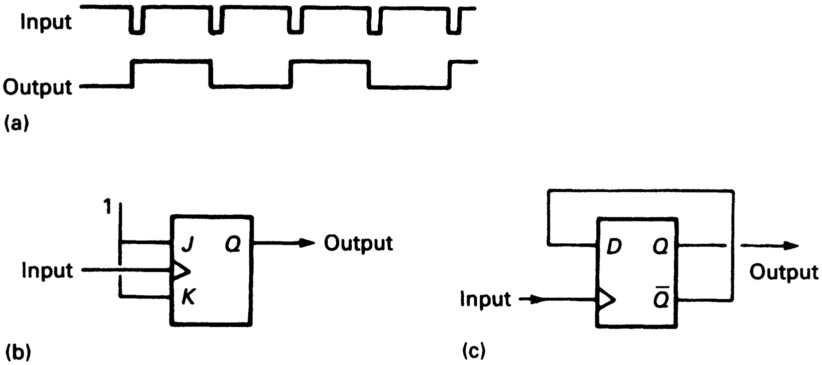


Figure 14.50 The toggle flip flop: (a) operation, the output changes state for each input pulse; (b) a JK toggle flip flop; (c) a D type connected to make a toggle flip flop

Another pulse will take it to state 0 again. It can be seen that *Figure 14.51* is counting up. To count down, the $\bar{Q}$ outputs are connected to the input of the following stage and the signal outputs taken from the Q lines.

There are two limitations to the speed at which a counter chain similar to *Figure 14.51* can operate. The first is the maximum speed at which the first (fastest) stage can toggle. The second restriction is not so obvious.

Consider the case of an 8 bit counter going from 01111111 to 10000000. The LSB toggling causes the next to toggle and so on to the MSB. The change has to propagate through all 8 bits of the counter, so circuits similar to *Figure.14.51* are called *ripple counters*. During the 'ripple' the counter will assume invalid states and cannot be sensibly read. Obviously the propagation delay through all the stages should be considerably less than the input period. High speed applications use synchronous counters, described below.

In *Figure 14.51* the frequency of output C is precisely one eighth of the input frequency. A simple ripple counter can therefore also act as a frequency divider. If we define

$$N = f_{\text{in}}/f_{\text{out}}$$

then $N = 2^m$ for m binary stages.

It will also be seen that the output of any stage of a binary counter has equal mark space ratio regardless of the input mark/space providing the input frequency is constant.

Although it is feasible to construct ripple counters with D type and JK flip flops it is usually more cost effective to use MSI ICs such as the TTL 7493 4 bit counter or the CMOS 4024 7 bit counter. These incorporate features such as a reset line to take the counter to a zero state.

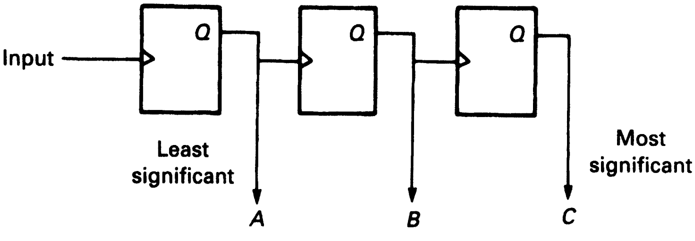


Figure 14.51 A simple three bit binary ripple counter constructed from three toggle flip flops

14.7.2 Synchronous counters

Ripple counters are limited in both speed and length by the cumulative ripple through propagation delay and also temporarily exhibit invalid outputs. Although these limitations are not important in slow speed applications, they can cause difficulties in high speed counting.

These restrictions can be overcome by the use of a synchronous counter where all required outputs change simultaneously. There is no ripple propagation delay through the counters and no transient false count stages. The only speed restriction is the toggling frequency of the first stage.

The building block of a synchronous counter is the JK flip flop/AND gate arrangement of *Figure 14.52*(*a*). If the T input is 1, the JK flip flop will toggle on the receipt of a clock pulse. If the T input is D, the flip flop will not respond to a clock pulse. The carry output is 1 if T is 1 and Q is 1.

A synchronous up counter is constructed as *Figure 14.52*(*b*), which is simply the circuit of *Figure 14.52*(*a*) repeated. Note that the clock input is common to all stages, and the carry from one stage is the T input of the next.

It will be seen that the T inputs, T_b, T_c, T_d will be 1 when all the preceding outputs are 1. T_c will be 1, for example, when A and B are both 1. This is the condition when a counter stage should toggle, taking $DCBA$ from, say, 0011 to 0100.

It is also possible to construct a synchronous down counter by counting the AND gate input of *Figure 14.52*(*a*) to the $\bar{Q}$ output rather than the Q, and observing the counter state on the Q output. A synchronous up/down counter with selectable direction can be constructed as *Figure 14.53*. If the direction line is a 1, gates 1, 2, 3 are enabled, the Q outputs pass to the next stage and the counter counts up. If the direction line is a 0, gates 4, 5, 6 are enabled, the $\bar{Q}$ outputs pass to the next stage and the counter counts down.

14.7.3 Non binary counters

Counting to non binary bases is often required, a BCD count is probably the most common requirement. When the required count is a subset of a straight binary count, (e.g. BCD), the circuit of *Figure 14.54*(*a*) can be used. The counter output is decoded by external logic. When the counter reaches the desired maximum count the decoder

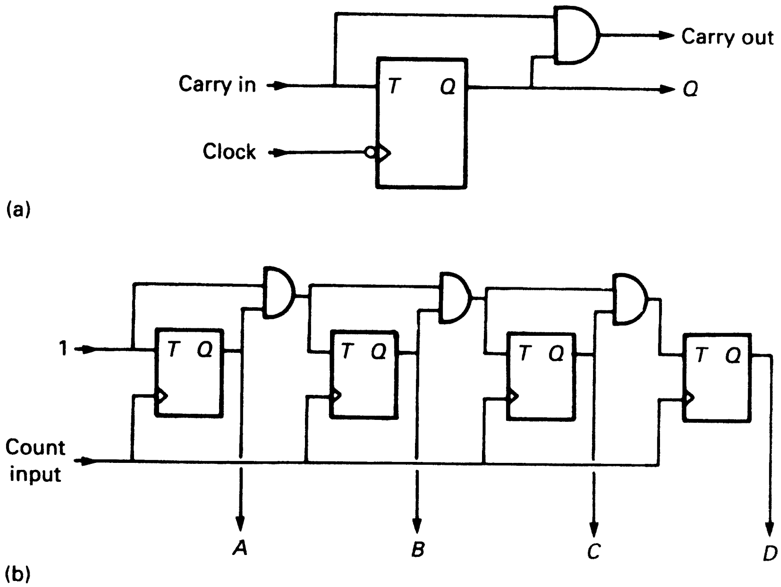


Figure 14.52 Synchronous counters: (a) basic circuit for a synchronous counter; (b) four bit series connected synchronous up counter

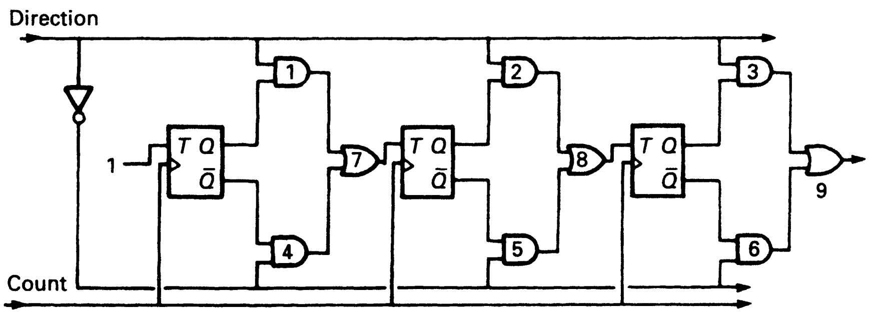


Figure 14.53 Synchronous selectable up/down counter

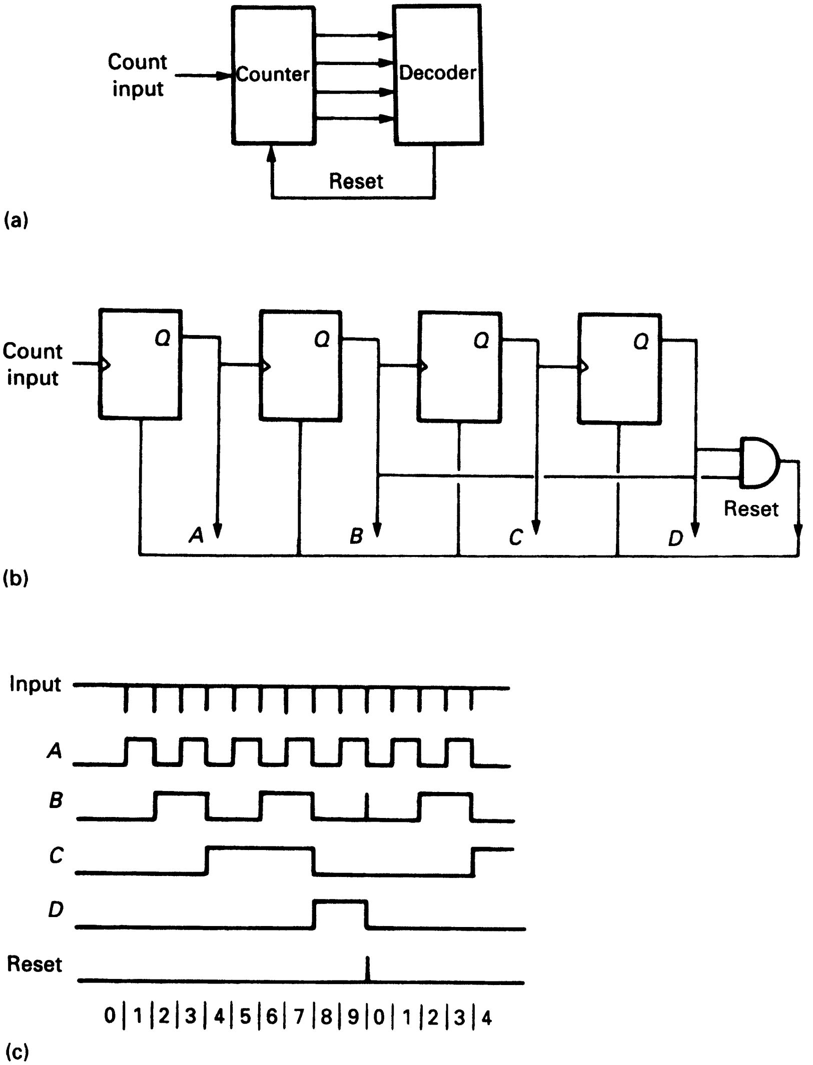


Figure 14.54 Non binary counters: (a) principle of operation; (b) logic diagram for a BCD up counter; (c) counter operation

output forces the counter to its zero state (which is 0000 for a BCD counter, but need not be for other counters).

A single BCD stage constructed on these principles is shown on *Figure 14.54(b)*. The circuit shown is a ripple counter, but could equally well be a synchronous counter. Gate A detects a count of ten (binary 1010) and resets the counter to zero via direct reset inputs on the JK flip flops. Waveforms are shown on *Figure 14.54(c)*.

Where a non binary count is needed (e.g. a Gray code count), it is best to use synchronous counters and an arrangement similar to *Figure 14.55*. This is drawn for D type flip flops, but JK based design is similar.

A combinational logic network looks at the counter outputs and sets the *D* inputs for the next state. If the counter, say, was required to step from 1101 to 0011, the combinational logic output to the *D* inputs would be 0011 for an input of 1101. Effectively there are four combinational circuits in the network, one for each *D* input.

14.7.4 Shift registers

A simple shift register is shown on *Figure 14.56(a)*. Data applied to the serial input, *S* in, will move one place to the right on each clock pulse as shown on the timing diagram of *Figure 14.56(b)*.

Shift registers are used for parallel/serial and serial/parallel conversions. They are also the basis of multiplication and division circuits as a shift of one place towards the MSB is equivalent to a multiplication by 2, and one place towards the LSB an integer division of 2.

14.8 Sequencing and event driven logic

Many logic systems are driven by randomly occurring external events, and follow a sequence of operations. In such systems, the output states do not depend solely on the input

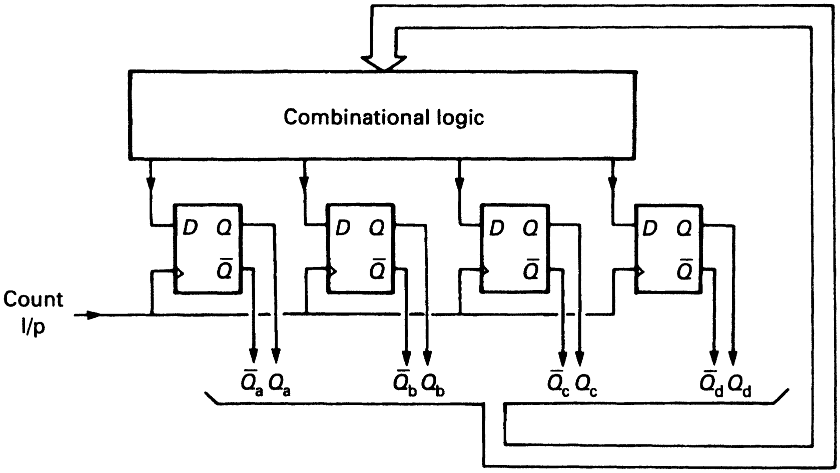


Figure 14.55 Generalised synchronous non binary counter using D type flip flops. Any count pattern can be produced with this arrangement. The principle can also be implemented using JK flips flops.

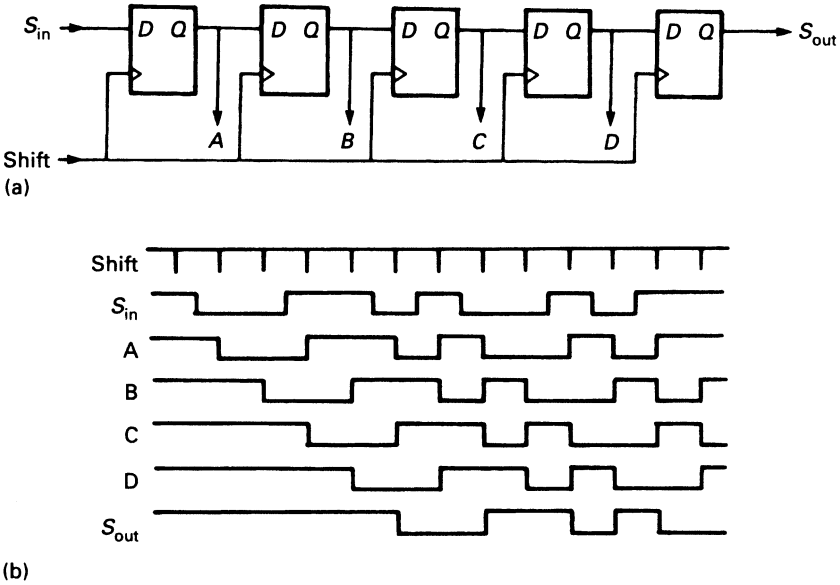


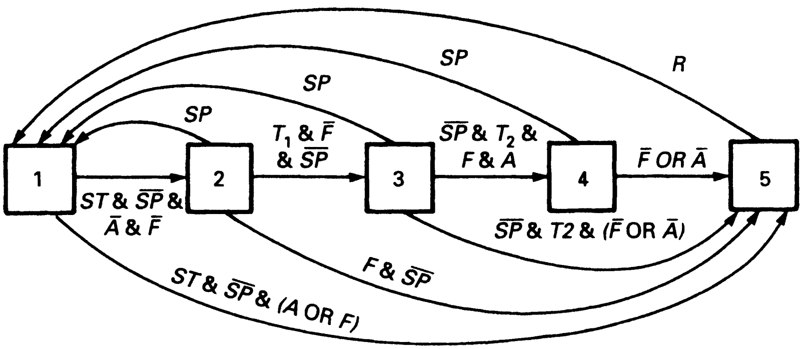
Figure 14.56 Simple shift register constructed from D type flip flops: (a) logic diagram; (b) operation

states, but also on what the system was doing last. These types of systems are said to be sequencing and event driven logic. Sequencing logic is designed using a state diagram. This shows the possible conditions the system can be in, the conditions that are required to move from one state to the next, and the outputs required in each state.

Figure 14.57 shows a possible state diagram for a gas burner control. When the start PB is pressed a 15 second air purge is given (set by timer 1). The pilot valve is opened, and the igniter started for 4 seconds (timer 2). If, at the end of this time, the flame detector shows the flame to be lit, the main gas valve is opened. At any time the stop button terminates the sequence. A non valid signal from the flame detector (i.e. flame present in states 1 and 2 or no flame in state 4) puts the system to an alarm state, as does the incorrect signal from the air flow switch. Note that these are checked for being 'unfrigged' at the start of the sequence.

Event driven logic is built around flip flops, usually one for each state. The flip flop corresponding to state 4 is shown on *Figure 14.58*(*a*), and is set by the required conditions from state 3 and reset by the possible next states (1 and 5). Outputs are simply obtained by ORing the necessary states. The pilot output, shown on *Figure 14.58*(*b*), is simply State 3 OR State 4.

It is possible to minimise event driven circuits to use fewer flip flops, but such an approach is usually not required as it makes the operation more difficult to understand. A straightforward state diagram similar to *Figure 14.57*


SP
R
SP
SP
T_1 & $\bar{F}$ & $\overline{SP}$
$\overline{SP}$ & T_2 & F & A
$\bar{F}$ OR $\bar{A}$
1
2
3
4
5
ST & $\overline{SP}$ & $\bar{A}$ & $\bar{F}$
$\overline{SP}$ & T2 & ($\bar{F}$ OR $\bar{A}$)
F & $\overline{SP}$
ST & $\overline{SP}$ & (A OR F)

Inputs: Start PB (ST), Stop PB (SP), Flame present (F), Reset PB (R), Timer 1 complete (T_1), Timer 2 complete (T_2), Air flow SW (A)

Outputs:

State	Description	Air	Pilot valve	Ignition	Gas valve	Start timer 1	Start timer 2	Alarm bell
1	Off	0	0	0	0	0	0	0
2	Air purge	1	0	0	0	1	0	0
3	Ignition	1	1	1	0	0	1	0
4	On	1	1	0	1	0	0	0
5	Alarm	1	0	0	0	0	0	1

Figure 14.57 State diagram and output table for control of a gas burner

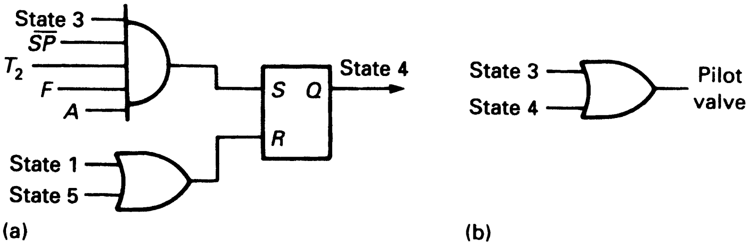


Figure 14.58 Implementation of a state diagram: (a) one of the five states of the gas burner control. Each state is represented by a flip flop and is set by transitions to the state and reset by transitions from the state; (b) one of the seven outputs. Each is simply an OR function of the states in which it is energised. The pilot valve is energised in states 3 and 4.

is easy to design, understand and modify and simplifies fault finding for maintenance personnel.

State diagrams are being formalised by the International Electrotechnical Commission (IEC) and the British Standards Institute (BSI), and already exist with the French Standard Grafset. These are basically identical to the approach outlined above, but introduce the idea of parallel routes which can be operated at the same time. *Figure 14.59(a)* is called a *divergence*; state 0 can lead to state 1 for condition 's' OR to state 2 for condition 't' with transitions 's' and 't' mutually exclusive. This is the form of the state diagrams described so far.

Figure 14.59(b) is a *simultaneous divergence*, where state 0 will lead to state 1 AND state 2 simultaneously for transition 'u'. States 1 and 2 can now run further sequences in parallel.

Figure 14.59(c) again corresponds to the state diagrams described earlier, and is known as a *convergence*. The sequence can go from state 5 to state 7 if transition 'v' is true OR from state 6 to state 7 if transition 'w' is true.

Figure 14.59(d) is called a *simultaneous convergence* (note again the double horizontal line) state 7 will be entered if the left-hand branch is in state 5 AND the right-hand branch is in state 6 AND transition 'x' is true.

The state diagram is so powerful that most medium size PLCs include it in their programming language in one form or another. Telemecanique give it the name Grafcet (with a 'c'), others use the name Sequential Function Chart (SFC) (Allen Bradley) or Function Block (Siemens). The IEC have adopted state diagrams as one of their formalised methods of PLC programming in IEC 1131.

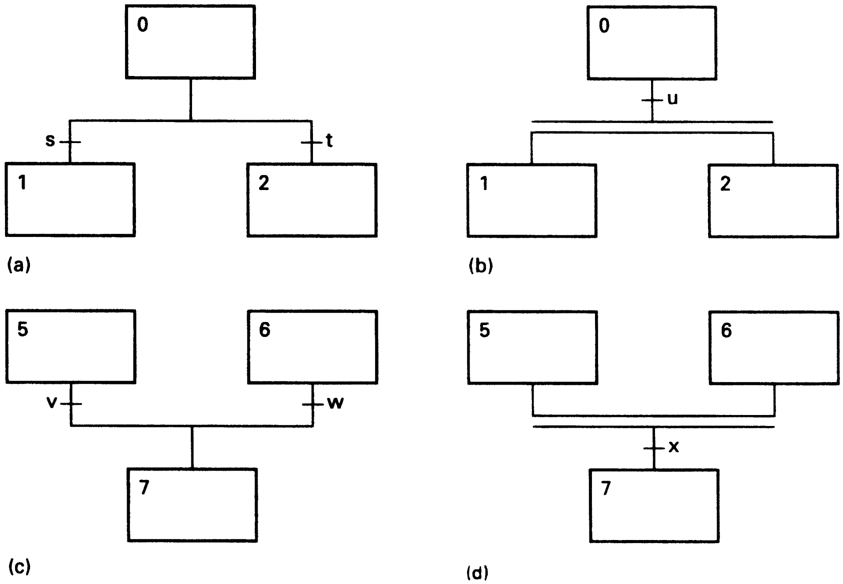


Figure 14.59 State transition diagram symbols: (a) divergence; (b) simultaneous divergence; (c) convergence; (d) simultaneous convergence

14.9 Analog interfacing

14.9.1 Digital to analog conversion (DAC)

A binary number can represent an analog voltage. An 8 bit number, for example, represents a decimal number from 0 to 255 (or −128 to +127 if two's complement representation is used). An 8 bit number could therefore represent a voltage from 0 to 2.55 V, say, with a resolution of 10 mV. A device which converted a digital number to an analog voltage is called a *digital to analog converter*, or DAC.

Common DAC circuits are shown on *Figure 14.60*, in each case the output voltage is related to the binary pattern on the switches. In practice, FETs are used for the switches, and usually an IC DAC is used. The R-2R ladder circuit is particularly well suited to IC construction.

14.9.2 Analog to digital converters (ADCs)

There are several circuits which convert an analog voltage to its binary equivalent. The two commonest are the *ramp ADC* and the *successive approximation ADC*. Both of these compare the output voltage from a DAC with the input voltage.

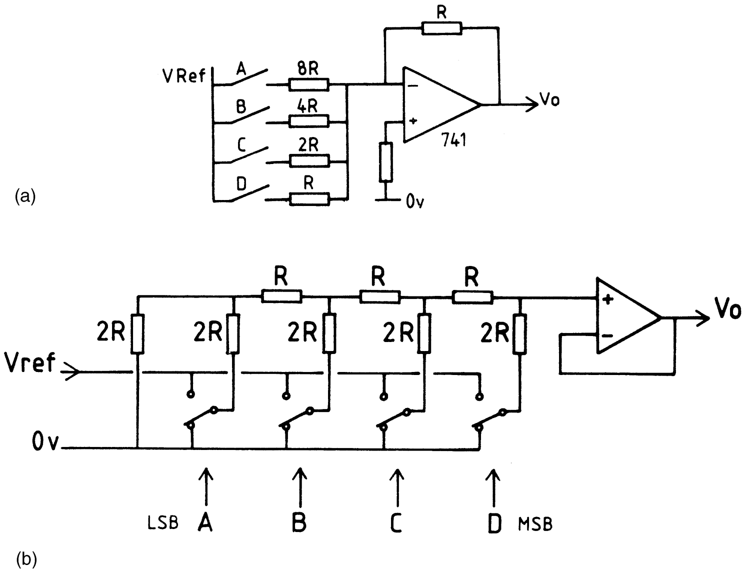


Figure 14.60 Digital to analog converters: (a) weighted resistors with OpAmp adder; (b) R-2R ladder

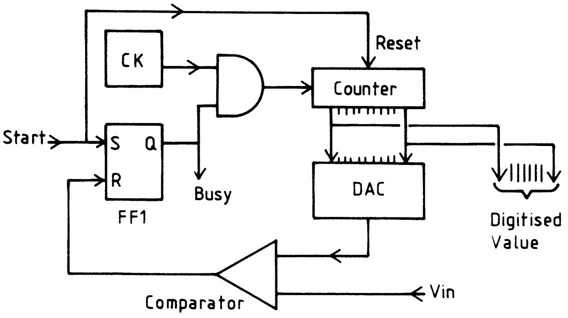


Figure 14.61 Ramp ADC block diagram

The operation of the ramp ADC, shown on *Figure 14.61*, commences with a start command which sets FF1 and resets the counter to zero. FF1 gates pulses to which counts up. The counter output is connected to a DAC whose output ramps up as the counter counts up. The DAC output is compared with the input voltage, and when the two are equal FF1 is reset, blocking further pulses and indicating the conversion is complete. The binary number in the counter now represents the input voltage. A variation of the ramp ADC, known as a *tracking ADC* uses an up/down counter that continuously follows the input voltage.

The ramp ADC is simple and cheap, but relatively slow (typical conversion time >1 mS). Where high speed, or high accuracy is required a successive approximation ADC is used. The circuit, shown on *Figure 14.62* uses an ordered trial and error process. The sequence, shown on *Figure 14.63*, starts with the register cleared. The MSB is set, and the comparator output examined. If the comparator shows the DAC output is less than, or equal to, V_{in}, the bit is left set. If the DAC output is greater than V_{in}, the bit is reset. Each bit is similarly tested, in order from MSB to LSB, causing the DAC output to quickly home in on V_{in} as shown. In total the number of comparisons is equal to the number of bits, so the conversion is much faster than the ramp ADC.

Successive approximation ADCs are fast (conversion times of a few μS) and accurate (0.01% is easily achievable). Unlike the ramp ADC, the conversion time is constant. They are, however, more complex and expensive than the simpler ramp ADC.

The *flash converter* is the fastest ADC available, but is not widely used for high accuracy applications because the circuit complexity increases rapidly with the number of bits. Commercial eight bit flash encoders such as the MC10135 are to be found in digital television and digital audio applications. *Figure 14.64* shows a simple three bit converter with a resolution of one part in eight.

The input signal is compared simultaneously with seven equally spaced voltages, for our simple example these are 1, 2, 3 V etc. If, for example, the input signal is 3.6 V, comparators a, b and c will all give a '1' output, and comparators d to g will give a '0' output.

The outputs from the seven comparators are converted to a three bit binary output by an encoder. This is simple combinational logic, output *C*, for example, being given by

$$C = \overleftarrow{d} + \overleftarrow{e} + \overleftarrow{f} + \overleftarrow{g}$$

The complexity of the combinational logic goes up considerably with the number of bits and the degree of internal checking required.

The flash converter is very fast with conversion times of a few nanoseconds, the only constraint being the propagation delays through the comparators and the encoder logic. It is, however, prone to giving invalid transitory states if the input is varying, going, say, 011 to 111 to 100 for a input change from three to four volts. For this reason a flash converter is usually preceded by a sample and hold circuit to freeze the analog input circuit whilst the measurement is being made.

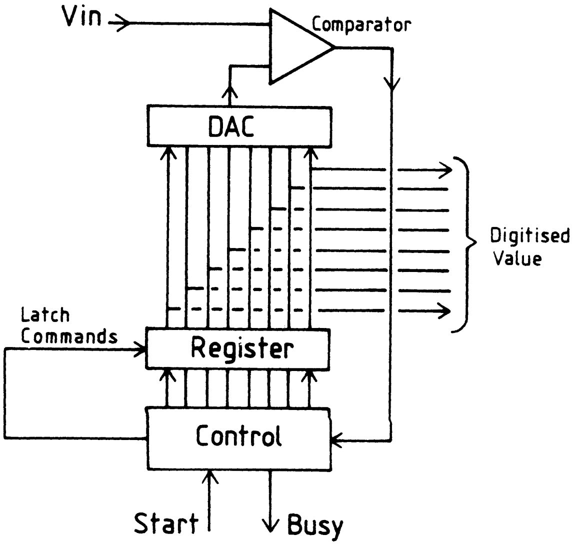


Figure 14.62 Successive approximation ADC block diagram

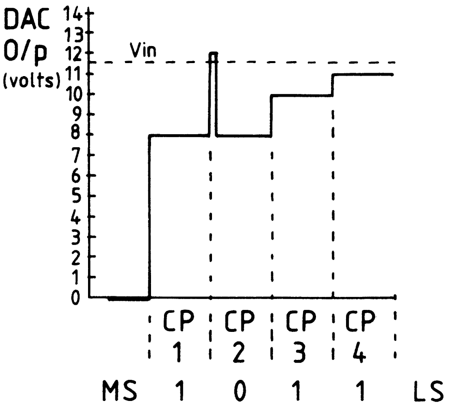


Figure 14.63 Operation of a successive approximation ADC

14.10 Practical considerations

Real life digital systems have to connect to the outside world, and this can often bring problems when noise and effects such as contact bounce are encountered. Precautions also need to be taken against inadvertent introduction of high voltages into logic systems via inter-cable faults on the plant.

All signals between a logic system and the outside world should use a technique called *opto isolation* when cable lengths are longer than a few metres. *Figure 14.65* shows

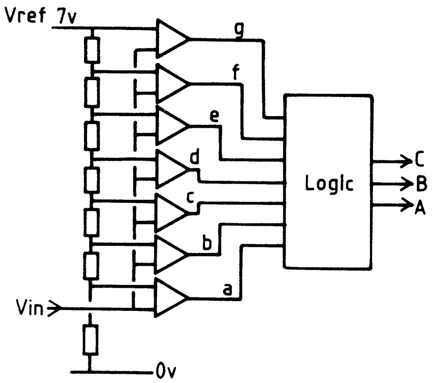


Figure 14.64 Three bit flash ADC

typical input and output circuits. In both, the signal is electrically isolated by using a coupled LED and phototransistor. Because the plant side power supply and digital power supply are totally separate, the system will withstand voltages of up to 1 kV without damage to the digital equipment (although such voltages would probably damage the plant side components of course). The absence of ground loops and relatively high current levels (around 20 mA) also gives excellent noise immunity.

Opto isolation devices (such as the TIL 107) are usually constructed in a 6 pin IC, and are characterised by a *current transfer ratio*. This is defined as the ratio between the photo transistor collector current to the LED current. A typical value is 0.3, so 20 mA input current will give 6 mA, output current. If Darlington phototransistors are used, transfer ratios as high as 1.2 can be obtained.

Noise can also enter digital systems via the power supply rails so excellent filtering is necessary, both the d.c. side and (with LC filter) on the a.c. supply side. It is particularly important to adopt a sensible segregation of 0 V rails such that digital logic, relays/lamps and analog circuits have separate 0 V returns to some common earth points. Under no circumstances should high currents flow along logic 0 V lines, or the logic 0 V be taken outside its own cubicle.

Digital ICs can also generate their own noise on power supplies (TTL is particularly troublesome). It is therefore highly desirable to provide each IC with its own local 0.01 μF capacitor. A single large value electrolytic has no effect as the noise is caused by rapid d*i*/d*t* and the PCB track inductance.

Mechanical contacts from switches, relays etc, do not make instantly but 'bounce' rapidly for 1 to 4 mS due to dirt and the uneven constant surfaces. In many purely combinational logic systems this does not matter, but where counting, sequencing or arithmetic circuits are used, trouble can ensue.

Contact bounce can be removed by RC filters, but the best solution is to use a bounce removing flip flop as *Figure 14.66*. Provided break before make contacts are used, the circuit gives totally bounce free true and complement

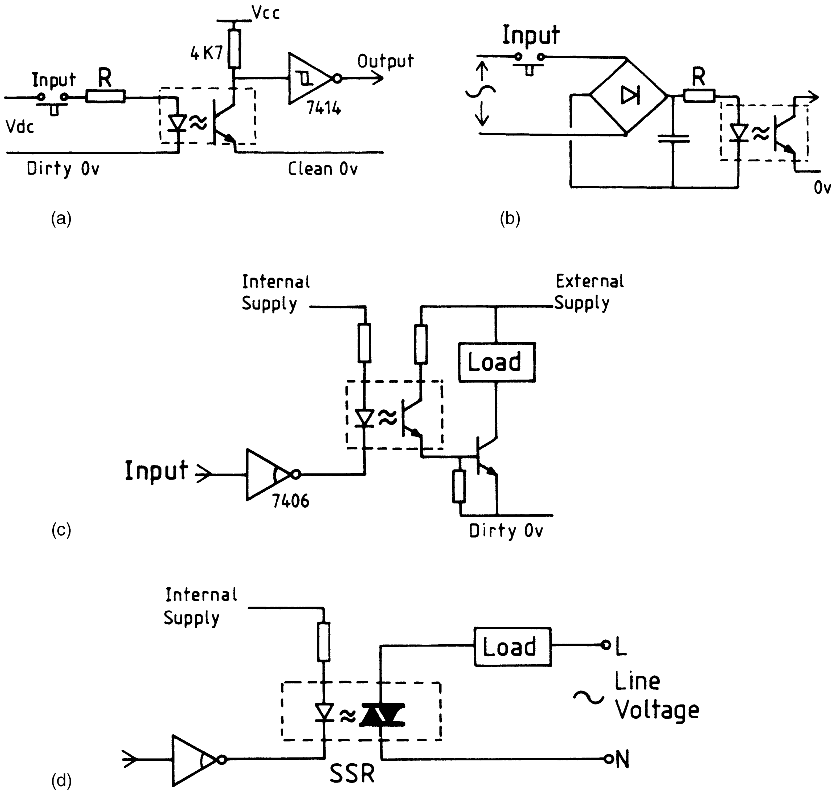


Figure 14.65 Optical isolation between digital system and outside world: (a) d.c. input circuit; (b) a.c. input circuit; (c) d.c. output circuit; (d) a.c. output circuit

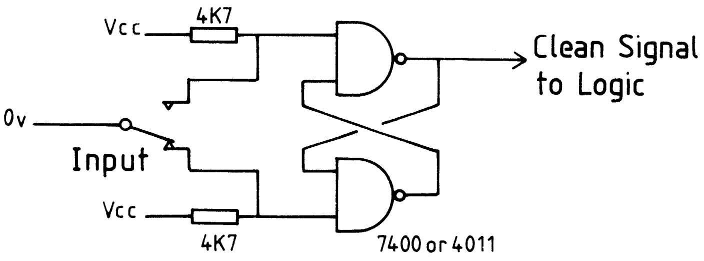


Figure 14.66 Bounce removing flip-flop

outputs. If the contacts are some distance from the digital system opto isolation should, of course, be used before the flip flop.

14.11 Data sheet notations

The following abbreviations are commonly (but by no means universally) used on logic data sheets:

A, B, C, D...	Data inputs. Where a number is implied, A = 1, B = 4, C = 8 etc.
a, b, c, d, e, f, g	Seven segments display signals
BCD	Binary coded decimal
BI	Blanking input
C	Capacitor for timer or monostable
Cin, Cout	Carry in, Carry out
CD	Count down input (on up/down counter)
CEP	Count enable parallel input
CER	Count enable ripple input
CK	Clock (often with >ς symbol)
CS	Chip select
CU	Count up input (on up/down counter)
CY	Carry out
D	Data input on D type flip flop
DEC	Decrement input (on up/down counter)
DIS	Disable tri-state output
EN	Enable
GND	Ground or 0 V
I/O	Input/output (often on bidirectional buffer)
INC	Increment input (on up/down counters)
INH	Inhibit
J, K	Inputs on JK flip flop
LE	Latch enable
LT	Lamp test
MR	Master reset
OEN	Output enable (for tri-state gate or buffer)
OF, OV	Overflow
PE	Parallel enable
PH	Phase input for liquid crystal display drivers
P/S	Parallel/Serial selection
Qn	Output with weighting (e.g. Qb)
R	Reset or Resistor for timer or monostable
RBI	Ripple blanking input
RBO	Ripple blanking output
RC	Resistor/capacitor for timer or monostable
RCO	Ripple carry out
S	Set input or sum output
SDL	Serial input data to shift register shift left
SDR	Serial input data to shift register shift right
Si	Serial input
SQ (or Qs)	Serial output
SR	Synchronous reset
ST	Strobe
T	Trigger
TC	Terminal count output
U/D	Up/Down mode control for counter
Vcc	Positive supply
Vdd	Positive supply
Vss	Usually 0 V
Vee	Negative supply
WE	Write enable
X	Data inputs for data selector
⊓	Schmitt trigger action
Σ	Sum output
>ς	Clock input

15

Microprocessors

D J Holding BSc(Eng), PhD, CEng, FIEE, MBCS, MIEEE
Aston University

Contents

15.1 Introduction 15/3

15.2 Structured design of programmable logic systems 15/3
  15.2.1 Design for test 15/5

15.3 Microprogrammable systems 15/6

15.4 Programmable systems 15/8
  15.4.1 The logic design of a digital computer system 15/8
  15.4.2 Processor architecture 15/9
  15.4.3 Central processing unit 15/9
  15.5.4 Control and timing unit 15/10
  15.4.5 Arithmetic logic unit 15/11
  15.4.6 Memory unit 15/11
  15.4.7 Interrupts 15/13
  15.4.8 Input/output 15/13
  15.4.9 Microprocessors 15/14

15.5 Processor instruction sets 15/14
  15.5.1 Types of instruction 15/14
  15.5.2 Data objects and data types 15/15
  15.5.3 Instruction formats 15/16
  15.5.4 Addressing data objects 15/16
  15.5.5 Addressing program code 15/17

15.6 Program structures 15/17
  15.6.1 Selection 15/17
  15.6.2 Repetition 15/18

15.7 Reduced instruction set computers (RISC) 15/19
  15.7.1 The reduced instruction set concept 15/19
  15.7.2 The reduced instruction set (RISC) processor 15/19
  15.7.3 Instruction pipelines 15/20

15.8 Software design 15/21
  15.8.1 Program development 15/21
  15.8.2 Assembly languages 15/21
  15.8.3 High-level languages 15/22
  15.8.4 Real-time processes 15/23
  15.8.5 Embedded real-time operating systems (RTOS) 15/23

15.9 Embedded systems 15/23
  15.9.1 Embedded processors 15/23
  15.9.2 System on chip (SoC) design 15/24

15.1 Introduction

Digital systems are used to process discrete elements of information. They are built from digital electronic circuits that process discrete electrical signals using simple logic and arithmetic operations. A digital electronic system can also be used to hold or store discrete elements of information and this gives the system a memory capability. The ability to store information or data and to process the data by logical or arithmetic operations is central to the design of nearly all digital information-processing systems including digital computers. The function of a digital system is determined by the sequence of operations that are performed on the information or data being processed. A digital system can be classified by the way in which its sequence of operations is implemented.

A digital system is considered to be hardwired if the sequence of operations is governed by the physical interconnection of the digital processing elements. For example, in hardwired logic systems the physical interconnections of the elements govern the routes by which data flows between the processing elements and thus the sequence of processing operations performed on the data. Conventionally, a hardwired system is considered to be inflexible because the design is specific to a particular processing function: if the processing function is changed, then the processing elements and their interconnections have to be altered.

The flexibility of hardwired systems has been much improved by the introduction of programmable (i.e. configurable) logic devices such as Programmable Logic Arrays (PLAs) and Field Programmable Gate Arrays (FPGAs) that can be programmed or configured to implement an application-specific digital signal processing function. Flexibility has been improved further with the introduction of re-programmable (i.e. re-configurable) devices that can be reprogrammed easily during fast-prototype system development, and can be reprogrammed after a product has been deployed to provide enhanced features or performance. Progress in this area has been rapid and the latest generation of re-programmable FPGA device can be configured with a wide variety of communication interfaces. This opens the possibility of using advanced communication technology, such as the Internet, to re-program or re-configure a remote hardwired system.

A digital system is considered to be genuinely programmable if a prescriptive program of instructions (i.e. software) can be used to control the data-processing function of the system. This type of system usually incorporates a general-purpose processing element which is programmed to implement a specific function in a predetermined way. The coded instructions are normally stored in the memory part of the system and the program forms an integral part of the system. The ability to define the function of the digital system by programming introduces considerable flexibility into the system because the programming operation can take place after the general-purpose digital elements have been designed. It also means that identical hardware designs can be used in a number of different applications, the system being tailored to the individual tasks by the applications program. A wide range of simple fixed-function programmable systems, such as sequencers and microprogrammed controllers, are used as controllers in embedded electronic systems. In this type of application the sequence of instructions is usually held in read-only memory (i.e. firmware) which increases the robustness of the system.

The digital computer is a very important class of stored program system. The computer or microprocessor is distinguished by the fact that its processing function depends on both the prescriptive sequence of coded instructions and the value of the data being processed. In effect, the program prescribes a number of possible sequences of operations and the conditions under which they may be carried out. The computer, under program control, assesses the data and determines which specific sequence of instruction is to be executed. It is the ability of the computer to take into account the value of the data being processed, when taking decisions about the type of processing to be performed, which makes the computer such a significant and powerful information-processing device.

All three forms of digital electronic system find widespread application. Traditionally hardwired logic has been used extensively to provide the control and interface logic for more complex digital components such as microprocessors and other very large scale integration (VLSI) devices. It is also used in the design of high-speed signal processing circuits for FPGA implementation. Increasingly, hardwired logic is used to provide the interface circuits between the main functional components within a complex FPGA. Where flexibility is required, it is common to use reconfigurable systems particularly in more complex applications.

Programmable systems are used in an extremely large range of applications. The simpler fixed-function programmable systems are often used in repetitive tasks such as input scanning and data acquisition. They are also used in mass-produced products and as components of larger systems such as telephony equipment. However, the continually increasing computational power of the microprocessor and its derivatives, such as digital signal processing (DSP) devices or powerful reduced instruction set (RISC) processors, has led to many of these applications being designed using fully programmable digital systems. In addition, commercial off-the-shelf (COTS) microprocessors are commonly used in both stand-alone and embedded systems. Such systems are providing economic solutions to design problems in an increasingly wide range of application.

The increase in size of VLSI logic circuits has led to a new generation of reconfigurable devices that are large enough to contain a complete digital processing system within a chip, called System on Chip (SoC) devices. An SoC device can be configured to include an embedded digital RISC processor, memory, communication interfaces, clock management, application-specific digital signal processing (hardwired logic functions), and appropriate internal interfaces and data buses. This allows the designer to partition a design into those parts that will be implemented as software executing on the embedded processor and those parts that will be implemented in hardware as high-speed application-specific logic circuits. This design approach, known as co-design or co-ware, has the significant advantage that established and high-performance parts of the design can be committed to application-specific hardware, and more adventurous parts of the design or low-speed functions can be committed to easily changed software. This minimises risk, facilitates time-to-market which gives competitive advantage, and provides a good path to post-deployment upgrades of the system's capabilities and performance.

15.2 Structured design of programmable logic systems

The design of an application-specific digital system typically involves the so-called 'top-down' approach and starts from a specification which includes a statement of the problem and the identification of the principal functional parts of the system. This can be elaborated as an architectural specification which identifies the major components of the data or signal processing system and a control specification which describes an algorithm or procedure for the functional control of the processing system.

Traditionally a systems-level design approach is adopted and the design is developed through a structured process of elaboration and refinement. During this process, the data- or signal-processing specification is translated into a set of digital signal processing modules or circuits. Similarly, the control algorithm is depicted as a finite state machine (FSM) and is translated into a sequential logic circuit that generates the sequence of control signals which coordinate and synchronise the signal processing modules. In large designs the process of design refinement through analysis and decomposition can be applied repeatedly to form a hierarchy of functional descriptions. The process of decomposition is conventionally terminated when the granularity of the description matches that of commonly used digital electronic building blocks such as arithmetic circuits and memory elements, or sets of logic gates. However, modern programmable devices, such as complex programmable logic devices (CPLDs) and FPGAs, have complex internal structures that are purpose designed for the efficient implementation of large functional units such as multipliers or ALUs. Therefore, it is often counter-productive to elaborate a design down to gate level without taking into account the logic structure of the target device.

Figure 15.1 shows the general organisation of a system designed using such an approach. It comprises external inputs and outputs, the controlled circuit which performs the data or signal processing, the controller which governs its behaviour, and internal signal paths which transfer condition or status information from the controlled circuit to the controller and control signals generated by the controller to the controlled circuit. The FSM controller is a simple sequential circuit that comprises: a state register which stores the current value of the state variables, combinational logic for generating the next value of the state variables, a clock signal which synchronises the transition from the current state to the next state, and combinational logic for generating the value of the outputs (which are either a function of the current state or a function of the inputs and the current state).

Structured design techniques are well suited to computer-aided design (CAD) or electronic design automation (EDA) procedures. In particular, the hierarchical decomposition techniques used during the design phase have a one-to-one correspondence with the hierarchical CAD techniques used in traditional schematic diagram-based approaches to the capture, simulation, layout and routing, implementation, test and validation of complex circuit designs. Increasingly CAD tools provide high level specification capture facilities, such as graphical state machine (FSM) editors, to help capture design features in a tangible and user friendly manner.

The trend is to write the system specification using either a formal notation, or a programming language such as C or concurrent extensions of C, or a hardware description language (HDL) such as VHDL or Verilog. These notations provide constructs that facilitate the description of complex logic systems or algorithms and an underlying mathematical structure that can be used to reason about the behaviour of the systems. The use of such abstract or high-level notations has been found to facilitate design by allowing the designer to focus attention on the functional aspects of the design without the need to bind the design to a particular implementation technology. This is supported by modern CAD tools that allow high-level behavioural specifications to be simulated to verify the function of the system (i.e. using symbolic simulation) before the high level description is compiled (i.e. synthesised) into a logic circuit. It is conventional to use the high-level description language to describe both the design (or unit-under-test) and a test-bench (test sequence generator and response analyser). Thus the highest level in the design hierarchy comprises both the design and a test-bench.

Design synthesis CAD tools are commonly used to translate high-level behavioural digital systems specifications into logic circuits. The synthesis process is not easy, and modern synthesis tools typically use artificial intelligence techniques and employ deep knowledge of the architecture of the FPGA in order to synthesise sensible, efficient and fast logic circuits. Once synthesised, the design can be incorporated into the conventional logic design process of post-synthesis simulation, routing, implementation, test and validation.

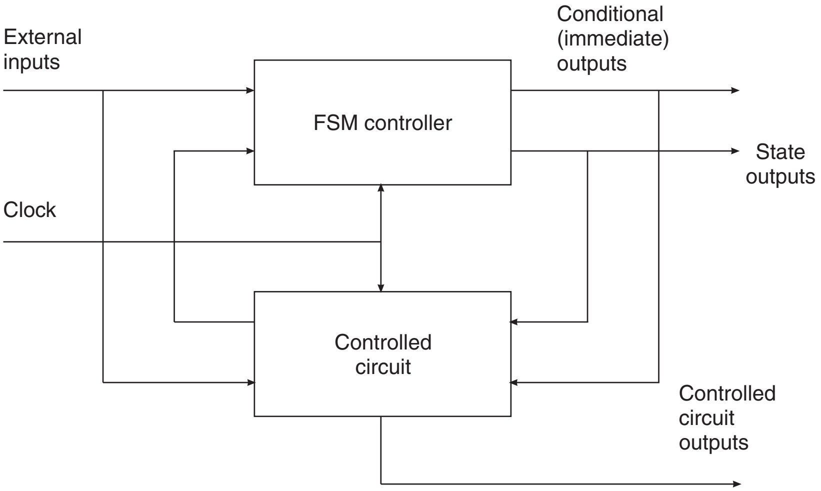


Figure 15.1 General organisation of a structured digital design

15.2.1 Design for test

It is normally a principal design requirement that the circuit should be testable. In most structured design methods the derivation of the functional specification goes hand-in-hand with the specification of tests to verify that the designed circuit functions as intended. The problem of testing a circuit is made more difficult when it is implemented as an ASIC because the limited number of pins on the integrated circuit restricts access to test points in the circuit. In particular, the constraint that test inputs must be applied via the external input pins limits the controllability of internal parts of the circuit under test. Similarly, the constraint that the response of the circuit must be observed using the output pins limits the ability to observe the state of internal parts of the circuit under test. For testability, it is necessary to ensure that the accessible or primary inputs can drive each node of the circuit (the property of controllability) and that each node can be observed from the accessible or primary outputs (the property of observability).

Combinational logic is tested by applying a set of test patterns to the inputs of the circuit, measuring the circuit's response at its outputs, and comparing its response with its predefined fault-free function. In order to test a testable circuit, it is necessary to generate a set of inputs (test vectors) which can be applied to the primary inputs and drive each node of the circuit. Observability problems may arise if redundant logic to added to a circuit to provide hazard cover.

The problem of testing sequential logic is considerably more complex because the state of a sequential circuit is a function of both the current inputs and the previous state of the circuit. To reduce the problem of testing such circuits, it is desirable to open the feedback paths (which are essential to sequential behaviour) and thus change the problem to one of testing the constituent next-state and state-output combinational logic. This requires the introduction of additional gates to inhibit the feedback paths, to allow the assertion of test states, to ensure the direct control of the clock, and to allow the observation of the next-state variables. This approach tests the combinational components and memory elements but does not provide a full-speed test of the actual sequential circuit and additional tests are required to ensure that the circuit is free of race hazards.

In the case of structured designs, a primary concern is the test of the controller which coordinates and synchronises the data or signal processing modules. In a typical FSM, the function of the controller is clearly defined by the control algorithm and the finite state machine controller is relatively easily tested once the feedback loops of the sequential part of the circuit are opened. Furthermore, the state register can easily be reconfigured in the test mode to form a shift register for the entry of test data and the capture of test results. This is shown in *Figure 15.2* for a controller with two state variables; the test mode select signal TMS causes the reconfiguration of the D-type state register to form a test vector shift register (shown in bold). This technique, which is known as scan path testing, can be applied

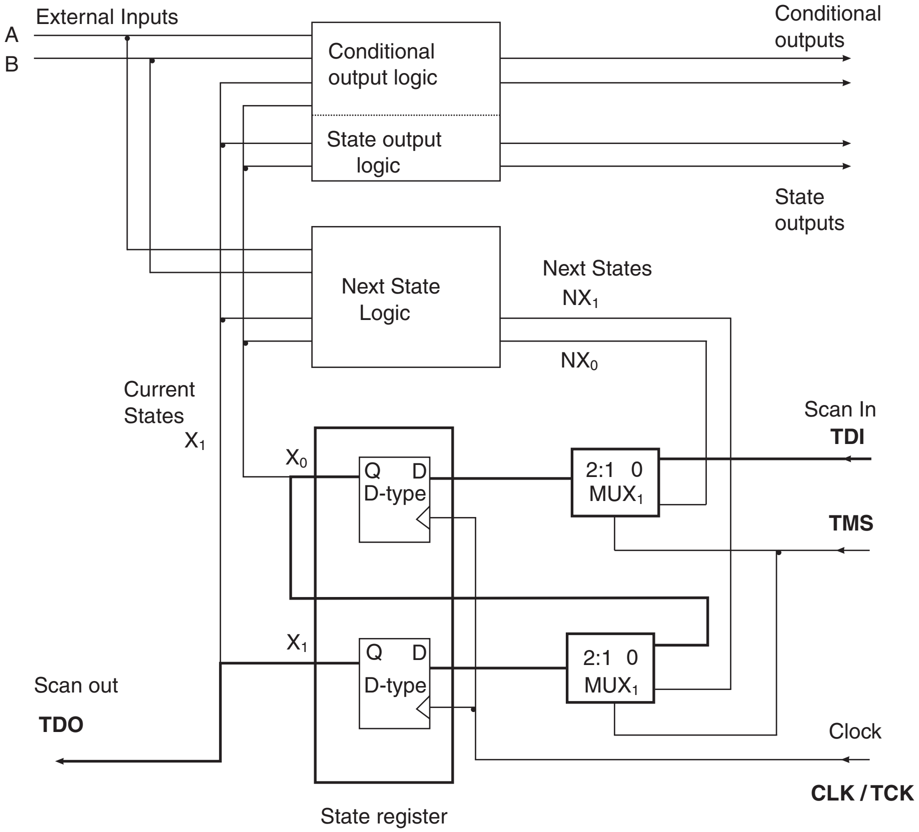


Figure 15.2 Scan-path design for an FSM controller

to a wide range of circuit and there are many variations of the theme, such as random scan and level sensitive scan techniques.

In general, significant increases in controllability and observability can be obtained if serial input and output techniques are used to load test data into an ASIC and remove capture response data from the ASIC. This can be achieved by incorporating a general-purpose shift register with serial-load/parallel-out, and parallel-load/serial-out facilities within the ASIC. In a typical test, the ASIC circuit would be put into test mode and the test stimuli or vector would be serial loaded into the shift register. The ASIC would then be switched into normal mode and the test data present at the parallel output of the shift register would be allowed to excite the circuit. When the response had stabilised, observation data would be latched into the shift register using the parallel-load facility. The ASIC would then be put into test mode and the response would be output using the serial-output facility of the shift register. Clearly, the additional test circuits must be built into the ASIC as part of the design. This method requires up to five pins on the integrated circuit to be dedicated for test purposes: test mode select (TMS), serial test data in (TDI), serial test data out (TDO), direct control of the relevant clock circuits (TCK), and a reset (TRST). The use of such signals is consistent with the JTAG/IEEE standard 1149.1 which provides a framework for test at board and chip test level, primarily using the Boundary Scan technique that is commonly provided as a built-in component of modern VLSI and ASIC devices.

15.3 Microprogrammable systems

A typical digital system's design can be decomposed into a set of signal or data processing elements and a set of Finite State Machines (FSMs) which coordinate and control the signal- or data-flow and processing. To do this, the FSM controllers monitor any necessary signals (such as inputs or status signals) and generate the control signals necessary to coordinate, synchronise and control the signal- or data-processing elements. It follows that the FSMs form a crucial part of such designs, and a variety of structured design methods have been devised for capturing and synthesising the FSM controllers. These include traditional algorithmic state machine or ASM design methods, and state-transition diagram methods that allow direct synthesis of the design from Mealy or Moore diagrams. In addition, many HDL synthesis tools include program analysis tools that are designed to detect FSM-like structures and specifically synthesise FSM components. A key point in all these design methods is the assignment of a unique binary coding to each state of the system and the design of a bespoke finite state machine to generate the required state sequences and state outputs.

The need for bespoke hardware can be removed by designing a general-purpose controller (or finite state machine) which can be 'programmed' to produce the necessary control functions. To accommodate such an approach, a simple FSM can be reduced to a networks of states, state outputs, and single-qualifier decisions by replacing any conditional or immediate outputs by state outputs and by inserting additional states where necessary to ensure that only one qualifier is associated with each decision. A unique state identifier is then assigned to each state according to a normal binary count (as far as possible) so that the state-machine design can then be implemented using a counter, with appropriate controls, instead of with a state register and combinational next state logic. The modified FSM can then be reduced to a set of 'instructions' by identifying commonly occurring structures and their associated counter control logic as follows:

(1) *Sequence of states*: increment counter unconditional (IUC).
(2) *Decision*: increment or branch conditional (IBC).
(3) *Wait until condition*: hold or increment conditional (HIC).
(4) *Branch unconditional*: branch unconditional (BUC).
(5) *Loop forever*: hold unconditional (HUC).

The counter is normally controlled using two control lines, 'counter enable' (CE) and 'counter load' (LD). On the next clock, the counter is incremented if CE is asserted or a branch address is loaded into the counter if LD is asserted. Thus, the counter control logic necessary for each construct or primitive instruction defined above can be readily determined. The use of mnemonics (such as IUC, HIC etc.) to represent commonly occurring structures allows the FSM to be replaced by a list of primitive symbolic 'instructions'. Each instruction will define the present state identifier or location count, the mnemonic describing the control operation to be performed on the counter, the identity of any qualifier, the name or value of any branch address, and the name or value of the state outputs. A typical instruction format is shown in *Figure 15.3*.

A suitable processing architecture for the above primitive instructions is shown in *Figure 15.4*. The state register is implemented with a controlled counter and, as only one instruction is needed per state, the input and output logic is efficiently implemented in ROM. The relevant input (or qualifier) for each state is chosen by addressing a multiplexer. Before the contents of the ROM (or RAM) can be defined, each instruction must be assigned a binary code or 'opcode' and each input must be assigned a MUX address. The instruction decoder is needed to translate the instructions into suitable control signals for the counter. In the case of conditional instructions, the counter control signals depend on both the instruction type and the qualifier or 'flag'.

Thus, each instruction stored in ROM comprises an opcode, the binary MUX address for the input qualifier, the binary branch-location address, and the binary values of the state outputs. This form of instruction is known as a *microinstruction*. The function of the controller can be changed by simply altering the microinstructions, and this process is known as *microprogramming*. Microprogramming is tedious and error prone and software development tools such as assembly language generators are often used to allow programming using symbolic notations.

In practice, a number of proprietary microprogrammable controllers have been developed. They are often equipped with a primitive stack to allow a limited procedure or subroutine facility. This requires additional instructions such as 'call procedure' or 'branch to procedure' and 'return from procedure' and mechanisms to increment the current counter (or ROM address) and save the incremented address in the stack, to load the procedure start address or value into

Location	Opcode	MUX address	Branch address	Outputs

Figure 15.3 Microinstruction format

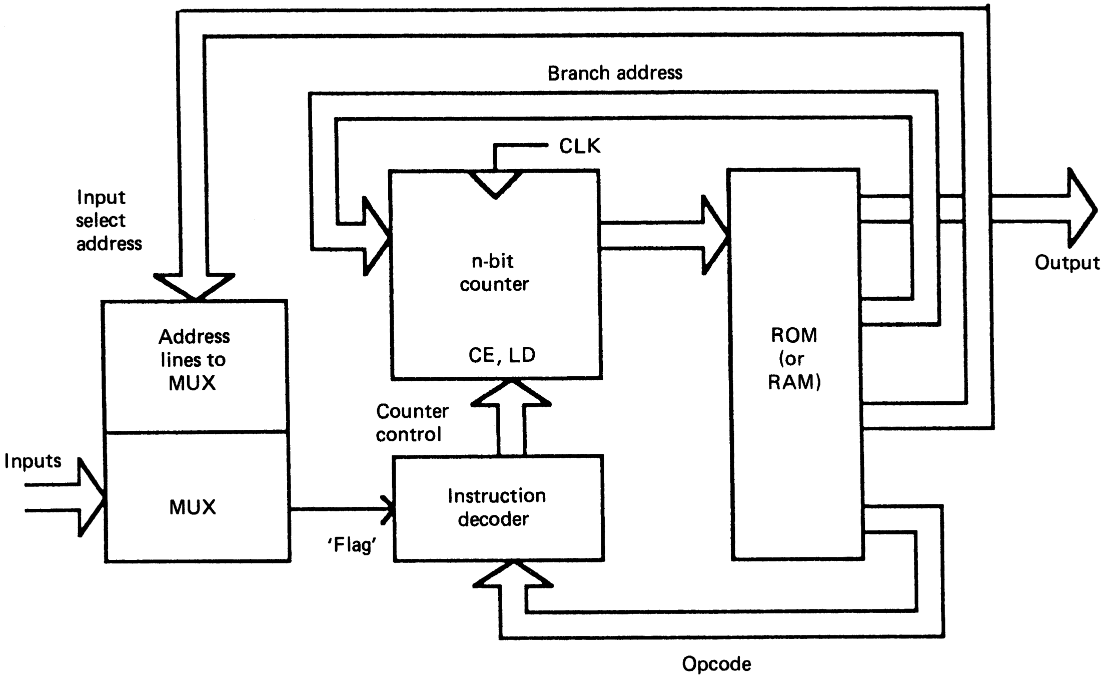

Branch address
CLK
Input
select
address
n-bit
counter
CE, LD
ROM
(or
RAM)
Outputs
Address
lines to
MUX
Counter
control
Inputs
MUX
Instruction
decoder
'Flag'
Opcode

Figure 15.4 Simple microprogrammable architecture

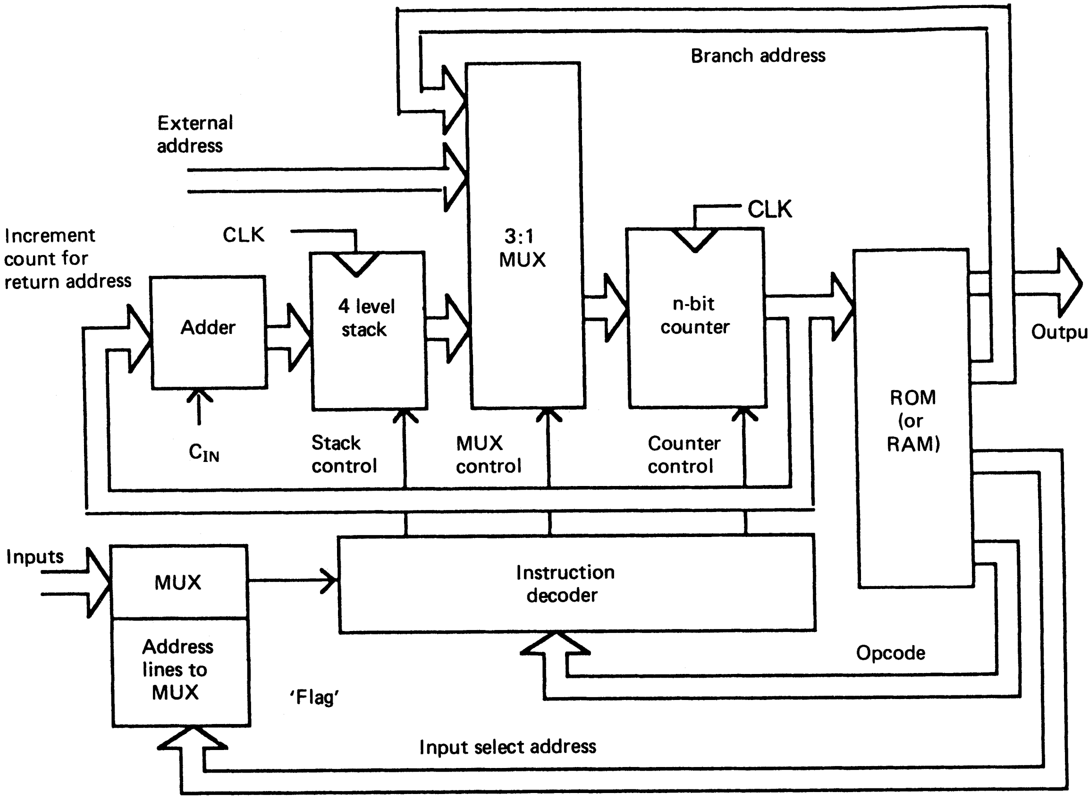

Branch address
External
address
Increment
count for
return address
CLK
3:1
MUX
CLK
Adder
4 level
stack
n-bit
counter
ROM
(or
RAM)
Outputs
C_{IN}
Stack
control
MUX
control
Counter
control
Inputs
MUX
Instruction
decoder
Address
lines to
MUX
'Flag'
Opcode
Input select address

Figure 15.5 Microprogrammable controller

the counter, and to retrieve the address from the stack and load it into the counter. An external load facility for the counter may also be provided so that the controller can be used in conjunction with another processor. This gives the architecture shown in *Figure 15.5*. Such processors often have regularised instruction sets and well-developed tools for software development including simulation and emulation facilities.

Since a microprogrammed controller is a relatively simple circuit; designers have often taken the opportunity to incorporate a data processing capability into the design. Significantly, such microprogrammed 'processors' often incorporate a specialised, very high performance, arithmetic processor. Typical examples include high performance floating-point arithmetic processors, and digital signal processing devices (which include a high-speed multiplier and accumulator for implementing the repetitive add-and-multiply operations found in digital filtering algorithms). Microprogrammed controllers are readily available as stand alone devices, or as pre-prepared HDL scripts, known as 'intellectual property (IP) cores', that can be embedding within an HDL design and synthesised for an FPGA target.

15.4 Programmable systems

A programmable system, such as a microprocessor or computer, comprises a general-purpose processing unit which processes data or digital signals. The processing operations performed are specified by a computer program, which consists of a set of logical instructions stored in the computer memory.

A programmable system will comprise digital electronic circuits to:

(1) *input* signals or data from external circuits or systems;
(2) *move* or transfer the data within the system;
(3) *store* the data before, during and after processing;
(4) *process* the data by logic, arithmetic or bit manipulation operations;
(5) *output* the processed data to external circuits or systems.

Each of these operations can be performed by an appropriate configuration of combinational and sequential logic circuits including memory elements. In practice, programmable systems comprise a general-purpose logic design which can be configured to perform a wide range of operations. The hardware is controlled by a program or sequence of instruction codes which define the operations necessary to implement a particular processing function. The instruction codes are normally stored in the memory part of the system and the function of the system can be changed simply by altering the stored program. This type of system can be considered to be composed of two parts: the hardware which is basically independent of the application and the software which defines the application function. Since the hardware part of such a system is invariant, it can be produced economically as a standard design or device.

There are basically two types of general-purpose programmable system. The fixed function programmable machine is a limited form of programmable system which is constrained to perform a prescribed and fixed sequence of instructions. This type of system does not have the capability under software control to select between two alternative sequences of instruction. A fixed function programmable machine is therefore forced to execute a fixed sequence of instructions in all circumstances and is properly regarded as a programmed machine rather than a computer. The application function of such a system can be altered only by reprogramming the system. These systems can be used in any applications in which the processing function does not depend on the nature or value of the data being processed.

The digital computer is the most powerful and flexible form of programmable system. A program of instructions which are executed in sequence again defines its function, but in the case of the computer the program may specify alternative sequences of instructions and the conditions under which they can be executed. For example, conditional expressions and alternative sequences are implicit in high-level programming constructs such as 'if condition-true then...', 'repeat... until condition-true', and 'while condition-true do...'. Special hardware is required to carry out this type of operation and a computer is equipped with a logic circuit, known as a status register, which is used to store information about the status of the processor after it has executed an instruction. When a program is executed, each conditional expression is evaluated (using the actual values of the data variables) and the resulting values of the bits or flags in the status register are used to identify which alternative sequence of instructions is to be executed next. Thus, the order of processing may be modified according to the result of the instruction. It is this mechanism which allows a program to take into account the nature of the information being processed.

In effect, the computer can be programmed to take a decision about the future courses of action that it may take, based on the actual value of the data being processed. It is this facility which characterises a proper computing system and which makes the computer such a powerful data processing system. Such systems are providing economic solutions to digital signal and data processing problems in an increasingly wide range of applications.

15.4.1 The logic design of a digital computer system

A digital computer system can be considered to be composed of two logic structures. The first is associated with the flow, storage and processing of data and consists of the data input and output subsystems, the data highways used to move or transfer data within the system, the memory which is used to store the data, and the arithmetic logic unit which is used to process the data. The function of the processor is prescribed by a program of instructions held in memory. The second structure is responsible for controlling the fetching of instructions from memory and for ensuring that they are executed in the correct sequence. It also governs the detailed execution of each instruction and controls the data-flow and data-processing elements so that they perform the required processing operation.

The control structure consists of the memory which is used to store the program code and mechanisms for identifying the location of the next instruction, for fetching and decoding the instruction, and for identifying the location of any inputs or stored data which form the operands in a processing operation. It controls all aspects of the execution of the instruction including fetching operand data and, when necessary, taking into account the status of the previous programming operation. It also identifies the destination location of any resultand data generated by the processing operation and stores the resultand or generates an output.

The data-processing and program-control structures are heavily interconnected and often share common hardware.

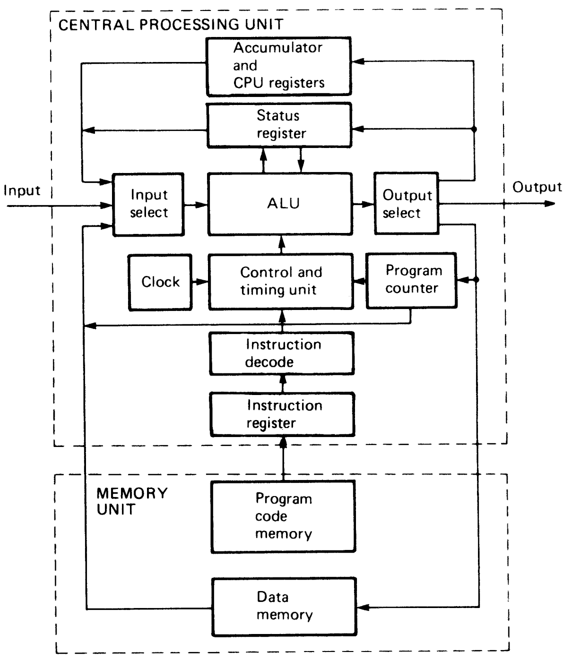

CENTRAL PROCESSING UNIT
Accumulator and CPU registers
Status register
Input
Input select
ALU
Output select
Output
Clock
Control and timing unit
Program counter
Instruction decode
Instruction register
MEMORY UNIT
Program code memory
Data memory

Figure 15.6 Processor architecture

In particular, the program code and the data being processed are usually stored in an identical binary format. It is usual for code and data to be stored in separate areas or segments within a common memory unit.

15.4.2 Processor architecture

Computers and microprocessors are general-purpose programmable systems which perform sequential processing operations. Classically, they are constructed using general-purpose functional units such as a central processing unit or CPU, a memory unit, and an input/output subsystem, as shown in *Figure 15.6*.

The CPU is the heart of the computer, it contains the control and timing unit (CTU) which controls the programmed operation of the system and the arithmetic logic unit (ALU) which processes the data. An external clock provides the timing pulses or reference signals required by the CTU. The CPU also contains a number of important registers such as the *program-counter* which points to the next instruction, the *instruction register* which holds the current instruction, and the *status register* or *flag register* which stores the status information about the result of the previous instruction executed by the CPU.

The memory unit provides storage for program code and data. The code and data are always considered to be separate entities although they may share physical memory. (Some processor designs enforce the conceptual separation of code and data by providing separate memories for code and data.) The memory is arranged into words which consist of several binary digits, typically 8, 16 or 32 bits. Each word can be individually addressed and operated on by the computer.

The input/output subsystem of a computer provides the interface to external circuits or systems. Data may be passed in and out of the computer via serial or parallel interfaces. The input/output system is used to input program code, to input data for processing, and to output results. It also provides the means of communication with the operator via a man-machine or human-computer interface (MMI or HCI). The processing power of the computer can only be used if the input/output subsystems allow efficient communication between the user or application and the processing system.

Many computer systems are configured around one or more general-purpose data highways. They typically consists of a common bus structure of address, data and control lines and are used to communicate to all devices external to the CPU including memory, input/output systems and backing stores. The simple bus-orientated architecture illustrated in *Figure 15.7* provides ease of access to the control, address and data highways which are used to interface any logical system to the CPU. Bus-oriented architectures are used in many designs to provide a flexible and easily expanded computer system.

15.4.3 Central processing unit

The CPU contains the CTU which coordinates, synchronises and controls the fetching of instructions from memory and the execution of the instructions. The execution of an instruction will typically involve fetching operand data from memory, processing the data in the ALU, and storing the result in memory. The CPU also contains at least the minimum set of internal registers necessary for the execution of a program. These include the program-counter

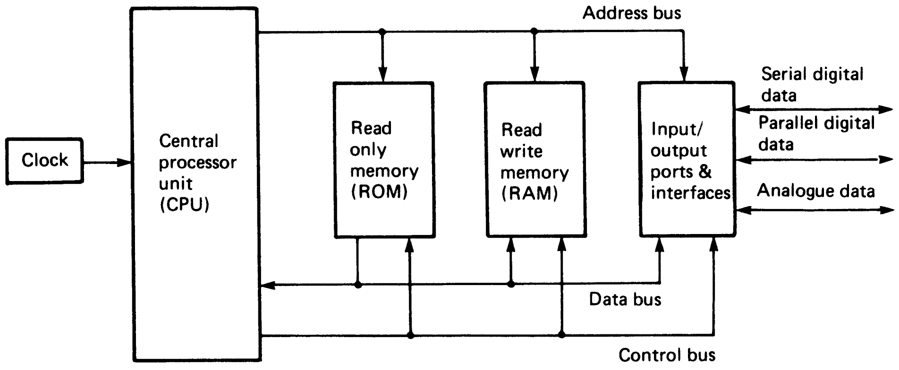

Address bus
Clock
Central processor unit (CPU)
Read only memory (ROM)
Read write memory (RAM)
Input/ output ports & interfaces
Serial digital data
Parallel digital data
Analogue data
Data bus
Control bus

Figure 15.7 Bus-orientated system architecture

register, the instruction register, the data memory reference registers, and the CPU data registers.

15.4.3.1 Program-counter register

This holds the address of the memory location containing the next instruction to be executed. It is the programmer's responsibility to initialise the program-counter correctly, so that it points to the first instruction in the program. During program execution, the program-counter is automatically incremented by the CTU to point to the next instruction. In this way the CPU is forced to execute instructions in strict sequence. However, some instructions are provided which modify the contents of the program-counter. For example, unconditional jump, branch, or 'go-to' instructions simply overwrite or modify the contents of the program-counter to effect a branch to another instruction. Similarly, conditional jump or branch instructions modify the contents of the program-counter if a particular condition is satisfied. This feature allows the program sequence to be modified if a specified condition is detected in the data being processed e.g. if the result of the previous instruction was negative.

15.4.3.2 Instruction register

This register holds the current instruction so that it can be decoded and input to the control and timing unit. Specifically, the instruction register holds the opcode which defines the type of instruction. Depending on the type of instruction, it may also hold immediate operand data or the addresses of operands and the address of the resultand. Since operand data and addresses comprise many bits, they are commonly held in temporary registers which can be considered as extensions to the instruction register. The contents of the instruction register can not be overwritten by the ALU, nor can they be accessed by a programmer.

15.4.3.3 Status register

This comprises a number of discrete status bits or flags and holds status data about the result of the previous instruction executed by the ALU. The data are used when computing 'decisions', such as selecting one of two possible future courses of action. Processors actually compute decisions of this type in two stages. In the first stage the ALU computes the condition which governs the decision; the Boolean result (yes/no or true/false) is held in the status register. In the second stage, this result is used as a qualifier in a conditional operation, such as a conditional branch or jump, such that the value of the qualifier is used to choose the appropriate future sequence of instructions. It is the responsibility of the programmer (or compiler writer) to ensure the correct and consistent use of the status register throughout the two-stage process of computing a decision.

15.4.3.4 Data memory reference pointers

These registers hold the addresses of the operands and resultand and are loaded during the instruction 'fetch'. They are used to access data objects during the execution phase of the instruction cycle and should be capable of accommodating the various addressing modes associated with the complex data types used in high-level programming languages. Therefore, it is perhaps more accurate to think of a data memory reference pointer as a mechanism which generates the address of a data object. The number of address registers available in a CPU is an important feature of computer architecture. Ideally, separate pointers or mechanisms are required for each operand and the resultand.

15.4.3.5 CPU data registers

These hold operands and resultand data during the execution of a program. Again, microprocessors differ, particularly in the internal storage provided in the form of CPU data registers. However, most processors have an accumulator register into which the ALU will automatically load the resultand of a processing operation. Many modern microprocessors have a larger number of CPU data registers which can operate as accumulators in complex arithmetic and logic operations.

Those parts of a processor which are of direct interest to a programmer are shown in the *programmer's model* which describes only those registers within the CPU which can be accessed by the programmer. *Figure 15.8* shows the programmer's model of a typical microprocessor system.

15.5.4 Control and timing unit

The basic operation of a computer or microprocessor is governed by the control and timing unit (CTU) which generates the signals necessary to coordinate, synchronise and control the movement and processing of all information within the system. A simple external clock usually drives the unit, and this provides a time-reference signal from which the CTU generates the timing and control signals for the various logic subsystems in the computer. Modern high-performance processors may include a separate clock management subsystem which generates multi-phase timing sequences for use by the CTU.

The control and timing unit is responsible for controlling the main operational cycle of the processor which is known as the 'instruction cycle'. The instruction cycle can be split

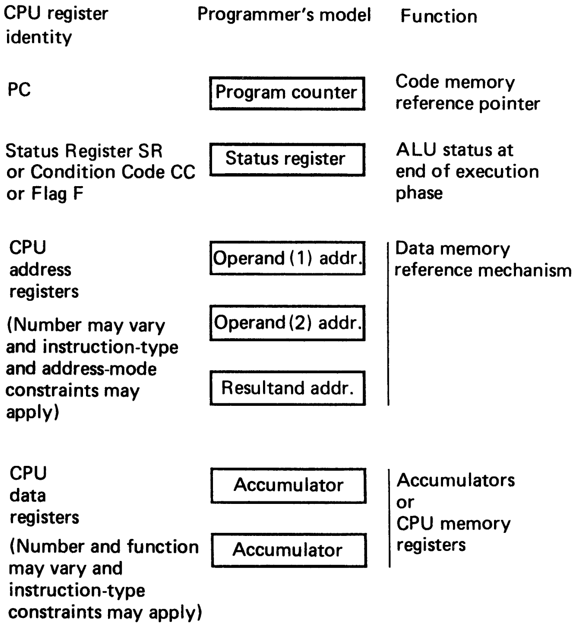

CPU register identity
Programmer's model
Function
PC
Program counter
Code memory reference pointer
Status Register SR or Condition Code CC or Flag F
Status register
ALU status at end of execution phase
CPU address registers
(Number may vary and instruction-type and address-mode constraints may apply)
Operand (1) addr.
Operand (2) addr.
Resultand addr.
Data memory reference mechanism
CPU data registers
(Number and function may vary and instruction-type constraints may apply)
Accumulator
Accumulator
Accumulators or CPU memory registers

Figure 15.8 Programmer's model

into two distinct phases, the instruction fetch and the execution of the instruction. During the instruction fetch the address of the next instruction is obtained from the program-counter mechanism and transferred to the memory address register (MAR). A memory reference operation is then performed on the code part or code segment of memory to read the opcode which is the first part of an instruction. The opcode data are transferred via the memory buffer register (MBR) to the instruction register where it is decoded and then input to the CTU. The program-counter is then updated to point to the next part of the instruction or to the next instruction.

The opcode identifies any further memory reference operations which are required to complete the instruction 'fetch'. The control unit uses the updated program counter to make reference to successive addresses in the code part of memory to fetch any further parts of the instruction, such as immediate data values or the addresses of the operands and the address of the resultand. This information is transferred to various temporary registers in the CPU for use during the 'execute' cycle. At the end of the instruction 'fetch', the CPU will contain all the information it requires to control the execution of the instruction and the program-counter will be pointing to the next instruction to be fetched (assuming that the execution cycle does not compute a new program-counter address). The various logic units used during the instruction 'fetch' cycle are shown in *Figure 15.9* in which the memory and input/output discriminator $\overline{M}$/IO is used to distinguish between memory reference operations and any operations involving peripheral systems which may use the same address and data bus.

The opcode also defines the sequence of operations necessary to execute the instruction. During the execution part of the instruction cycle the control and timing unit will synchronise the transfer of data within the system and control the operation of the ALU. The control unit will access operand data by transferring the operand addresses from the temporary registers to the memory address register to perform memory reference operations. In practice, many processors have a complex data reference pointer which will compute the address of the data object using not only the temporary register but also base or segment registers, offset registers, and index registers according to the addressing mode specified in the instruction. If the computer has a memory-to-memory architecture, then operand data can be transferred direct from immediate access memory to the arithmetic logic unit and resultands can be returned direct to storage in immediate access memory. However, if the computer has a register-to-register architecture, then the operand data is normally transferred to a CPU register before being processed by the arithmetic logic unit and resultand data is held in the accumulator or transferred to another CPU register. The register-to-register architecture has distinct performance advantages, particularly when used with a multiple-instruction pipeline CTU, as in modern reduced instruction set (RISC) processors.

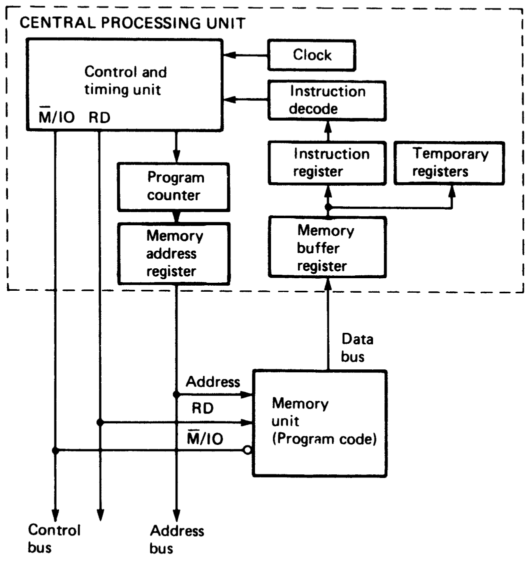


Figure 15.9 Instruction 'fetch' logic structure

15.4.5 Arithmetic logic unit

The actual data processing operations are performed by the ALU, which is a general-purpose logic system and can normally perform logical, arithmetic and bit manipulation operations. The ALU operates under the control of the control and timing unit and its function is defined by the current instruction held in the instruction register. The ALU can perform both monadic (single operand) and dyadic (two operand) operations and, therefore, has two input data paths. It generates status information in the status register and has an output data path for the resultand. Depending on the architecture of the processor, the operand data inputs may be from either immediate access memory registers or CPU data registers. The resultand is usually output to a special register, known as the accumulator, which is normally a multi-function register which can participate fully in the processing operations. In some systems the accumulator is used to store one of the operands before a processing operation and is subsequently used to store the resultand. This technique removes the need to have two operand registers and may increase the operational speed of the processor. However, the need to minimise the number of CPU registers is no longer a major design objective and many modern microprocessors have a number of CPU registers of advanced design which can act as operand registers or accumulators.

The ALU also contains the status register which is also known as a *flag register* or *condition-code register*. This register consists of a number of flip flops (flags) whose state reflects the result or state of the processing element at the end of the previous processing operation. This is illustrated in *Figure 15.10* which shows in schematic form the structure of a typical ALU and the other logical systems associated with the execution part of the instruction cycle.

15.4.6 Memory unit

The memory unit provides storage for program code and data. Computers commonly use two types of memory, fast immediate-access memory and backing store memory, which have different roles and functions.

The immediate-access memory is considered to be the primary memory unit of a computer, it is used to store program code and the data associated with the program so that it may be readily accessed during the execution of the program. Read only memory (ROM) devices may be used to store information which does not alter, such as program code or constant data, and random-access read-write memory (RAM) devices are used to store data which may be altered, such as the value of variables. The immediate

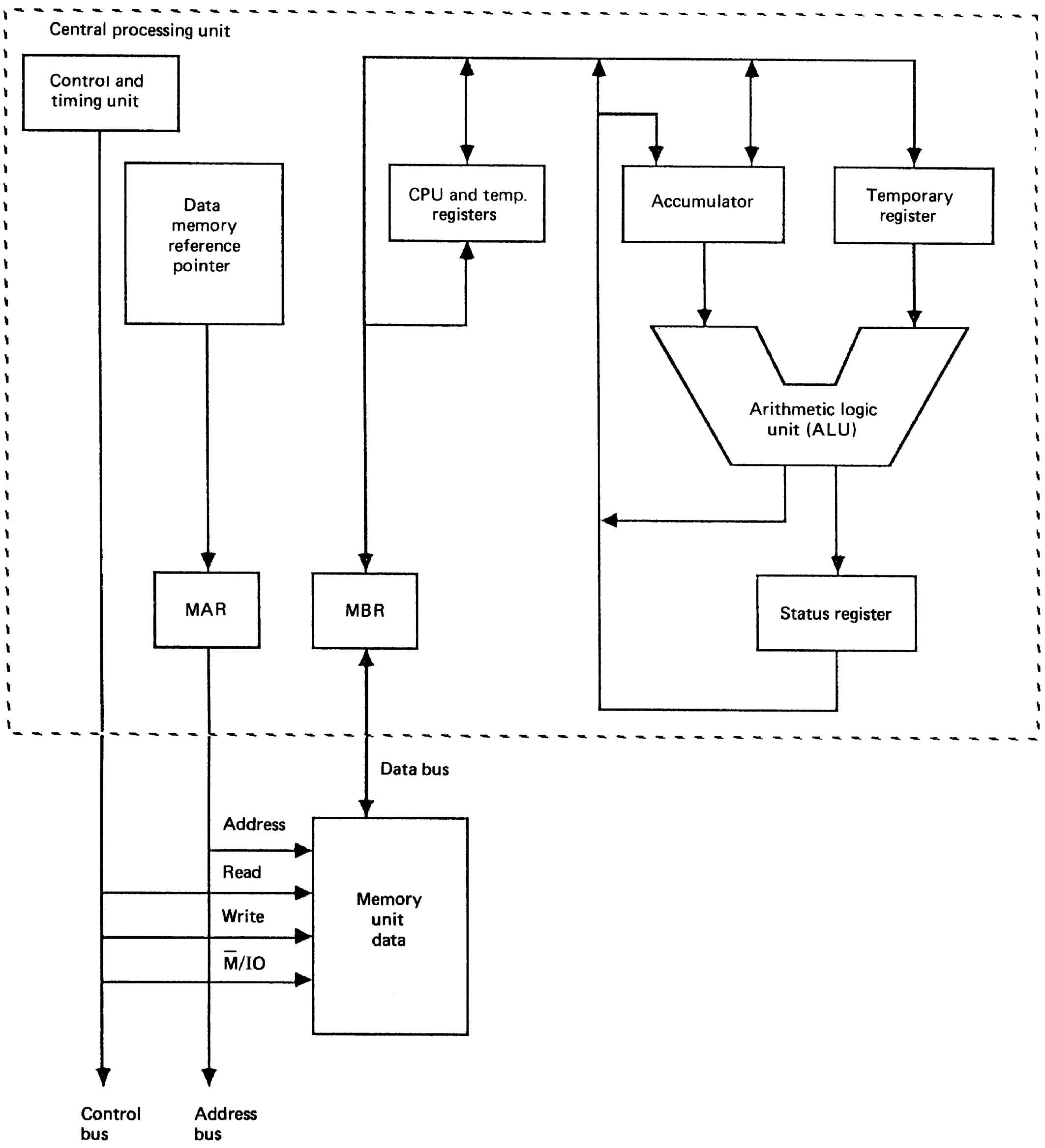

Central processing unit
Control and timing unit
Data memory reference pointer
CPU and temp. registers
Accumulator
Temporary register
Arithmetic logic unit (ALU)
MAR
MBR
Status register
Data bus
Address
Read
Write
M̄/IO
Memory unit data
Control bus
Address bus

Figure 15.10 Instruction 'execution' logic structure

access memory is normally constructed using semiconductor memory devices. Typically, large memories are built using dynamic RAM memory devices which provide 'high-density' storage at relatively low cost.

The immediate-access memory is normally limited in size by the computer architecture. For example, simple microprocessors may have a 16-bit memory address and this limits the size of the immediate-access memory to 64k registers or elements. While a 64 k × 8-bit memory is often sufficient for embedded applications such as dedicated control systems, larger memories are often required to support general-purpose applications software. Typically, modern microprocessors use 32-bit effective memory addressing and commonly have a 128 M byte or 256 M byte memory which is sufficient to run modern operating systems and applications software. However, the performance of memory-intensive applications, such as image processing, computer aided design (CAD), and interactive computer games, may benefit from larger memories. Therefore, more advanced microprocessors have the capability to physically address 1 G or more of memory.

Many computers have facilities for using an area of memory as a stack. This is a block of RAM memory which is used on a last-in/first-out (LIFO) basis for storing context information, such as the values of the program-counter, status register, and other CPU registers, or for storing data such as the parameters passed to subroutines. This facility is particularly useful for storing addresses and register contents during subroutine operations or during the context switches which take place following an interrupt.

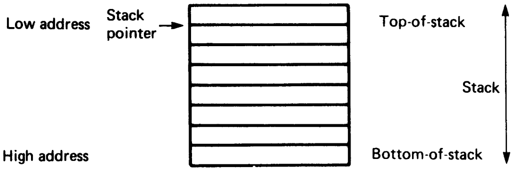


Figure 15.11 Stack layout in RAM memory

A stack is organised by a stack pointer, which is a CPU register holding the address of the last item placed on to the stack, also called the 'top-of-the-stack'. Instructions are provided which automatically decrement the stack pointer before data is stored (PUSH) or automatically increment the stack pointer after data is retrieved (POP) from the stack. A stack grows downwards in the store as shown in *Figure 15.11*.

Stacks can be used to implement nested subroutine calls by simply putting return addresses and CPU register contents on to the top of the stack during successive calls. Recursive subroutine calls can be implemented in a similar manner provided parameters are passed to the subroutine on the stack and local variables are stored on the stack at each level of call. Typically, the parameters are pushed onto the stack before the subroutine call; following the call the parameters are accessed by the subroutine using indirect addressing (using any CPU address register other than the stack pointer so that the stack pointer is available for the next recursive call). Results can be returned using the same technique. The stack can be employed in a similar fashion to store register and address information following an interrupt, including multiple and re-entrant interrupts.

Large immediate-access memories are relatively expensive to implement and are unsuited to the long-term storage of large program or data files. Most computers are therefore equipped with auxiliary or backing stores which are normally sequential access storage systems such as magnetic memory hard discs, and removable sequential stores such as optical storage discs, 650 M byte compact discs or CDs (such as write-once, read-only CD-R or read-write CD-RW discs), or 4.7 G byte digital video discs or DVDs (such as write-once, read-only DVD-R or read-write DVD-RW discs). These systems provide economic storage for the large volumes of data which are commonly used in database or image processing applications. Most computers are still equipped with a floppy disc drive, though this legacy technology is increasingly irrelevant for backing store purposes and is usually reserved for system start-up (or boot-up) during installation or fault recovery. Magnetic tape cartridges still provide economic storage for backing up large-scale systems such as servers.

15.4.7 Interrupts

In many applications the computer must respond rapidly when an external event occurs. This is usually achieved by an interrupt facility. The CPU is provided with a special input, the interrupt control line, which is used to notify the processor of the occurrence of an asynchronous external event. When the event occurs, e.g. a key is depressed or a switch is closed, the interrupt control line is driven to a specified logical state and the CPU is interrupted.

On detection of an interrupt, the processor carries out a sequence of operations that transfers control to a special form of subroutine, called the 'interrupt handler' or 'interrupt service routine', which is located at a pre-determined address in memory. The actions taken to invoke an interrupt service program vary from computer to computer, but in general terms the following sequence occurs:

(1) At the end of the current instruction the contents of the program counter and the status register are automatically stored in the stack and the interrupt line is disabled.
(2) The program counter is loaded with the address of an interrupt service routine, either directly or following interrogation of the interrupt source to determine the identity of the service routine so that the CPU can be vectored to one of a number of interrupt entry addresses appropriate to the particular interrupt.
(3) The interrupt service program is entered. Care should be taken to ensure that the interrupt service program does not alter the context of the interrupted program. Therefore, the CPU registers needed by the service routine are stored in the stack, this may be an automatic hardware facility or may be performed by the interrupt handler software.
(4) When the interrupt service program is complete, the context of the interrupted program is restored (by restoring the contents of registers saved in the stack) and control is returned to the interrupted program by restoring the contents of the status register and the program-counter. Also, interrupts are re-enabled if this has not already been done as part of the interrupt service above.

15.4.8 Input/output

The role of the input/output subsystem is to interface the computer to external logic devices. There are several ways of controlling input and output. Normally, data are input or output under program control at prescribed points in a program. In an event-driven environment, data can be input or output in response to an interrupt under the control of an interrupt service program. However, in both cases, the speed of data transfer is governed by the interface logic and by the speed of the input/output control program which executes in the CPU. In high-speed applications the restrictions due to the control program can be removed if the external logic circuits can access the immediate-access memory directly using a suitable access mechanism and input/output protocol.

15.4.8.1 Program controlled input/output

There are two commonly used methods for connecting input and output systems to a processor for program or interrupt controlled input/output. The most elegant technique treats all input and output ports as if they were memory registers in the memory unit. The input and output ports are connected to the address, data and control bus structures as if they were memory elements and are designed to operate to the same electrical and functional specification as a memory register. Data can then be output using a memory reference 'write' instruction at the output address, or input using a memory reference 'read' instruction at the input address. This method, which is known as *memory-mapped input/output*, is used in a wide range of processors. It gives fast input and output and is compatible with other software data-transfer instructions.

An alternative approach connects all inputs and outputs to a separate input/output bus structure which normally consists of a limited number of address lines and the usual control signals. In bus-orientated systems a subset of the memory address lines is used and an additional memory or input/output discriminator signal (M′/IO) is used to generate unambiguous addresses. Input/output-mapped input/output is not compatible with memory reference operations and special instructions such as IN or OUT are often used to distinguish this mode of operation.

Serial communications are usually interfaced using a universal asynchronous receiver and transmitter device (USART) which contains a serial-to-parallel receiver buffer, a parallel-to-serial transmitter buffer, a mode control register, and a status register which indicates valid communications. The data, control, and status register are accessed using either memory-mapped or input–output mapped techniques according to the architecture of the processor.

15.4.8.2 Interrupt-driven input/output

An interrupt can be used to force a processor to suspend its current task and execute an interrupt service program, as described in Section 15.4.7. Interrupt driven input/output is implemented by connecting the control logic of the external device to an interrupt line so that the device can demand the CPU's attention. Following the generation of an interrupt, the CPU is forced to respond immediately and execute a program which services the input or output requirements of the interrupting device. Interrupt-driven input/output maximises the utilisation of the external device, but causes suspension of the current task. Interrupt-driven input/output is commonly used to interface intermittent inputs such as keyboards. However, in some embedded applications it is undesirable to interrupt an executing task, and the preferred approach is to regularly inspect (or poll) an external device for the availability of an input.

15.4.8.3 Direct memory access

The use of direct memory access (DMA) allows an external device to transmit data directly into the computer memory without involving the CPU. The CPU is provided with control facilities which allow the DMA controller (external to the CPU) to gain control of the CPU data bus. The DMA controller must provide a memory address, the data, and bus control signals to effect a data transfer. The DMA controller then transfers data directly over the bus to or from the memory. DMA transfers are commonly used to send blocks of data, rather than individual items of data, between backing stores or peripheral devices and memory. The controller contains a counter to increment the memory address and count the number of transfers made within the data block. The DMA process is also referred to as *cycle-stealing*, since it proceeds simultaneously with program execution, the only effect being that the instruction execution time is increased by the number of memory cycles used when a transfer is in progress.

The relative merits of DMA over other means of input/output is that it is fast, uses the minimum amount of computer time per data word transferred and operates autonomously. The loss of instruction execution time is not usually significant unless a very large number of devices are under DMA control. The major disadvantage of DMA is that the computer program is not explicitly aware of changes in data or the completion of a DMA transfer and it is usually necessary to make the DMA controller invoke an interrupt to inform the CPU that a data block transfer is complete.

15.4.9 Microprocessors

Advances in microelectronics and computing science have provided the technologies necessary to construct the complete central processing unit of a computer on a single integrated circuit; this device was called a microprocessor. The microprocessor, which was developed in 1971, realised a step change in the cost, performance, power consumption and reliability of a minimum computer system. Further advances in VLSI design led to the development of integrated circuits containing both the CPU and the memory unit; the so-called single chip computer. In effect, these advances had resulted in the miniaturisation of the computer.

The microprocessor can also be viewed as an advanced programmable logic device. Special microprocessors and other advanced programmable systems have been developed to carry out specific computational functions. These processors are often designed to work in conjunction with a general CPU and are known as co-processors. A number of devices such as fast floating-point arithmetic units, communication or local area network co-processors, and multimedia units such as audio processors and graphics and video display generators, are available and can be used in the design of powerful processor architectures. To prevent such high-bandwidth processing elements making significant demands on immediate-access memory, they are often provided with separate application-specific memories as in the case of video display subsystems. A typical system architecture of this type is shown in *Figure 15.12* which illustrates the use of programmable systems including microprocessors in the design of an advanced information processing system.

15.5 Processor instruction sets

Most general-purpose computers or microprocessors are designed to execute sequences of instructions or more complex programs of instructions which prescribe the actions necessary to input, store, and process data, and output computed results. The instruction set of a processor defines the machine code operations, which the processor can perform. However, the range of instructions available with a particular processor depends to a considerable extent on the design objectives of the particular manufacturer. The range and capability of the instruction set provided may have a considerable influence on the choice of a computer for a particular application task.

15.5.1 Types of instruction

Although there is no standardisation of computer instructions, most processors provide primitive operations or instructions for the following.

15.5.1.1 Program flow control

The sequence in which instructions are executed is defined implicitly by the program which comprises an ordered list of instructions held in successive memory locations. Unconditional branch or jump instructions can be used to jump to a program in another part of memory. Also, repetitive or loop structures can be formed by jumping back to an instruction which has already been processed. However, the true power of a programmable system is provided by conditional instructions. The flags in the status register can be used as qualifiers for conditional branches, either on their own, as in 'branch if zero', or in combinations as in 'branch if greater than' or 'branch if less than or equal'.

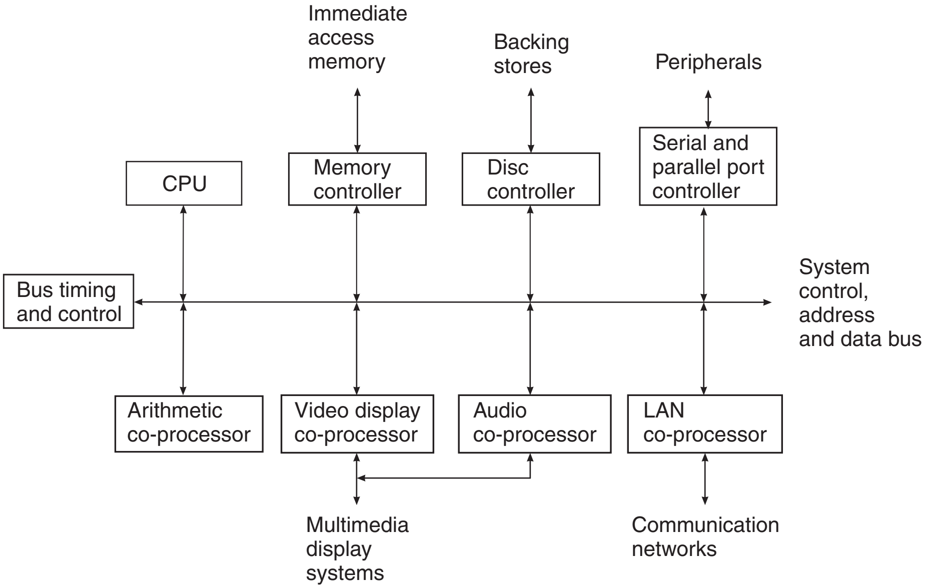


Figure 15.12 A multiprocessor system architecture

When computing decisions (selecting between alternative sequences of code) conditional branch or jump instructions are used to determine whether to continue the present sequence or jump to the start of an alternative sequence. Similarly, conditional jumps can be used to form conditional exits from repetitive structures.

Although most processors are unable to implement directly the flow control constructs found in high-level programming languages, the implicit sequence, unconditional branch and conditional branch instructions form the primitives from which constructs such as 'if-then-else', 'while-do', and 'repeat-until' can be formed.

15.5.1.2 Data-flow including input and output

Data-transfer instructions, such as MOVE, can be used to input external data to a CPU data register or an immediate-access memory register, to transfer data between such registers, and to output data from such registers to the outside world. Although these instructions do not necessarily make use of the ALU and may not alter the status register, they can be classed as data processing instructions in the sense that they assign values to the variables (registers).

15.5.1.3 Data-processing instructions involving the ALU

Most processors provide a range of arithmetic instructions including addition, subtraction, multiplication and division. These instructions are used in mathematical applications including the data 'sorting' operations used in data base applications. Simple processors often implement these operations using two's complement integer arithmetic. More complex algorithms involving floating point arithmetic can be programmed using these primitive operations. However, these programs often make intensive use of the processor and are relatively slow and it is common to enhance the performance of such processors by adding arithmetic co-processors. More advanced processors have powerful built-in arithmetic capabilities including floating point arithmetic units.

Most processors can also implement logic operations such as NOT, AND, OR, and EXCLUSIVE OR (XOR) which are implemented 'bit-wise' by performing the operation simultaneously on each corresponding pair of bits in the operands. These operations are used to perform the 'compare' or 'find' operations used in database applications. Many processors can also implement shift/rotate instructions which involve moving all the bits in a computer word either to the right or the left. There are several possible form of shift, such as arithmetic and logical shifts and logical rotations. (Few conventional high-level languages give direct access to primitives for physical bit-level manipulation.)

15.5.1.4 Machine control

These instructions control the mode of operation of the processor. Many machine control instructions, such as START/RESTART, HALT, STOP have a profound influence on the behaviour of the processor. Similarly, in event driven systems, machine control instructions such as INTERRUPT ENABLE/DISABLE and interrupt priority control instructions affect the ability of the processor to respond to external stimuli. Therefore, some processors classify certain machine control instructions as 'privileged instructions' which can only be used if the processor is in a special 'systems' or 'supervisor' mode that is used by systems programmers. (Few conventional high-level programming languages give direct access to machine-level primitives for machine or interrupt control.)

15.5.2 Data objects and data types

At machine level textual, numerical and logical information is represented by codes of binary digits and the processor in not able to infer the context of any particular binary data object. Thus, the concept of *data typing*, in the high-level sense, does not exist at machine level.

A processor handles such data in terms of the contents of registers. Thus low-level primitive instructions transfer and process data by reference to the architecture and registers of the processor, such as input/output device registers, CPU data registers and/or accumulators, and immediate-access memory registers. Data typing at machine level is restricted to specifying the length of a data object. Most low-level assembly languages provide assembler directives which allow the programmer to declare data objects by length, assign symbolic names (identifiers) to the objects, and provide initialising values for variables. When an assembly language source program is translated into machine code, the assembler enforces the data-type rules on the usage of the declared data objects and allocates storage space at machine level for all data objects.

Although limiting, low-level data types provide the building blocks for accommodating (storing and processing) the more complex data types normally associated with high-level programming languages. However, the efficient use of high-level data types also depends on the availability of suitable addressing mechanisms for accessing data objects.

15.5.3 Instruction formats

Each computer instruction is stored in memory as binary numbers and can be considered to comprise a number of fields:

(1) *Operation code (op-code)*: this part of the instruction identifies the type of operation which is to be performed (such as 'add' or 'jump'), the number and addressing mode of the operands, and the addressing mode of the resultand (if any).

(2) *Operand field*: this specifies either an immediate data value (if immediate addressing) or the address of the operand on which the instruction operation is to be performed. The processor's data memory reference mechanism will use the address information in conjunction with the addressing mode to compute the effective address (physical address) of the operand.

(3) *Resultand field*: this specifies the address of the resultand (corresponding to the addressing mode used). In some processors, the resultand address is, by default, the same as that of one of the operands, and when the instruction is executed the resultand overwrites the operand concerned.

The format of a typical instruction, such as ADD, for a memory-to-memory architecture processor in which the operands and resultands reside in immediate access memory is shown in *Figure 15.13*. This type of instruction format has the potential to generate multi-word instructions. For example, a 16-bit microprocessor may have a 16-bit op-code and either a 16, 20, 24 or 32 bit memory addressing capability. The resulting instruction would be long and the corresponding instruction fetch would require many memory reference operations, which is inefficient. Many CPU architectures force the resultand to overwrite one of the operands, this gives some gain in efficiency since the resultand address is, implicitly, the same as one of the operands. The format of a typical instruction of this type is shown in *Figure 15.14*.

In practice, many processors have a register-to-register architecture where the operands and resultands are stored in CPU data registers, which being few in number can be addressed using very short direct addresses. This gives some gain in efficiency, although separate MOVE instructions are required to load data from memory into the CPU registers (the LOAD operations) and to return results to memory (the STORE operation). The so-called LOAD-STORE architecture, or register-to-register architecture processor, is the default architecture for modern reduced instruction set (RISC) processors. These processors commonly have a CPU register file comprising 32 general-purpose 32-bits registers that can act as source registers for operands and accumulators for resultands.

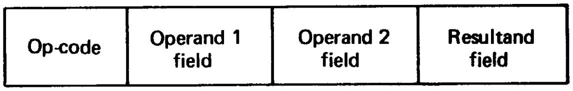


Figure 15.13 Instruction format—explicit resultand

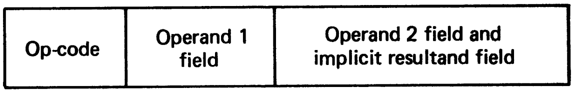


Figure 15.14 Instruction format—implicit resultand

15.5.4 Addressing data objects

During the execution of an instruction, operands are fetched from the addresses indicated in the operand field of the instruction, and resultands are returned to the address shown in the resultand field of the instruction. A number of different methods of addressing operands have been developed. These address modes are used to introduce flexibility by decoupling the logical address from the physical address, to extend the address range of the memory that can be accessed from an instruction, and to provide support at a primitive level for the addressing mechanisms required in advanced data structures. The data-memory reference system is used to generate the physical address of a data object from knowledge of the addressing mode and the values in the operand field of the instruction and any associated address registers.

The address modes commonly encountered for accessing operands are as follows.

15.5.4.1 Immediate addressing

In this mode, the actual value of the operand is included in the instruction, i.e. the operand address field is a literal. This allows rapid access to the operand, but the value of the operand is fixed by the program code. The operand may be a data object, such as an integer constant, or an address object. Typically, it is used to load small integer constants into a register. (To avoid repeatedly using immediate addressing to load the commonly used value zero into CPU registers, many RISC processors have one CPU 'register' permanently hardwired to the value zero.)

15.5.4.2 Direct addressing

In this mode the value in the operand address field is the address of the operand. There are two main variants of direct addressing. In *CPU register direct addressing* the address of the CPU register is given, as an explicit value, in the address field of the instruction. Since most CPUs have a small number of CPU registers, the address field is restricted and this allows single word instructions. CPU Register direct addressing is fast because the instruction is short and the operands are already held in the CPU. It is used extensively in register-to-register architecture processors, including RISC processors. In *memory direct addressing* the immediate access memory address of the operand is given, as an explicit value, in the address field of the instruction. In limited-word-length computers this means that only a small area of memory can be assessed directly. Also, this addressing mode is inflexible because the address is embedded in the program code. It is commonly used to access address constants, such as input/output ports.

15.5.4.3 Indirect addressing

In this mode, the operand field of the instruction identifies an address register (usually a CPU register) which holds the address of the operand. The address register must be initialised before use. In effect, the address-register acts as a 'pointer' to the operand. This removes the need for the instruction to specify the absolute address, which introduces flexibility. It also allows the operand to be accessed by a one-word instruction. If the address register has accumulator capabilities and can participate in arithmetic operations, then the indirect address or pointer can be manipulated to give access to complex data structures. For example, the indirect address could be incremented during successive passes through a loop of instructions.

Some processors have special mechanisms which allow an indirect register to be decremented or incremented immediately before use (e.g. pre-decrement) or immediately following use (e.g. post-increment). Such pointers can be used as stack pointers which are automatically updated to point to the 'top-of-stack' following a PUSH or POP operation.

15.5.4.4 Indexed addressing

Indexing is used to address sequential data structures. The effective address of a data object in the structure is formed from the sum of two components; the address of (the start of) the data structure and the index or offset of the object relative to the start of the structure. The data-memory reference mechanism computes the effective address as the instruction is executed.

Indexed addressing is normally implemented using two CPU registers. An address-register is used (as in indirect addressing) to point to the start of the data structure (the address of the first object the structure) and a second register, known as the *index register*, holds the offset address of the data object. The index register usually has an accumulator capability such that the index can be readily modified or incremented. Modern processors often have a number of address registers which can be used as pointers to data structures and a number of index registers. An indexed addressing instruction for such a processor would specify both the pointer and index register.

15.5.4.5 Base and (relative) offset addressing

In this system of addressing, an address register is used as a 'base pointer' and points to the segment of memory allocated to the data associated with a program. All references to data objects are made relative to the base address. The data memory reference mechanism automatically adds the base address of data segment when calculating the effective address of a data object. Thus, in this system, direct addresses, indirect address, and indexed addresses are assumed to be relative to the base address. Thus all relative addresses associated with the operand data can be calculated when the program is compiled or assembled and do not require further alteration when the program is located. This has the significant advantage that, as far as references to data objects are concerned, the data segment can be relocated easily because the executable program code remains unchanged and only the value of the base pointer has to be altered.

15.5.5 Addressing program code

The program-counter points to the location of the current instruction and is incremented as each instruction is executed in sequence. However, the value in the program-counter is overwritten during unconditional and conditional branch or jump instructions. In effect, the operand of branch and jump instructions may modify the value of the program counter. Branch and jump instructions employ a variety of addressing modes:

15.5.5.1 Direct addressing

The value in the operand field of the instruction is the destination address of the branch or jump.

15.5.5.2 Indirect addressing

The operand field of the instruction identifies an address-register which holds the destination address of the branch or jump.

15.5.5.3 Relative addressing

The value in the operand field is interpreted as a positive or negative binary number which is added to the current contents of the program-counter to determine the destination address of the branch or jump. This is usually quite efficient because most destination addresses will be fairly close to the instruction being executed. Since the offset of the relative address is independent of the location of the code, the offset can be determined as a constant when the program is assembled or compiled and does not need to be altered when the program is located. This has the significant advantage that, as far as references to code locations are concerned, the executable program code can be relocated and only the initial value of the program counter has to be changed.

15.6 Program structures

Certain program structures occur so frequently in program design that it is worth looking at methods for implementing them both at high-level and at assembly or machine level. The implementations use both unconditional and conditional jumps.

15.6.1 Selection

The '*if-then-else*' selective construct is used to select between two alternative instructions (or processes). This high-level construct specifies the alternative processes and the conditions under which they can be executed, for example:

```
if x > 0 then P1 else P2;
```

The decision part of this construct is implemented at machine level by two distinct instructions:

(1) the evaluation of the conditional expression, which must be a relational operation that returns a Boolean result. When this is computed, the result is reflected by the setting or resetting of one or more flags in the ALU status register.
(2) the conditional branch which uses the relevant flags in the status register as operands. When the conditional branch instruction is executed it passes program flow control to the selected process (i.e. if the condition is true then branch to P1 else continue with P2). Note that in the low level implementation an unconditional branch instruction has to be inserted at the end of the 'else process' P2 to allow both alternative processes to be stored in sequential memory, as shown in *Figure 15.15*.

15.6.2 Repetition

Consider a process which must be executed several times. If the number of iterations is known, then a 'for' loop would be indicated; otherwise the loop structure could be implemented using either as a 'while... do' or a 'repeat... until' construct. The 'repeat... until' construct should be used if the process is to be executed at least once, otherwise the 'while... do' construct, which allows the possibility of an exit before the process is executed, should be used. Since both the 'for' and 'repeat... until' constructs can be derived from the 'while... do' construct, the 'while... do' construct is the primitive and is found in all high-level sequential and concurrent programming languages.

15.6.2.1 'Repeat-until' construct

This repetitive construct allows a process to be executed at least once. The number of times the process is executed depends on a value of a control variable, for example:

```
count :=number;
repeat
  P1
until count <ς0 do ;
```

where the process P1 must update the loop control variable, as in:

```
count :=count - 1;
```

Implementation of the 'repeat-until' construct at low-level requires an explicit loop control mechanism, with initialisation and termination phases. The 'repeat-until' construct tests the exit condition at the end of the loop as shown in *Figure 15.16*.

15.6.2.2 'While-do' construct

This construct provides for a process not to be executed, or to be executed one or more times. The number of times the process is executed depends on a value of a control variable, for example:

```
count :=number;
while count >ς=0 do
          P1;
```

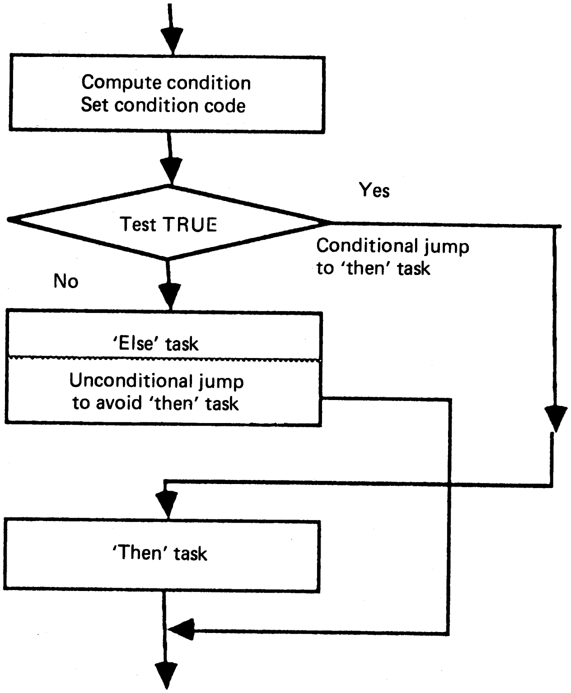


Figure 15.15 'If-then-else' construct

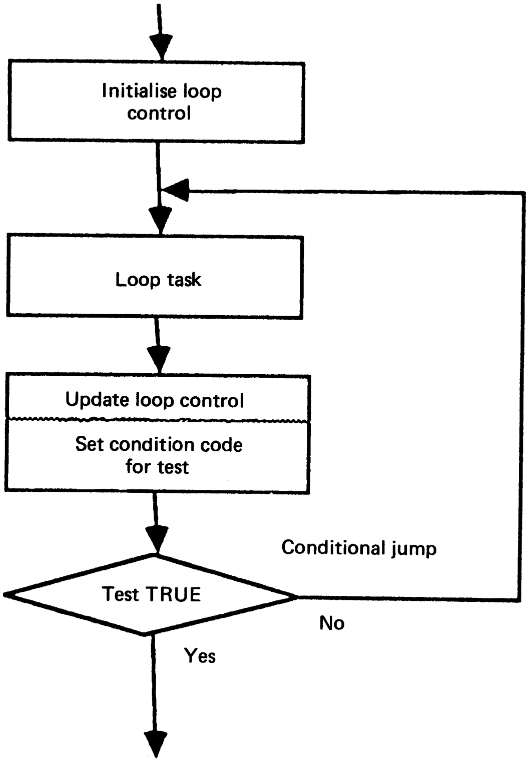


Figure 15.16 'Repeat-until' construct

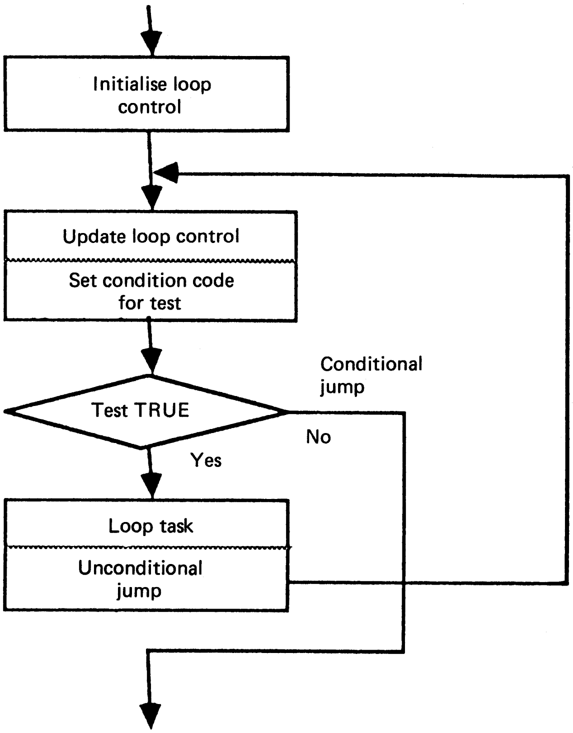


Figure 15.17 'While-do' construct

where the process P1 must update the loop control variable, as in:

```
count := count - 1;
```

Implementation of the 'while-do' construct at low-level requires an explicit loop control mechanism, with initialisation and termination phases. The terminating condition must be evaluated at the start of the loop. The general form of the 'while-do' construct is shown in *Figure 15.17.*

15.7 Reduced instruction set computers (RISC)

15.7.1 The reduced instruction set concept

The evolution in processor design during the 1980s and early 1990s led to increasingly complex processor architectures. The emerging 32-bit architectures accommodated a good range of data types and a wide range of instruction types and addressing modes. In addition, the flexible instructions often allowed the user a free choice of the addressing mode for the operands which resulted in many permutations of (variable-length) instructions. Such processors were characterised by the complexity of their instruction sets, the complex sequences necessary to 'fetch' the variable-length instructions, and the complex and very large instruction-decode logic. These processors were known as complex instruction set computers (CISC) processors.

Analysis of the actual use of these processors showed that the instruction set complexity often exceeded the needs of many applications. Many users simply preferred to use a familiar and well understood subset of the data types, addressing modes, and instructions. In addition, it was found that common programming constructs, such as conditional expressions, were often formed using simple relational operations, such as 'equals' and simple and short constants, such as 'zero'. Similarly, program branches often had a short relative offset that could be accommodated using short (16-bit) relative addressing. This led to the notion of developing high-performance processors with a reduced set of appropriate instructions.

15.7.2 The reduced instruction set (RISC) processor

The reduced instruction set computer (RISC) processor has a relatively simple register-to-register architecture which focuses on a CPU register file of, say, 32 general purpose 32-bit registers. Each register may hold data or addresses and can provide source operands and/or act as an accumulator for resultands. Since data can not be transferred directly between immediate-access memory and the ALU, simple data transfer (LOAD) instructions are used to transfer operand data from immediate-access memory to CPU registers so that it can be processed subsequently by the ALU. Similarly, simple data transfer (STORE) instructions must be used to transfer result data from CPU registers to immediate-access memory. Thus, if X, Y, and Z are stored in immediate-access memory, the high level expression $X:=Y+Z$ must be implemented using four separate instructions.

$R1:=Y$	LOAD operation, load CPU register R1 with Y from memory
$R2=Z$	LOAD operation, load CPU register R2 with Z from memory
$R3=R1+R2$	ALU operation, write into CPU register R3 sum R1 + R2
$X:=R3$	STORE operation, store CPU register R1 as X in memory.

The apparent disadvantage of using four low-level instructions to implement one higher-level memory-to-memory instruction is partially offset by the simple format of the RISC instructions. Register-to-register architecture processors have relatively short instructions since there is at most only one immediate-access memory reference operation (to load or store an operand) per instruction. This leads to the notion of using fixed-length instructions. A typical 32-bit RISC processor with 32 CPU registers will use a 32-bit fixed length instruction with, say, a 6-bit op-code, 5-bit direct addressing of CPU registers, and relative (base and offset) addressing of immediate-access memory using 16-bit relative addresses or immediate data. Thus LOAD and STORE instructions may consist of the 6-bit op-code, a 5-bit CPU register address for the operand or resultand, and an immediate-access memory address comprising a 5-bit base (register) address and a 16-bit relative offset. Similarly, an ALU instruction may comprise the 6-bit op-code, and either three 5-bit CPU register addresses (plus scope for instruction extensions), or two 5-bit CPU register addresses and a 16-bit immediate-data object. The use of a 32-bit instruction, which can be 'fetched' in a single memory reference operation, results in a fast instruction cycle and much simplified instruction decoder logic. The performance of RISC processors is further enhanced by the use of multi-stage instruction-pipelines.

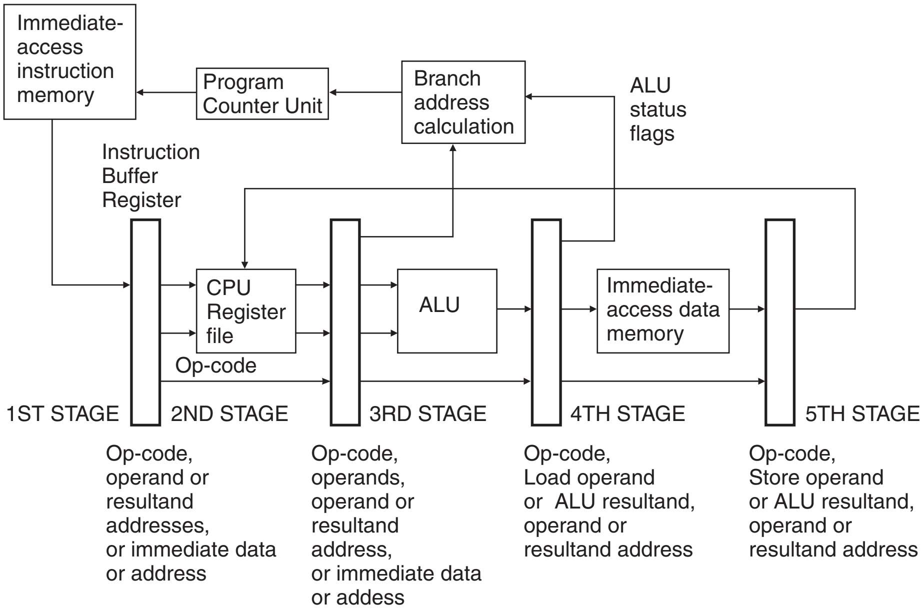

Immediate-access instruction memory
Program Counter Unit
Branch address calculation
ALU status flags
Instruction Buffer Register
CPU Register file
ALU
Immediate-access data memory
Op-code
1ST STAGE
2ND STAGE
3RD STAGE
4TH STAGE
5TH STAGE
Op-code, operand or resultand addresses, or immediate data or address
Op-code, operands, operand or resultand address, or immediate data or addess
Op-code, Load operand or ALU resultand, operand or resultand address
Op-code, Store operand or ALU resultand, operand or resultand address

Figure 15.18 Simplified instruction pipeline

15.7.3 Instruction pipelines

The conventional instruction ‘fetch’ and ‘execute’ cycle imposes speed constraints on a processor. The primary problem is the time taken to access immediate-access memory, both to ‘fetch’ the instruction and to read operands during LOAD instructions, or to write operands during STORE instructions. In each case the processor ALU is idle (not processing data) while the memory reference operation is taking place. This limitation can be overcome, to a large extent, by the notion of pipelining.

The basic idea of pipelining is to partition the processor into autonomous operational or functional units that can operate concurrently, (such as instruction ‘fetch’ from immediate-access memory, program counter update, CPU register-file read, ALU operation, write operation on immediate-access data memory, read operation on immediate-access memory read, CPU register-file write). Typically, this allows the process of ‘fetching’, decoding and executing an instruction to be split into the following stages.

(1) Instruction ‘fetch’ and program counter update.
(2) Operand read from CPU register file.
(3) ALU operation or effective address calculation (for LOAD and STORE).
(4) Immediate-access data memory read (for LOAD) or write (for STORE).
(5) Resultand (ALU operations) or operand (for LOAD) write to CPU register file.

In effect, the traditional instruction register is replaced by a pipeline of instruction buffers each of which store the information (opcode, CPU register addresses, memory reference address, operands, resultands, ALU status) necessary to carry out subsequent phases of the instruction cycle for each instruction, *Figure 15.18*. Thus, such a pipeline will normally contain five sequential instructions, and each stage of the pipeline will have the information and resources required to autonomously implement its stage of the instruction cycle. In addition, each stage includes at most one immediate-access memory reference operation, thus allowing fast (average) instruction throughput.

In practice, the pipeline speed-up is restricted by physical resource constraints (resource hazards) and temporal data access constraints (data hazards). For example, if the instruction memory and data memory share the same address and data bus (as in von Neumann architecture processors), then an immediate-access data memory reference operation for an instruction in stage (4) of the above pipeline can not take place concurrently with the instruction ‘fetch’ operation for a successor instruction in stage (1) of the pipeline, and a resource hazard occurs. This results in the need to stall the pipeline (with respect to the successor instructions) until the data buses are free. This problem can be overcome at some cost by adding resources, such as separate instruction and data memories (as in Havard architecture processors). Data hazards typically occur when an instruction attempts to use the results of a preceding instruction, but due to pipelining the results are unavailable. For example, the results of the preceding instruction in the pipeline may still be being computed or may not have been written back into the CPU register file at the point at which they are required as operands by the successor instruction. Such problems can be solved either by invoking pipeline stalls or by designing additional fast-track data paths which make results available direct from the ALU. Data hazards may occur in sequences of data processing operations, or in flow control operations when branches may be conditional on data generated by previous instructions.

A related problem affecting the performance of pipelined processors arises when computing decisions, such as IF-THEN-ELSE constructs, that involve alternative sequences. The instruction pipeline usually operates on the assumption that the instructions form a natural sequence, and no allowance is made for the presence of alternative sequences. Consequently, when the condition governing the branch-decision is computed (in stage (4) of the above pipeline), presumed-sequence instructions are already being processed in stages (1), (2) and (3) of the pipeline. If the branch is taken, the presumed-sequence instructions have to be deleted (or 'flushed') from the pipeline and the 'branch-taken' instructions 'fetched' and processed.

Modern RISC processors use many techniques to overcome these performance restrictions. Multi-stage pipelines may be used (up to 20 stages) to maximise concurrency. Multiple pipelines may be used to accommodate both of the alternative sequences for a decision, and this removes the delay associated with flushing and re-filling a pipeline. Persistent data (regularly used data objects) may be maintained in the CPU register file to reduce the need for multiple LOAD and STORE operations. First and second level high-speed access cache memories may be used for instructions and data to overcome the delay associated with accessing immediate-access memory. In all cases, the processor design process involves compromise between performance, complexity and cost. General purpose microprocessors and workstations often seek to maximise performance. Alternatively, in embedded applications, performance and complexity are often costly in terms of silicon area or 'real estate'. In an interesting development, proprietary embedded RISC processors are now available in a range of variants so that the designer can choose the level of performance, complexity, word length, and silicon area.

15.8 Software design

The process of software development involves turning the specification of the task to be performed into a program in a form which the processor can directly execute. The starting point for the design of such a system is the derivation of the system requirements specification. The specification should state what the system should do in a formal and precise manner. However, the requirements specification should not state how the task should be carried out or how it should be implemented.

The design procedure is normally a 'top-down' approach in which the requirements specification is translated into a design by a process of elaboration. The description of what the system should do is elaborated until the description comprises a set of easily implemented activities. In most design methods, the formal system specification will be analysed and decomposed on a functional basis. Careful consideration should be given to determining when to take decisions that bind or constrain the design, such as the choice of programming notation or processing architecture.

The design procedure often consists of a compromise between taking an early decision to map the problem onto a known implementation, such as a particular high-performance architecture, or delaying such decisions to retain the freedom of choice in design and implementation. It is desirable that the process of analysis is not subject to implementation constraints before the analysis has revealed the characteristics of the problem. This is particularly important if the designer is to exploit fully the advantages that can be obtained by using modern programming notations and processing hardware.

In applications which involve safety functions or have implications for safety, the system must perform in a reliable and safe manner. Ideally, the designer should prove the correctness of the design and the design should be translated into an implementation using proven translators. Finally, the implementation should be verified to show that it is fit for its intended use. However, current formal proof techniques require high levels of skill; they are also lengthy and are not efficient for complex systems.

15.8.1 Program development

The design and development of the computer programs or software is not a trivial task. Even a relatively simple program can have a sophisticated logical structure. Large software systems can have considerable complexity and special software engineering and management techniques have been developed to ensure that such systems can be designed and developed to the required quality within a specified time-scale and budget. The methods place considerable emphasis on the need to document relevant aspects of a system design including the test phase, they also address the important problem of maintaining a system throughout its lifecycle. They also provide a range of computer-aided software engineering tools to support the design activity.

Good software design techniques, such as structured programming, are used to produce readable, reliable and understandable programs. A restricted set of programming constructs or processes is allowed: sequential processes, selection, and repetition (see Section 15.6). Each such process has a single input and output and can be readily documented, tested and understood. Complex processes can be decomposed into a hierarchy of simpler sub-processes, each of which can be declared as subprograms or procedures to hide unnecessary detail. This improves the legibility of a program and helps the problem of managing complexity.

The structured programming approach leads naturally to a modular approach to program construction. The program is divided into modules, each comprising separate code and data. The data within any module is local to that module and communication between the modules has to be declared and is strictly controlled. This limits the interaction between modules and helps prevent error migration. It also encourages the documentation of module interfaces and inter-module actions; an understanding of this is essential when a program is modified.

15.8.2 Assembly languages

Programming languages may be either high-level and oriented to the solution of a particular class of problems, or low-level and oriented towards the architecture of a particular machine.

Assembly-language allows the designer to program in terms of the machine instructions that a specific processor can perform. Since binary machine-code instructions are difficult to understand directly, assembly-language programs are expressed in a symbolic notation. There is a one-to-one correspondence between each assembly-language instruction and a machine-code instruction.

Assembly-language programs have to be written in terms of the specific processor's instruction set and architecture, such as its CPU registers, memory locations, and input/output device registers. Also, memory storage has to be allocated explicitly for data objects using primitive data types. Assembly-language uses mnemonics for each machine level instruction. The mnemonics are usually specific to one processor or a family of processors and are chosen such that the function of the instruction is fairly obvious (e.g. ADD, SUB, MOV, etc.). In addition, the user has to define symbolic names for data objects such as variables (memory addresses), data constants, and labels (code locations).

The low-level code is translated into machine-code by a simple process of transliteration, this is usually carried out by a program known as an *assembler*. The assembler checks the syntax and usage of each of the instructions in the source text (the 'source code'), and produces cross-references for any jump, branch, or data access instructions.

The output of the assembler (the 'object code') can then be 'linked' with any library routines or external subroutines which are called from the program, and 'located' by inserting into the file the absolute addresses of the memory locations where the program code and data will loaded in the target system.

The unstructured nature of assembly-language programs increases the problem of testing such software. It is made even more difficult because testing conventionally takes place at machine level, rather than at the level of the symbolic assembly language. Normally, the software is executed on simulator, or on a target processor equipped with monitor program, with facilities for memory and CPU register examination, and for the insertion of breakpoints which allow the programmer to inspect and change memory or register contents at specified points in the program. Such testing requires skill, care and good management if it is to produce usable results.

Assembly-language programming, which requires a detailed understanding of instruction sets and processor architecture, is normally only necessary in applications where it is critical that the processing models and programming constructs used in the design are supported properly at machine level. Typical examples are compilers, the kernels of operating systems, interface software including interrupt handling, and certain aspects of real-time software. This is the province of the 'systems programmer' rather than the 'applications programmer'.

Knowledge of assembly language programming is not essential for general applications programming. Specifically, due to the lack of high-level constructs, assembly-language programs normally comprise an intimate mix of low-level program flow-control instructions and architecture-dependent data-processing instructions. Such programs are often difficult to design or comprehend. Thus, a programmer is advised to always use the highest level programming language appropriate for an application.

15.8.3 High-level languages

Instead of writing the programs in the assembly language of a processor, a high-level language can be used. The advantages to writing programs in a high-level language are as follows:

1. *Hardware independence*: the language is independent of the implementation hardware and can be compiled for a range of target processor.
2. *High-level notation*: the language comprises unambiguous statements which are often close to those used to express problems in natural language. This aids comprehension and increases the speed of programming.
3. *Structured code*: most high-level languages support structured programming and modular program construction.
4. *Data types*: most high level languages support a wide range of data types which allow checks on expression validity to be applied by the compiler.
5. *Maintenance*: clear program and data structures give easier program maintenance.

Few conventional high-level languages give direct access to machine-level primitives that are used for physical input and output, bit-level manipulation, machine control, and interrupts. Therefore, programs written in conventional high-level languages are normally restricted to using idealised input and output (e.g. to files) and are run in a protected environment provided by the operating system. However, this is not always sufficient for 'systems' programmers who are concerned with how the software, including the operating system and user programs, operates and performs and how it interfaces with the outside world. Thus, the programming languages used by systems programmers, such as 'C', typically include both high-level constructs and machine-level primitives.

A high-level language program is prepared as a source text file using a text editor (or using a word processing facility capable of producing a text file in the required format). The high-level language program is translated into an object program (i.e. nearly executable machine code) by a compiler. At the start of the compilation process, the compiler will check the syntax of the source program and any errors detected will be reported. During the compilation process, each construct or statement in the source program is translated into one or more lines of object code. The compiler may require several passes to convert the source program into the object code. Also, error checking may be included within the object code to detect run-time errors, such as array bounds exceeding predefined limits.

The compiler must be informed of the identity of the processor on which the compiled program is to run, so that it can produce processor-specific object code. Normally, the object code is generated for the processor on which the compiler runs. However, in the case of microprocessor systems, it is common to prepare programs on larger computers with better software engineering support facilities and to compile the source program for the intended target processor (this is known as cross-compilation).

The object code produced by the compiler requires linking, locating and loading before it can be executed,

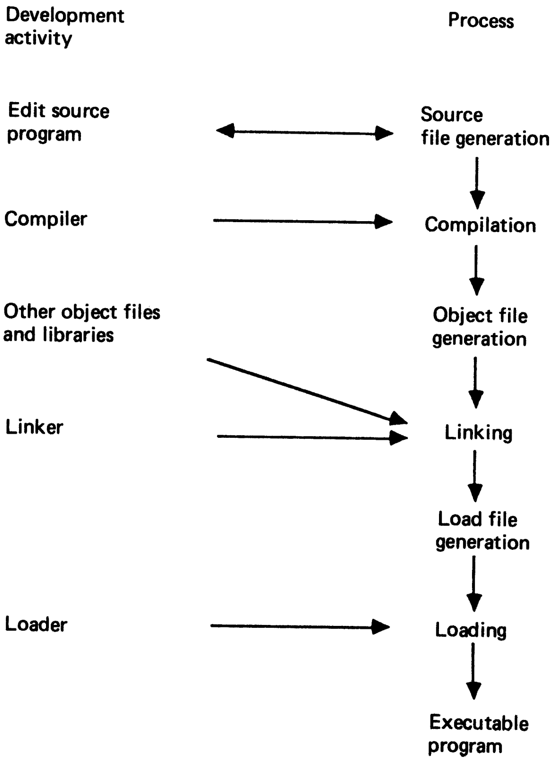


Figure 15.19 Stages in program development

as shown in *Figure 15.19*. The linker resolves all external references in the object code, such as references to library subroutines or other program modules held as object files, and combines all these modules into a single executable program. It also locates the program by adjusting the addresses all data memory references (variables) and code location references (labels) to the values of locations at which the code and data will be loaded. (In the case of relocatable code and data, it simply calculates the correct base addresses.) Finally, the linked object modules are loaded into memory at a defined location by the loader ready for execution. (For ease of communication, the object program is often downline loaded into the target processor in HEX form, and converted into binary by the loader during the loading process.) When the program is loaded, it is resident in memory in the target machine ready for execution.

15.8.4 Real-time processes

Many engineering systems, such as embedded computer real-time systems, are required to maintain synchronism with an asynchronous external system, or to respond to stimuli from such a system, within a finite and specified delay. In real-time programming there is a primary need for a mechanism for handling the concept of time. Real-time sequential programming languages include an additional primitive construct that allows the formal inclusion of time.

An application is said to be 'time-critical' if it must perform activities and produce responses at times dictated by an external environment. Typically in real-time control, a precise time-window is specified during which sensors must be sampled, a satisfactory control response computed, and output values sent to actuators. This schedule may re-occur periodically, or be initiated at irregular intervals by stimuli from the external system. Failure to perform the required functions in time is a fault; this may lead to system failure and may be hazardous.

The software for a time-critical application will comprise processes that must be properly synchronised with each other and with the external system. Synchronism with the external system is usually imposed by a 'real-time clock' driven schedule; these times will not be dictated by the optimum use of computing resources. To ensure that component processes do not overrun, it is common to place (critical) timing requirements on software execution and to provide a 'time-out' mechanism to warn of timing violations.

It is conventional to monitor the performance of a time-critical application process. Traditionally, this is done using a real-time time-lapse counter built as an external circuit. The counter is preset to trip after a pre-determined time and is initiated to run concurrently with the time-critical process. The first process to complete causes the other to abort. The mechanism is known as a 'watch-dog' timer.

If the application process has been properly designed, it will produce results well before the maximum allowed time and the 'watch-dog' timer will be aborted. The expiry of a 'watch-dog' timer or 'trip' indicates the presence of a fault (which may be a software design fault or a transient or permanent malfunction of the system) and appropriate fault recovery activities should be invoked. It is therefore necessary to set the pre-determined trip period to somewhat less than the time-critical time so that fault recovery can take place and the system can still provide a timely and satisfactory response.

15.8.5 Embedded real-time operating systems (RTOS)

A number of proprietary real-time operating systems (RTOS) are available for use in embedded microprocessor systems. These operating systems typically provide input/output handling, deterministic real-time task scheduling, watchdog facilities, and default behaviour under fault conditions. They are also characterised by having modest memory requirements. Significantly, these operating systems are targeted at a range of processors including conventional microprocessors, commercial off-the-shelf (COTS) microcomputers, embedded controllers (microprocessors with built-in analogue and digital signal acquisition and output generation), and the more recently introduced system-on-chip (SoC) systems with on-chip embedded processors.

15.9 Embedded systems

Many products have computers, microprocessors, or microcontrollers hidden or embedded within them. Some devices, such as digital organisers or personal digital assistants (PDAs), resemble small computers and the user may be provided with limited programming facilities for the embedded processor. In the case of other products, such as video games machines or consoles, it is fairly obvious that the device contains embedded processors and video display generators, even though the user is given little or no facility for programming the device. A much wider range of products contains embedded processors that are hidden or invisible to the user. For example embedded processors are commonly found in the following: cellular or mobile telephones, automobile electronics (such as engine management, braking systems, active suspension, intelligent sensing for lights or windscreen wipers, navigation systems, and both in-car and in-seat entertainment systems), office automation products (such as faxes, scanners, printers, copiers or duplicators, multimedia display projectors, computer network switches), domestic appliances (such as washing machines, dishwashers, tumble driers, cookers or ovens, microwave cookers, food blenders and processors, weigh scales, and vacuum cleaners), home entertainment units (such as radios, televisions, satellite receivers, digital set-top boxes, video cassette recorders, digital video disc (DVD) players, and hi-fi units), photographic equipment (such as film and digital cameras, analogue and digital video cameras) and personal or 'wearable' electronics (such as portable radios and CD players, MP3 players, watches, and fitness monitors). The trend is to increase significantly the sophistication of such systems, including user adaptation or personalisation, and to increase very significantly the data handling and processing requirements of the embedded processor. This is leading to the development of extremely powerful processors that are designed specifically for embedded applications, including small battery-powered products.

15.9.1 Embedded processors

The designer of embedded systems can choose from a wide range of processors. Classical microcontrollers, with integrated analogue inputs and outputs, are often used in control applications for domestic appliances or automotive systems. Specialist microcontrollers have been developed for applications such as electric motor control, or servo control. Traditional microprocessors or RISC processors may be used as stand alone devices within a more complex design.

Increasingly, the trend is to minimise the number of components in an embedded system. Traditionally, conventional microprocessors tend to be poorly suited to integration within, say, an FPGA since they are relatively complex and are not scalable (i.e. smaller versions can not be readily generated), although these constraints may become less important as the size of FPGAs increase. RISC processors are much better suited to integration because they have powerful ALUs, small instruction sets and instruction decoders, and are scalable in terms of word-length, concurrent resource, instruction pipeline, and clock speed. Embedded RISC processors are readily available as intellectual property (IP) cores.

Traditionally, IP cores have taken the form of hardware description language (HDL) software macros, which a designer incorporates within an application-specific design. However, the performance of such IP cores is layout dependent, and very careful design is necessary to achieve good speeds. Therefore, the preferred method of deriving high-performance designs is to use a proprietary ‘hard’ IP core that provides guaranteed performance (i.e. an IP core that has been pre-mapped into a fixed and full-tested hardware layout on the intended FPGA target). The second advantage of using proprietary ‘hard’ IP core RISC processors is that they are usually fully supported by proprietary embedded real-time operating system (RTOS) software and software tools for writing applications (such as editors and compilers). Thus the embedded systems design approach typically involves:

(1) Determining the application’s processing requirements and selecting a proprietary ‘hard’ IP RISC processor of appropriate performance, complexity, word length, and silicon area.
(2) Selecting a proprietary embedded RTOS that supports the processor.
(3) Developing application-specific software for software-implemented functions.
(4) Developing digital designs for any hardware-implemented functions and interfaces.

15.9.2 System on chip (SoC) design

The increase in size of VLSI logic circuits has led to a new generation of reconfigurable FPGA devices that provide a platform for ‘hard’ IP cores and are large enough to contain a complete high-performance digital processing system within a chip. These FPGAs are commonly called System on Chip (SoC) devices. An SoC device can be configured to include both ‘hard’ and ‘soft’ IP cores, plus user-designed digital circuits, *Figure 15.20*. At the heart of most SoC designs are an embedded proprietary RISC processor and a block of RAM memory (for the software that runs on the processor). This allows the designer to partition a design into those parts that will be implemented as software executing on the embedded processor (under an appropriate embedded RTOS) and those parts that will be implemented in hardware as high-speed application-specific logic circuits. This design approach, known as co-design or co-ware, has the significant advantage that established and high-performance parts of the design can be committed to application-specific hardware, and more adventurous parts of the design or low-speed functions can be committed to easily changed software.

The high-performance hardware implemented application-specific logic circuits may make extensive use of both ‘hard’- and ‘soft’- IP cores. In particular, SoC devices commonly provide co-processor support for digital signal processing, typically using a proprietary ‘hard’ IP digital signal processor (DSP). Alternatively, specific algorithms may be implemented using either ‘hard’- or ‘soft’- IP digital signal processing circuits. Similarly, signal acquisition may be facilitated using mixed-signal (digital/analogue) components such as analogue-to-digital (A/D) or digital-to-analogue (D/A) convertors. The problem of interfacing such major components is eased by the on-chip provision of advanced data highways or buses, standardised bus interfaces (for the major components such as the RISC/DSP processors and the memory), sophisticated clock generators, and a clock management unit.

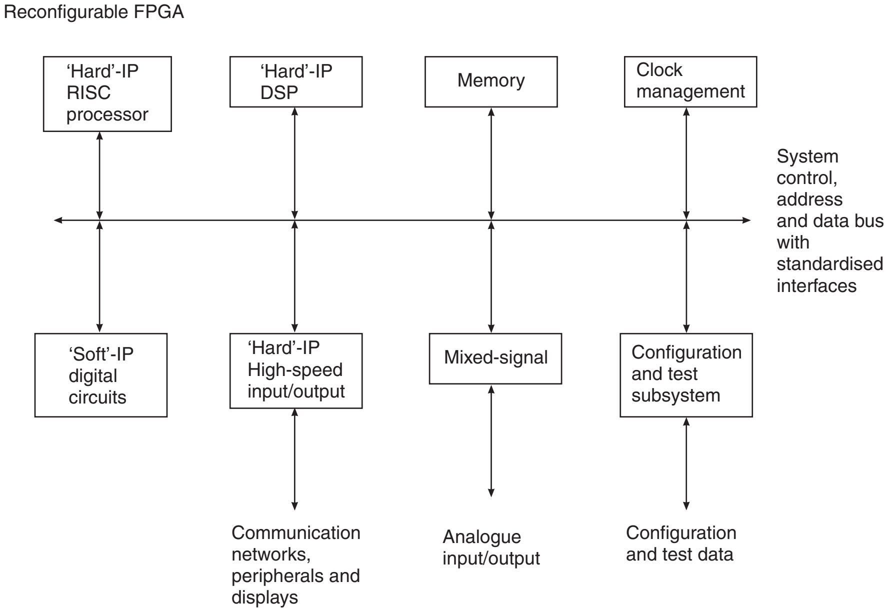

Reconfigurable FPGA
‘Hard’-IP RISC processor
‘Hard’-IP DSP
Memory
Clock management
System control, address and data bus with standardised interfaces
‘Soft’-IP digital circuits
‘Hard’-IP High-speed input/output
Mixed-signal
Configuration and test subsystem
Communication networks, peripherals and displays
Analogue input/output
Configuration and test data

Figure 15.20 Typical system-on-chip (SoC) configuration

External interfacing is facilitated by the use of ‘hard’ IP cores that support a wide range of high-bandwidth interface standards (for high-speed serial and parallel interfaces, including either system- or source-synchronous parallel interfaces). This is an area of on-going development, and both the SoC vendors and third-party IP providers are developing IP cores to support the ever widening range of communications protocols and interface standards found in the embedded systems market.

Modern SoC devices and the co-design approach, (involving embedded RTOS software, ‘soft’- and ‘hard’-IP cores, and bespoke digital design), provides a manageable and flexible route to embedded systems design. It may also change the relative roles of the hardware and software designer in the development of SoC designs. The reconfigurable SoC devices can be deployed in a wide range of applications and the common-platform nature of the devices helps avoid the non-recurring costs associated with fully bespoke designs. The extensive use of ‘hard’- and ‘soft’-IP from SoC vendors and third party suppliers minimises risk and facilitates time-to-market which gives competitive advantage. However, it also involves entering into IP license arrangements for both product development and deployment. It will be interesting to see whether the rapidly decreasing SoC product design cycles are complemented by corresponding decreases in the time, complexity, and cost of negotiating multi-party IP license agreements.

16

Programmable Controllers

E A Parr MSc, CEng, MIEE, MInstMC
CoSteel Sheerness

Contents

16.1 Introduction 16/3
  16.1.1 The computer in control 16/3
  16.1.2 Requirements for industrial control 16/3
  16.1.3 Enter the PLC 16/4
  16.1.4 The advantages of PLC control 16/5

16.2 The programmable controller 16/6
  16.2.1 Modern PLC systems 16/6
  16.2.2 I/O connections 16/6
  16.2.3 Remote I/O 16/10
  16.2.4 The program scan 16/10

16.3 Programming methods 16/13
  16.3.1 Introduction 16/13
  16.3.2 I/O identification 16/14
  16.3.3 Ladder logic 16/16
  16.3.4 Logic symbols 16/17
  16.3.5 Statement list 16/18
  16.3.6 Bit storage 16/20
  16.3.7 Timers 16/22
  16.3.8 Counters 16/23
  16.3.9 Combinational logic 16/24
  16.3.10 Event driven logic and SFCs 16/24
  16.3.11 IEC 1131 16/27

16.4 Numerics 16/29
  16.4.1 Numerical applications 16/29
  16.4.2 Numeric representations 16/30
  16.4.3 Data movement 16/31
  16.4.4 Data comparison 16/33
  16.4.5 Arithmetical operations 16/34
  16.4.6 Analog signals 16/35
  16.4.7 Closed loop control 16/38
  16.4.8 Intelligent modules 16/39

16.5 Distributed systems and fieldbus 16/41
  16.5.1 Introduction 16/41
  16.5.2 Transmission lines 16/41
  16.5.3 Network topologies 16/41
  16.5.4 Network sharing 16/42
  16.5.5 A communication hierachy 16/42
  16.5.6 Proprietary systems 16/43
  16.5.7 Ethernet 16/43
  16.5.8 Towards standardisation 16/43

16.6 Graphics 16/45

16.7 Software engineering 16/48

16.8 Safety 16/48

16.1 Introduction

16.1.1 The computer in control

A computer can be considered as a device that follows predetermined instructions to manipulate input data in order to produce new output data as summarised on *Figure 16.1*(*a*). Early computer systems tended to be based on commercial functions; payroll, accountancy, banking and similar activities. The operations tended to be batch processes; a daily update of stores stock for example.

A computer can also be used as part of a control system as *Figure 16.1*(*b*). The input data will be the operator's commands and signals from the plant (limit switches, flows, temperatures). The output data are control actions to the plant and status displays to the operator. The instructions will define what action is to be taken as the input data (from both the plant and the operator) changes.

The first industrial computer application was probably a system installed in an oil refinery in Port Arthur USA in 1959. The reliability and mean time between failure of computers at this time meant that little actual control was performed by the computer, and its role approximated to a simple monitoring subsystem.

16.1.2 Requirements for industrial control

Industrial control has rather different requirements than other computer applications. It is worth examining these in some detail.

A conventional computer takes data, usually from a keyboard, and outputs data to a screen or printer. The data being manipulated will generally be characters or numbers (e.g. item names and quantities held in a stores stock list).

An industrial control computer is very different. Its inputs come from a vast number of devices. Although some of these will be numeric (flows, temperature, pressures and similar analog signals) the majority will be single bit, on/off, digital signals representing valves, limit switches, motor contactors etc.

There will also be a similar large amount of digital and analog output signals. A very small control system may have connections to about twenty input and output signals; figures of over two hundred connections are quite common on medium sized systems.

Although it is possible to connect this quantity of signals into a conventional machine, it requires non-standard connections and external boxes. Similarly, although programming for a large amount of input and output signals can be done in Pascal, BASIC or C, the languages are being used for a purpose for which they were not really designed, and the result can be very ungainly.

In *Figure 16.2*(*a*), for example, we have a simple motor starter. This could be connected as a computer driven circuit as *Figure 16.2*(*b*). The two inputs are identified by addresses 1 and 2, with the output (the relay starter) being given the address 10.

If we assume a program function bitread (*N*) exists which gives the state (on/off) of address *N*, and a function bitwrite (*M*, var) which sends the state of program variable var to address *M*, we could give the actions of *Figure 16.2* by

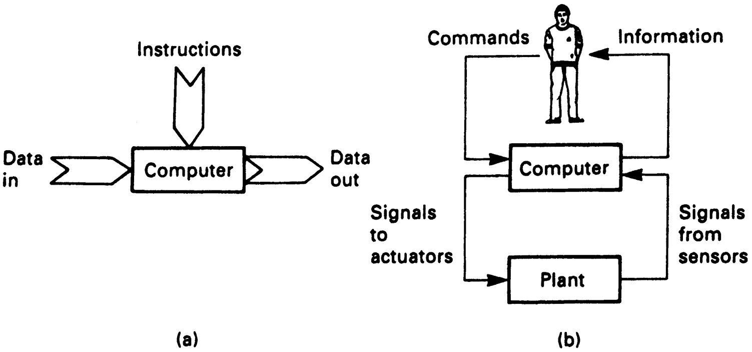


Figure 16.1 The computer as part of an industrial control system: (a) a simple overview of a computer; (b) the computer as part of a control system

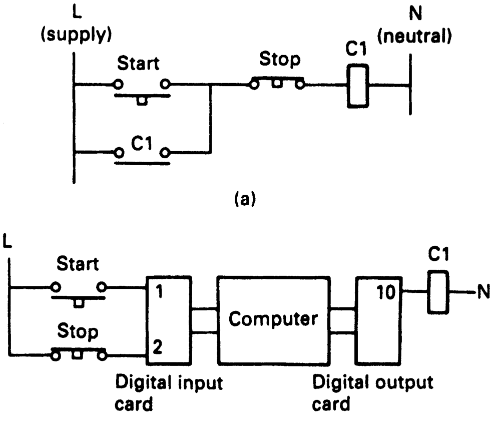


Figure 16.2 Comparison of hardwire and computer based systems: (a) hardwire motor starter; (b) computer based motor starter

```
repeat
    start:= bitread(1);
    stop:= bitread(2);
    run:= ((start) or (run)) & stop;
    bitwrite (10,run);
until hellfreezesover
```

where start, stop and run are one bit variables. The program is not very clear, however, and we have just three connections.

An industrial control program rarely stays the same for the whole of its life. There are always modifications to cover changes in the operations of the plant. These changes will be made by plant maintenance staff, and must be made with minimal (preferably none) interruptions to the plant production. Adding a second stop button and a second start button into *Figure 16.2* would not be a simple task.

In general, computer control is done in real time, i.e. the computer has to respond to random events as they occur. An operator expects a motor to start (and more important to stop!) within a fraction of a second of a button being pressed. Although commercial computing needs fast computers, it is unlikely that the difference between a one second and two second computation time for a spreadsheet would be noticed by the user. Such a difference would be unacceptable for industrial control.

Time itself is often part of the control strategy (e.g. start air fan, wait 10 secs for air purge, open pilot gas valve, wait 0.5 s, start ignition spark, wait 2.5 s, if flame present open main gas valve). Such sequences are difficult to write with conventional languages.

Most control faults are caused by external items (limit switches, solenoids and similar devices) and not by failures within the central control itself. The permission to start a plant, for example, could rely on signals involving cooling water flows, lubrication pressure and temperatures all being within allowable ranges. For quick fault finding the maintenance staff must be able to monitor the action of the computer program whilst it is running. If, as is quite common, there are ten interlock signals which allow a motor to start, the maintenance staff will need to be able to check these quickly in the event of a fault. With a conventional computer, this could only be achieved with yet more complex programming.

The power supply in an industrial site is shared with many antisocial loads; large motors stopping and starting, thyristor drives which put spikes and harmonic frequencies onto the mains supply. To a human these are perceived as light flicker; to a computer they can result in storage corruption or even machine failure.

An industrial computer must therefore be able to live with a 'dirty' mains supply, and should also be capable of responding sensibly following a total supply interruption. Some outputs must go back to the state they were in before the loss of supply, others will need to turn off or on until an operator takes corrective action. The designer must have the facility to define what happens when the system powers up from cold.

The final considerations are environmental. A large mainframe computer generally sits in an air conditioned room at a steady 20 °C with carefully controlled humidity. A desk top PC will normally live in a fairly constant office environment because human beings do not work well at extremes. An industrial computer, however, will probably have to operate away from people in a normal electrical substation with temperatures as low as −10 °C after a winter shutdown, and possibly over 40 °C in the height of summer. Even worse, these temperature variations lead to a constant expansion and contraction of components which can lead to early failure if the design has not taken this factor into account.

To these temperature changes must be added dust and dirt. Very few industrial processes are clean, and the dust gets everywhere. The dust will work itself into connectors, and if these are not of a highest quality, intermittent faults will occur which can be very difficult to find.

In most computer applications, a programming error or a machine fault can often be humorous (bills and reminders for 0*p*) or at worse expensive and embarrassing. When a computer controlling a plant fails, or a programmer misunderstands the plants operation, the result could be injuries or fatalities. It behoves everyone to take extreme care with the design.

Our requirements for industrial control computers are very demanding, and it is worth summarising them:

- They should be designed to survive in an industrial environment with all that this implies for temperature, dirt and poor quality mains supply.
- They should be capable of dealing with bit form digital input/output signals at the usual voltages encountered in industry (24 V d.c. to 240 V a.c.) plus analog input/output signals. The expansion of the I/O should be simple and straightforward.
- The programming language should be understandable by maintenance staff (such as electricians) who have no computer training. Programming changes should be easy to perform in a constantly changing plant.
- It must be possible to monitor the plant operation whilst it is running to assist fault finding. It should be appreciated that most faults will be in external equipment such as plant mounted limit switches, actuators and sensors, and it should be possible to observe the action of these from the control computer.
- The system should operate sufficiently fast for real-time control. In practice, 'sufficiently fast' means a response time of around 0.1 sec, but this can vary dependent on the application and the controller used.
- The user should be protected from computer jargon.
- Safety must be a prime consideration.

16.1.3 Enter the PLC

In the late 1960s the American motor car manufacturer General Motors was interested in the application of computers to replace the relay sequencing used in the control of its automated car plants. In 1969 it produced a specification for an industrial computer similar to that outlined at the end of the previous section.

Two independent companies, Bedford Associates (later called Modicon) and Allen Bradley (now owned by Rockwell) responded to General Motors specifications. Each produced a computer system similar to *Figure 16.3* which bore little resemblance to the commercial minicomputers of the day.

The computer itself, called the central processor, was designed to live in an industrial environment, and was connected to the outside world via racks into which input, or output cards could be plugged.

Each input or output card could connect to 16 signals. A typical rack would contain eight cards and the processor could connect to eight racks, allowing connection to 1024

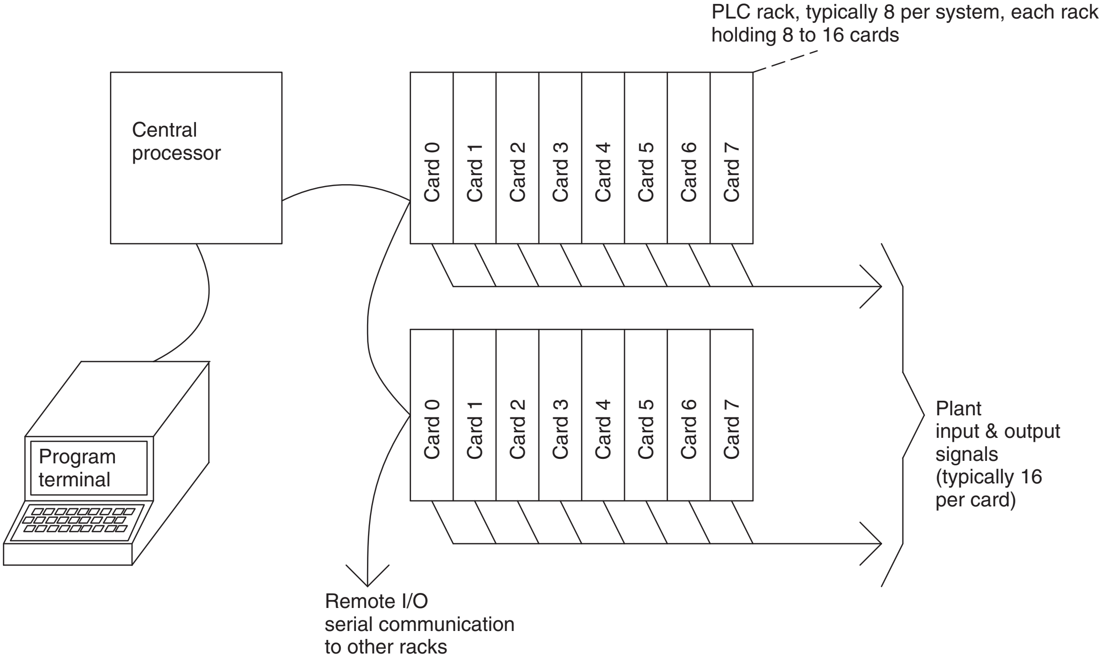

PLC rack, typically 8 per system, each rack holding 8 to 16 cards
Central processor
Card 0
Card 1
Card 2
Card 3
Card 4
Card 5
Card 6
Card 7
Program terminal
Plant input & output signals (typically 16 per card)
Remote I/O serial communication to other racks

Figure 16.3 The component parts of an early PLC system

devices. It is very important to appreciate that the card allocations were the user's choice, allowing great flexibility.

The most radical idea, however, was a programming language based on a relay schematic diagram, with inputs (from limit switches, pushbuttons, etc.) represented by relay contacts, and outputs (to solenoids, motor starters, lamps, etc.) represented by relay coils. *Figure 16.4*(*a*) shows a simple hydraulic cylinder which can be extended or retracted by pushbuttons. Its stroke is set by limit switches which open at the end of travel, and the solenoids can only be operated if the hydraulic pump is running. This would be controlled by the computer program of *Figure 16.4*(*b*) which is identical to the relay circuit needed to control the cylinder. These programs look like the rungs on a ladder, and were consequently called 'Ladder Diagrams'.

The program was entered via a programming terminal with keys showing relay symbols (normally open/normally closed contacts, coils, timers, counters, parallel branches, etc.), with which a maintenance electrician would be familiar. *Figure 16.5* shows the programmer's keyboard for an early PLC. The meaning of the majority of the keys should be obvious to any maintenance electrician. The program, shown exactly on the screen as *Figure 16.4*(*b*), would highlight energised contacts and coils allowing the programming terminal to be used for simple faultfinding.

The name given to these machines was *Programmable Controllers* or PCs. The name *Programmable Logic Controller* or PLC was also used, but this is, strictly, a registered trade mark of the Allen Bradley Company, now part of Rockwell. Unfortunately in more recent times the letters PC have come to be used for Personal Computer, and confusingly the worlds of programmable controllers and personal computers overlap where portable and lap-top computers are now used as programming terminals. To avoid confusion, we shall use PLC for a programmable controller and PC for a personal computer.

16.1.4 The advantages of PLC control

Any control system goes through several stages from conception to a working plant.

The first stage is *Design* when the required plant is studied and the control strategies decided. With conventional systems every '*i*' must be dotted before construction can start. With a PLC system all that is needed is a possibly (usually!) vague idea of the size of the machine and the I/O requirements (so many inputs and outputs). The cost of the input and output cards are cheap at this stage, so a healthy spare capacity can be built in to allow for the inevitable omissions and future developments.

Next comes *Construction*. With conventional schemes, every job is a 'one-off' with inevitable delays and costs. A PLC system is simply bolted together from standard parts.

The next stage is *Installation*, a tedious and expensive business as sensors, actuators, limit switches and operator controls are cabled. A distributed PLC system (discussed in Section 16.5) using serial links and pre-built and tested desks can simplify installation and bring huge cost benefits. The majority of the PLC program is usually written at this stage.

Finally comes *Commissioning*, and this is where the real advantages are found. No plant ever works first time. Human nature being what it is, there will be some oversights. (We need a limit switch to only allow feeding when the discharge valve is 'shut' or 'Whoops, didn't we say the loading valve is energised to UNLOAD on this system' and so on.) Changes to conventional systems are time consuming and expensive. Provided the designer of the PLC systems has built in spare memory capacity, spare I/O and a few spare cores in multi-core cables, most changes can be made quickly and relatively cheaply. An added bonus is that all changes are inherently recorded in the PLC's program and commissioning modifications do not go unrecorded.

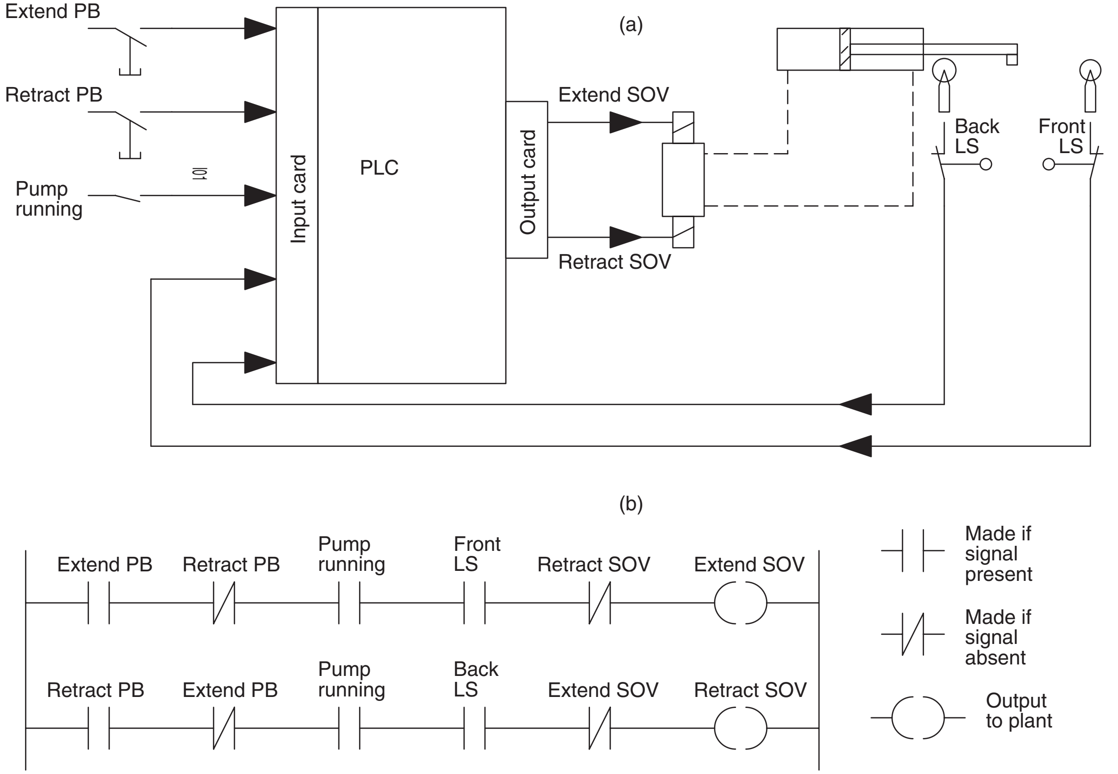


Figure 16.4 A simple PLC application: (a) a hydraulic cylinder controlled by a PLC; (b) the 'Ladder Diagram' program used to control the cylinder

There is an additional fifth stage called *Maintenance* which starts once the plant is working and is handed over to production. All plants have faults, and most tend to spend the majority of their time in some form of failure mode. A PLC system provides a very powerful tool for assisting with fault diagnosis.

A plant is also subject to many changes during its life to speed production, ease breakdowns or because of changes in its requirements. A PLC system can be changed so easily that modifications are simple and the PLC program will automatically document the changes that have been made.

16.2 The programmable controller

16.2.1 Modern PLC systems

This chapter is written around five manufacturers' ranges:

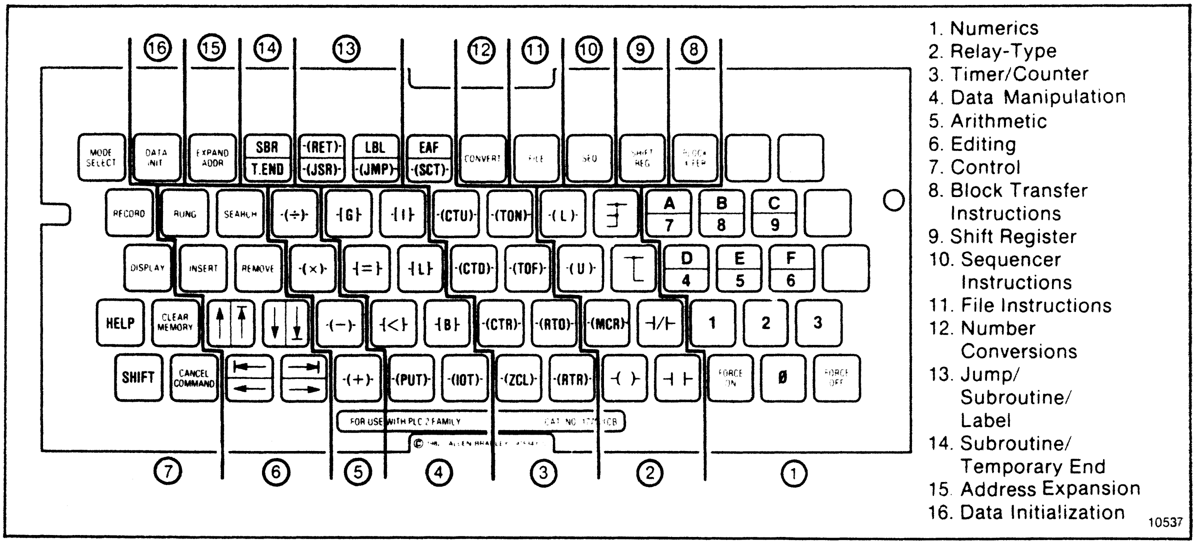
- The Allen Bradley PLC-5 series. Allen Bradley, now owned by Rockwell, were one of the original PLC originators (and actually has the US copyright on the name PLC). They have been responsible for much of the development of the ideas used in PLCs and have succeeded in maintaining a fair degree of upward compatibility from their earliest machine without restricting the features of the latest.
- The Siemens Simatic 55 range which is probably the commonest PLC in mainland Europe.
- The British GEM-80, originally designed by GEC from a long association with industrial computers dating back to English Electric. This part of GEC is now known as CEGELEC and is part of a French group in which Alsthom are a major shareholder.
- The ASEA Master System, now manufactured by the ABB company formed by the merger of ASEA and Brown Boveri. The Master system has features more akin to a conventional computer system and its programming language has some interesting and powerful features.

The above four PLCs are shown on *Figure 16.6*. Many PLC systems are now very small and as an example of this bottom end of the market we shall also consider the Japanese Mitsubishi F2-40.

16.2.2 I/O connections

Internally a computer usually operates at 5 V d.c. The external devices (solenoids, motor starters, limit switches, etc.) operate at voltages up to 110 V a.c. The mixing of these two voltages will cause irreparable damage to the PLC electronics. A less obvious problem can occur from electrical 'noise' introduced into the PLC from voltage spikes, caused by interference on signals lines, or from load currents flowing in a.c. neutral or d.c. return lines. Differences in earth potential between the PLC cubicle and outside plant can also cause problems.

There are obviously very good reasons for separating the plant supplies from the PLC supplies with some form of barrier to ensure that the PLC cannot be adversely affected by anything happening on the plant. Even a cable fault putting 415 V a.c. onto a d.c. input would only damage the input card; the PLC itself (and the other cards in the system) would not suffer.


⑯
⑮
⑭
⑬
⑫
⑪
⑩
⑨
⑧
MODE SELECT
DATA INIT
EXPAND ADDR
SBR
T.END
-(RET)-
-(JSR)-
LBL
-(JMP)-
EAF
-(SCT)-
CONVERT
FILE
SEQ
SHIFT REG
BLOCK XFER
RECORD
RUNG
SEARCH
-(÷)-
-[G]-
-[I]-
-(CTU)-
-(TON)-
-(L)-
A 7
B 8
C 9
DISPLAY
INSERT
REMOVE
-(×)-
-[=]-
-[L]-
-(CTD)-
-(TOF)-
-(U)-
D 4
E 5
F 6
HELP
CLEAR MEMORY
-(−)-
-[<]-
-[B]-
-(CTR)-
-(RTO)-
-(MCR)-
-|/|-
1
2
3
SHIFT
CANCEL COMMAND
-(+)-
-(PUT)-
-(IOT)-
-(ZCL)-
-(RTR)-
-( )-
-| |-
FORCE ON
0
FORCE OFF
FOR USE WITH PLC-2 FAMILY
CAT. NO.
© ALLEN-BRADLEY
⑦
⑥
⑤
④
③
②
①
1. Numerics
2. Relay-Type
3. Timer/Counter
4. Data Manipulation
5. Arithmetic
6. Editing
7. Control
8. Block Transfer Instructions
9. Shift Register
10. Sequencer Instructions
11. File Instructions
12. Number Conversions
13. Jump/ Subroutine/ Label
14. Subroutine/ Temporary End
15. Address Expansion
16. Data Initialization
10537

1. Numerics—provides addresses and decimal or hexadecimal values for instructions. It also provides force instructions.
2. Relay-Type—examines and controls the status of individual bits in specified memory areas.
3. Timer/Counter—allows the user to select various time-incremented and count-incremented and decremented functions.
4. Data Manipulation—used to transfer and compare BCD or octal values in the user program.
5. Arithmetic—performs the four indicated math functions.
6. Editing—used to locate, display and change instructions in the user program.
7. Control—directs the operation of the industrial terminal and its communication with the PLC-2 family processors and peripherals. Also provides HELP information.
8. Block Transfer Instructions—used to program block transfer Instructions in block format.
9. Shift Register—used to shift a word (all 16 bits) up or down one word in the shift register file.
   - used to shift a bit in the shift register file to the left or right one position.
   - used to create FIFO stacks.
10. Sequencer Instructions—used to establish and maintain user sequencer tables.
11. File Instructions—used to establish and manipulate user files.

Figure 16.5 A programming keyboard from an early PLC programming terminal. The link between the keys and relay symbols can be clearly seen. Figure courtesy of Allen Bradley

(a)

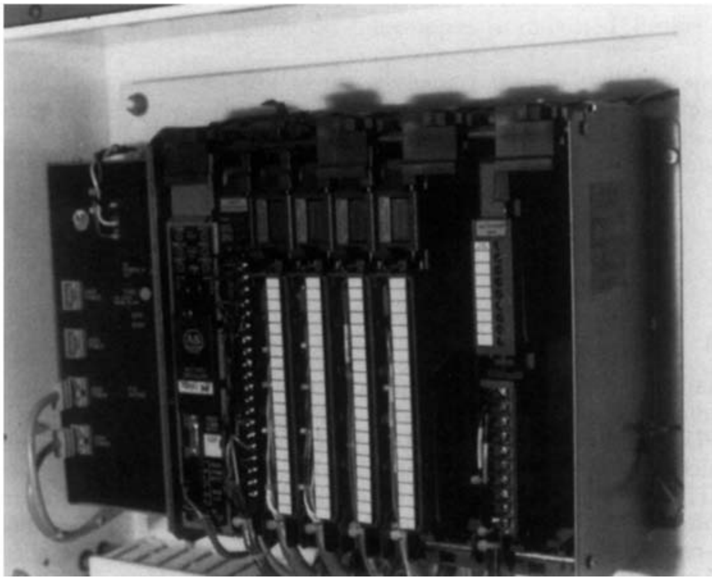


Figure 16.6 Four medium sized PLCs: (a) the Allen Bradley PLC-5; (b) the Siemens 115U; (c) the CEGELEC GEM-80; (d) the ABB Master

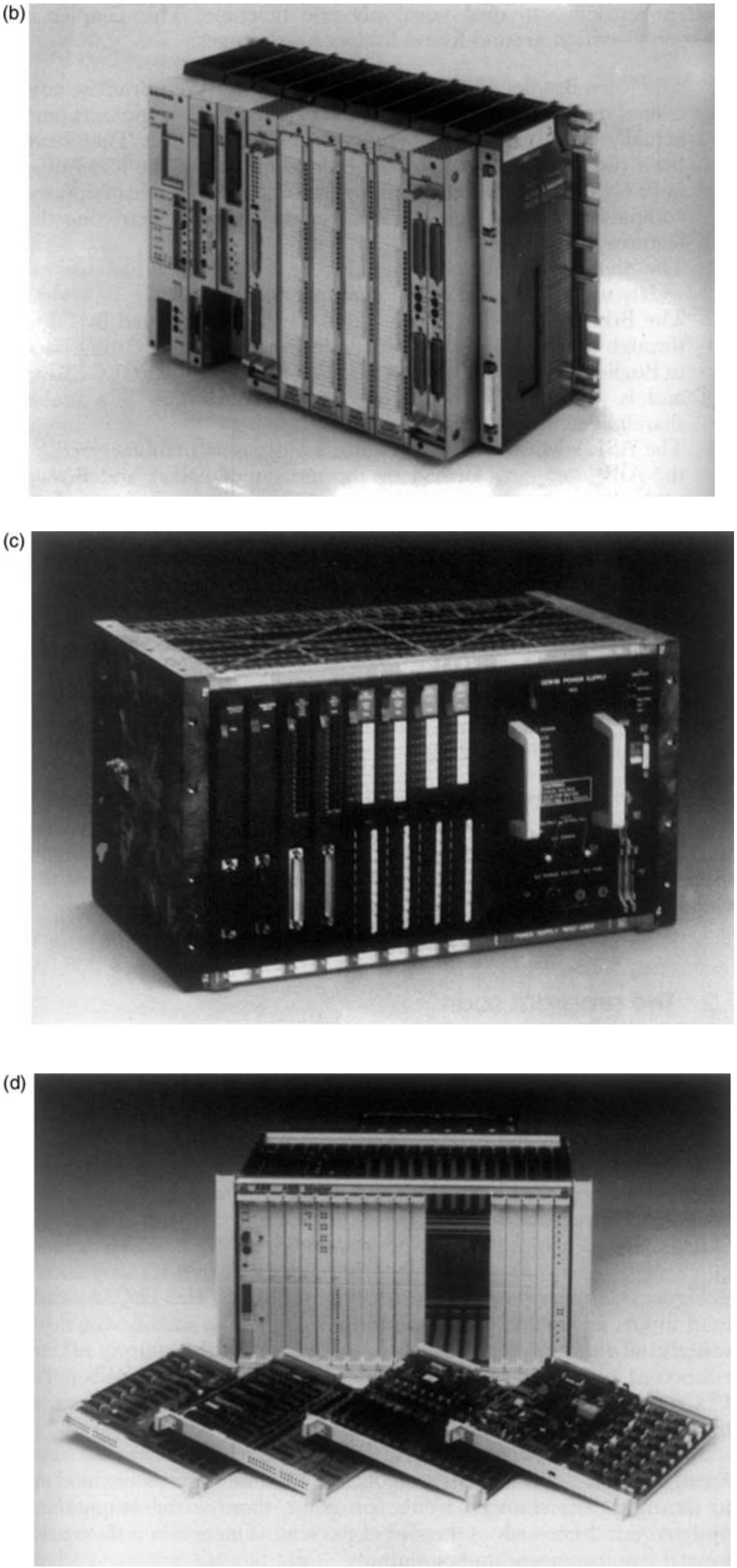


Figure 16.6 (*continued*)

This isolation is achieved by optical isolators consisting of a linked light emitting diode and photoelectric transistor. When current is passed through the diode it emits light causing the transistor to switch on. Because there is no electrical connections between the diode and the transistor, very good electrical isolation (typically 1–4 KV) is achieved.

A d.c. input can be provided as *Figure 16.7*(*a*). When the push button is pressed, current will flow through D1 causing TR1 to turn on passing the signal to the PLC internal logic. Diode D2 is a light emitting diode used as a fault finding aid to show when the input signal is present. Such indicators are present on almost all PLC input and output cards. The resistor R sets the voltage range of the input. D.c. input cards are usually available for three voltage ranges; 5 V (TTL), 12–24 V, 24–50 V.

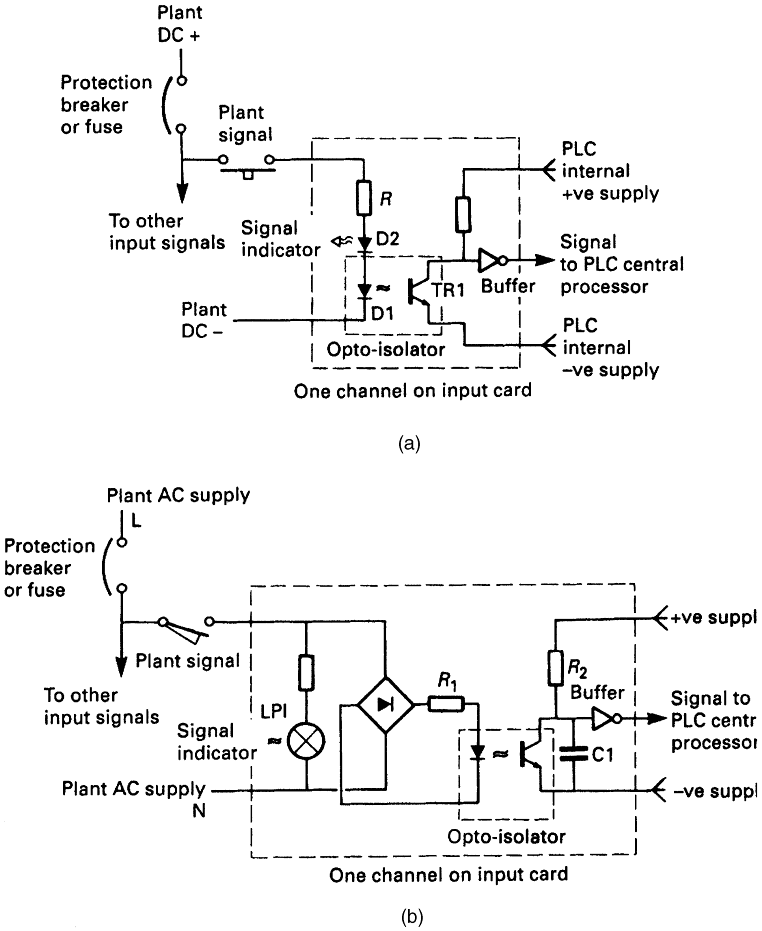
A possible a.c. input circuit is shown on *Figure 16.7*(*b*). The bridge rectifier is used to convert the a.c. to full wave rectified d.c. Resistor R2 and capacitor C1 act as a filter (typically 50 ms time constant) to give a clean signal to the PLC logic. As before a neon LP1 acts as an input signal indicator for fault finding, and resistor R1 sets the voltage range.

Output connections also require some form of isolation barrier to limit damage from the inevitable plant faults and to stop electrical 'noise' corrupting the processor's operations. Interference can be more of a problem on outputs because higher currents are being controlled by the cards and the loads (solenoids and relay coils) are often inductive.

In *Figure 16.8*, eight outputs are fed from a common supply, which originates local to the PLC cubicle (but separate from the supply to the PLC itself). This arrangement is the simplest and the cheapest, to install. Each output has its own individual fuse protection on the card and a common circuit breaker. It is important to design the system so that a fault, say, on load 3 blows the fuse FS3 but does not trip the supply to the whole card shutting down every output. This is known as '*discrimination*'.

Contacts have been shown on the outputs in *Figure 16.8*. Relay outputs can be used (and do give the required isolation) but are not particularly common. A relay is an electromagnetic device with moving parts and hence a finite limited life. A purely electronic device will have greater reliability. Less obviously, though, a relay driven inductive load can generate troublesome interference and lead to early contact failure.

A transistor output circuit is shown on *Figure 16.9*(*a*). Opto-isolation is again used to give the necessary separation between the plant and the PLC system. Diode D1 acts as a spike suppression diode to reduce the voltage spike encountered with inductive loads as shown on *Figure 16.9*(*b*). The output state can be observed on LED1. *Figure 16.9*(*a*) is a



Plant
DC +
Protection breaker or fuse
Plant signal
PLC internal +ve supply
R
To other input signals
Signal indicator
D2
Signal to PLC central processor
TR1
Buffer
Plant DC −
D1
PLC internal −ve supply
Opto-isolator
One channel on input card
(a)
Plant AC supply
L
Protection breaker or fuse
+ve suppl
Plant signal
R2
Buffer
R1
To other input signals
LPI
Signal to PLC centr processoı
Signal indicator
C1
Plant AC supply N
−ve suppl
Opto-isolator
One channel on input card
(b)

Figure 16.7 Optical isolation of input signals: (a) d.c. input; (b) a.c. input

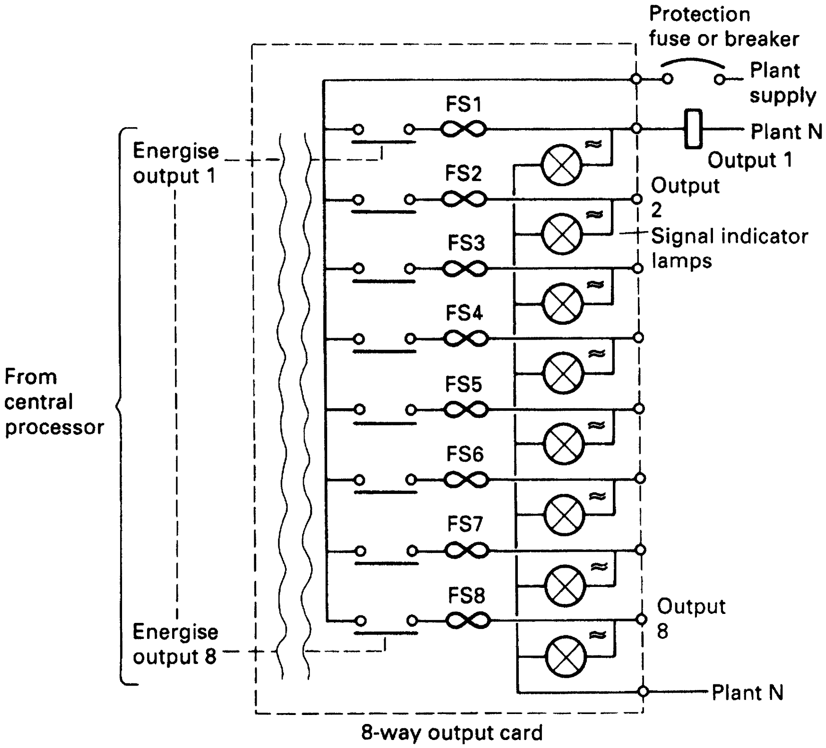


Figure 16.8 Schematic of an 8 way output card with common supply

current sourcing output. If NPN transistors are used, a current sinking card can be made as *Figure 16.9*(*c*).

A.c. output cards invariably use triacs. a typical circuit being shown on *Figure 16.10*. Triacs have the advantage that they can be made to turn on at zero voltage and inherently turn off at zero current in the load. The zero current turn off eliminates the spike interference caused by breaking the current through an inductive load. If possible, all a.c. loads should be driven from triacs rather than relays.

An output card will have a limit to the current it can supply, usually set by the printed circuit board tracks rather than the output devices. An individual output current will be set for each output (typically 2 A) and a total overall output (typically 6 A). Usually the total allowed for the card current is lower than the sum of the allowed individual outputs.

16.2.3 Remote I/O

So far we have assumed that a PLC consists of a processor unit and a collection of I/O cards mounted in local racks. Early PLCs were arranged like this, but in a large and scattered plant, all signals had to be brought back to some central point in expensive multi-core cables. This also makes commissioning and fault finding rather difficult, as signals can only be monitored effectively at a point distant from the plant device being tested.

In all bar the smallest and cheapest systems, PLC manufacturers therefore provide the ability to mount I/O racks remote from the processor, and linked with simple (and cheap) screened single pair or fibre optic cable. Racks can then be mounted up to several kilometres away from the processor.

There are many benefits from this. It obviously reduces cable costs as racks can be laid out local to the plant devices and only short multi-core cable runs are needed. The long runs will only be the communication cables (which are cheap, easy to install and only have a few cores to terminate at each end) and hardwire safety signals.

Less obviously, remote I/O allows complete plant units to be constructed, wired to a built in PLC rack, and tested off site prior to delivery and installation. Typical examples are hydraulic skids, desks and even complete control pulpits. The use of remote I/O in this way can greatly reduce installation and commissioning time and cost.

The use of serial communication for remote I/O means some form of sequential scan must be used to read input and update outputs. This scan, typically 30–50 ms, introduces a small delay in the response to signals discussed further in the following section.

If remote I/O is used, provision should be made for a program terminal to be connected local to each rack. It negates most of the benefits if the designer can only monitor the operation from a central control room several hundred metres from the plant. Fortunately, manufacturers have recognised this and most PLCs have programming terminals which can be remotely connected to the processor.

16.2.4 The program scan

A PLC program can be considered to behave as a permanent running loop similar to *Figure 16.11*(*a*). The user's instructions are obeyed sequentially, and when the last instruction has been obeyed the operation starts again at the first instruction. A PLC does not, therefore, communicate

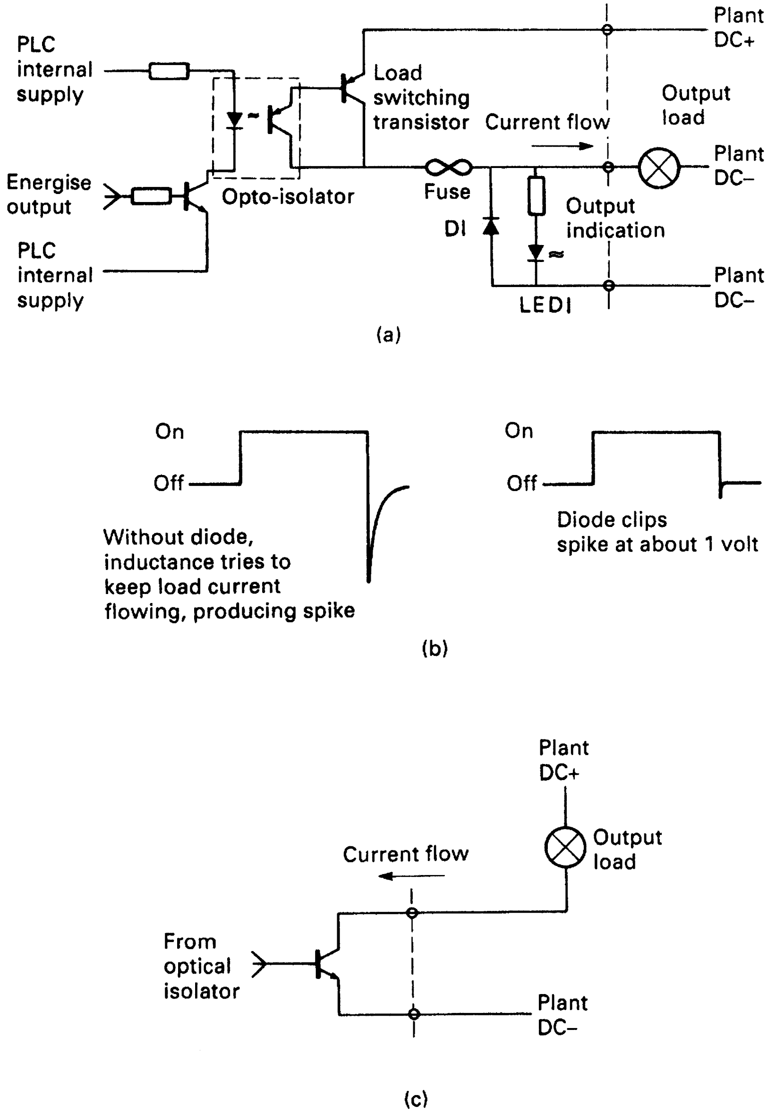

PLC internal supply
Energise output
PLC internal supply
Opto-isolator
Load switching transistor
Current flow
Fuse
D1
Output indication
LED1
Plant DC+
Output load
Plant DC–
Plant DC–
(a)
On
Off
Without diode, inductance tries to keep load current flowing, producing spike
On
Off
Diode clips spike at about 1 volt
(b)
Plant DC+
Output load
Current flow
From optical isolator
Plant DC–
(c)

Figure 16.9 D.c. output circuits: (a) isolated output circuit, current sourcing; (b) the effect of an inductive load and the reason for including diode D1; (c) current sinking output

continuously with the outside world, but acts, rather, by taking 'snapshots'.

The action of *Figure 16.11*(*a*) is called a *program scan*, and the period of the loop is called the *program scan time*. This depends on the size of the PLC program and the speed of the processor, but is typically 2–5 ms per K of program. Average scan times are usually around 10–50 ms.

Figure 16.11(*a*) can be expanded to *Figure 16.11*(*b*). The PLC does NOT read inputs as needed (as implied by *Figure 16.11*(*a*)) as this would be wasteful of time. At the start of the scan it reads the state of ALL the connected inputs and stores their state in the PLC memory. When the PLC program accesses an input, it reads the input state as it was at the start of the current program scan.

As the PLC program is obeyed through the scan, it again does not change outputs instantly. An area of the PLC's memory corresponding to the outputs is changed by the program, then ALL the outputs are updated simultaneously at the end of the scan. The action is thus:

```
Read Inputs,
Scan Program,
Update Outputs.
```

The PLC memory can therefore be considered to consist of four areas as shown on *Figure 16.11*(*c*). The inputs are read into an input mimic area at the start of the scan, and the outputs updated from the output mimic area at the end of the scan. There will be an area of memory reserved for internal signals which are used by the program but are not connected directly to the outside world (timers, counters, storage bits (e.g. fault signals) and so on). These three areas are often referred to as the *data table* (Allen Bradley) or the *database* (ASEA/ABB).

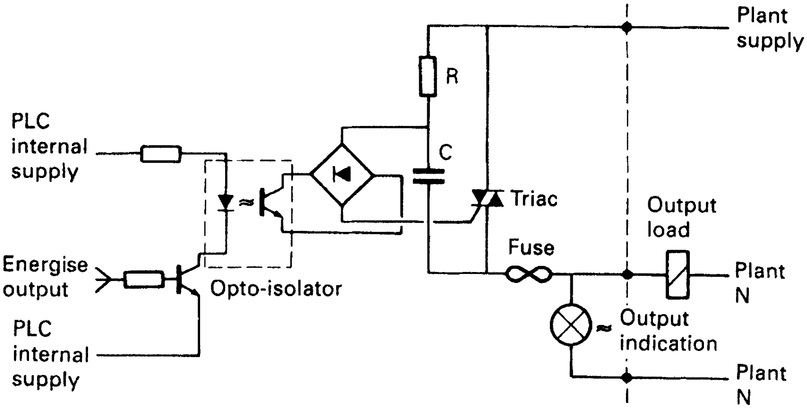


Figure 16.10 A.c. isolated output. The triac switches on at zero voltage and off at zero current which minimises interference

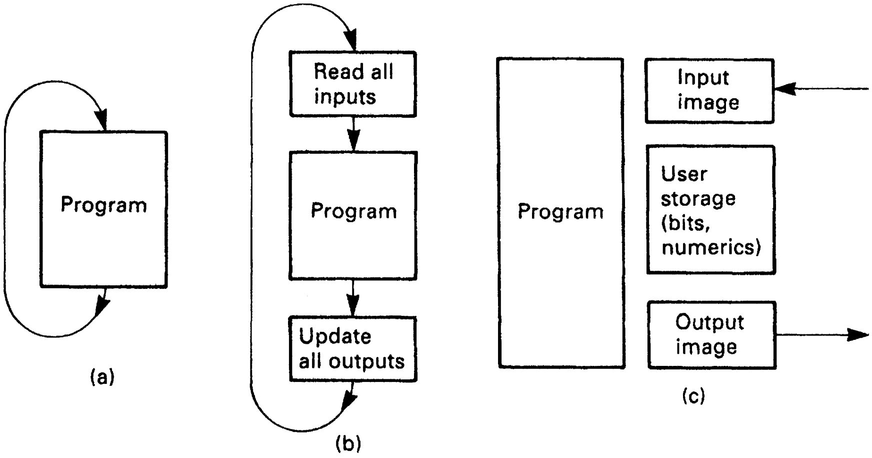


Figure 16.11 The program scan and memory organisation: (a) simple view of PLC operation; (b) more detailed view of PLC operation; (c) memory organisation

This data area is smaller than may be at first thought. A medium size PLC system will have around 1000 inputs and outputs. Stored as individual bits in a PLC with a 16 bit word this corresponds to just over 60 storage locations. An analog value read from the plant or written to the plant will take one word. Timers and counters take two words (one for the value, and one for the preset) and sixteen internal storage bits take just one word. The majority of the store therefore, is taken up by the fourth area, the program itself.

The program scan limits the speed of signals to which a PLC can respond. In *Figure 16.12*(*a*) a PLC is being used to count a series of fast pulses, with the pulse rate slower than the scan rate. The PLC counts correctly. In *Figure 16.12*(*b*) the pulse rate is faster than the scan rate and the PLC starts to miscount and miss pulses. In the extreme case of *Figure 16.12*(*c*) whole blocks of pulses are totally ignored.

In general, any input signal a PLC reads must be present for longer than the scan time; shorter pulses may be read if they happen to be present at the right time but this cannot be guaranteed. If pulse trains are being observed, the pulse frequency must be slower than $1/(2 \times \text{scan period})$. A PLC with a scan period of 40 ms can, in theory, just about follow a pulse train of $1/(2 \times 0.04) = 12.5$ Hz. In practice other factors such as filters on the input cards have a significant effect and it always advisable to be conservative in speed estimates.

Less obviously, the PLC scan can cause a random 'skew' between inputs and outputs. In *Figure 16.13* an input I is to cause an 'immediate' output O. In the best case of *Figure 16.13*(*a*), the input occurs just at the start of the scan, resulting in the energisation of the output one scan period later. In *Figure 16.13*(*b*) the input has arrived just after the inputs are read, and one whole scan is lost before the PLC 'sees' the input, and the rest of the second scan passes before the output is energised. The response can thus vary between one and two scan periods.

In the majority of applications this skew of a few tens of milliseconds is not important (it cannot be seen, for example, in the response of a plant to pushbuttons). Where fast actions are needed, however it can be crucial. If, for example, material travelling at 15 m/s is be cut to length by a PLC with the cut being triggered by a photocell a 30 ms scan time would result in a $0.03 \times 15\,000 = 450$ mm variation in cut length.

PLC manufacturers provide special cards (which are really small processors in their own right) for dealing with

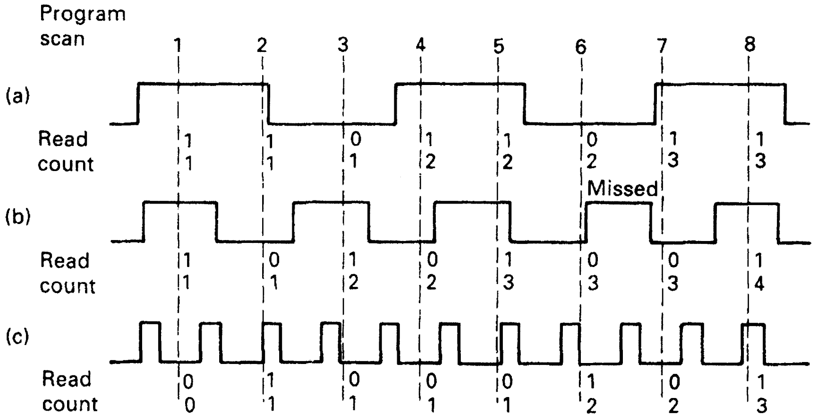
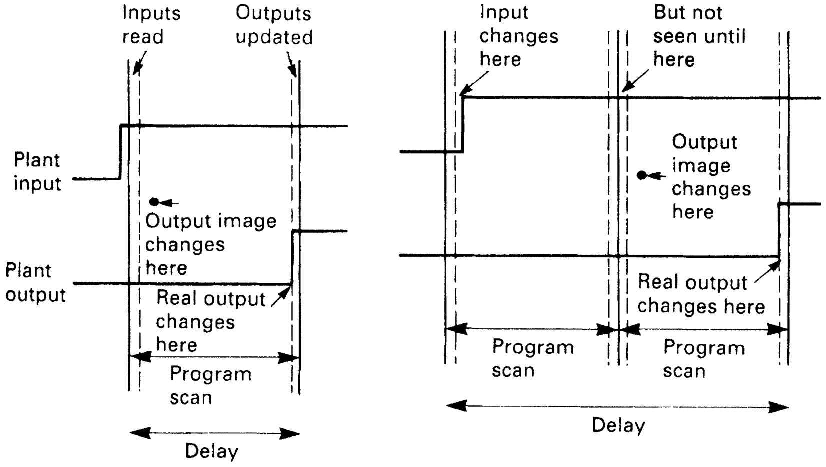

Program scan
1
2
3
4
5
6
7
8
(a)
Read count
1 1
1 1
0 1
1 2
1 2
0 2
1 3
1 3
Missed
(b)
Read count
1 1
0 1
1 2
0 2
1 3
0 3
0 3
1 4
(c)
Read count
0 0
1 1
0 1
0 1
0 1
1 2
0 2
1 3

Figure 16.12 The effect of program scan on a fast pulse train


Inputs read
Outputs updated
Plant input
Plant output
Output image changes here
Real output changes here
Program scan
Delay
Input changes here
But not seen until here
Output image changes here
Real output changes here
Program scan
Program scan
Delay

(a) (b)

Figure 16.13 The effect of program scan on response time: (a) best case; (b) worst case

this type of high speed application. We will return to these later in Section 16.4.8

The layout of the PLC program itself can result in undesirable delays if the program logic flows against the PLC program scan. The PLC starts at the first instruction for each scan, and works its way through the instructions in a sequential manner to the end of the program when it does its output update, then goes to read its inputs and run through the program again.

In *Figure 16.14*(*a*), an input I again causes an output O, but it goes through five steps first (it could be stepping a counter or seeing if some other required conditions are present). The program logic, however, is flowing against the scan. On the first scan the input I causes event A. On the next scan event A causes event B and so on until after 5 scans event D causes the output to energise. If the program had been arranged as *Figure 16.14*(*b*) the whole sequence would have occurred in one single scan.

The failings of *Figure 16.14*(*a*) are self-evident, but the effect can often occur when the layout of the program is not carefully planned. The effect can also be used deliberately to ensure sequences operate correctly.

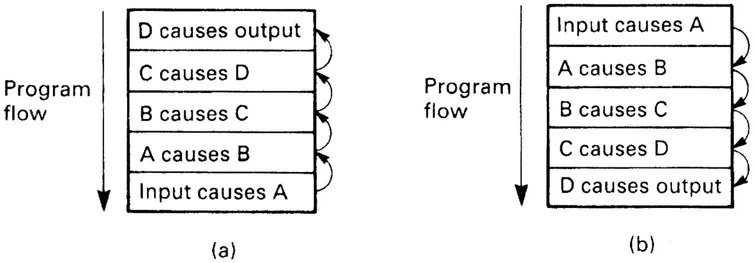
The effect of scan times can become even more complex when remote serially scanned I/O racks are present. These are generally read by an I/O scanner as *Figure 16.15* but the remote I/O scan is not usually synchronised to the program scan. In this case with, say, a program scan of 30 ms and a remote I/O scan of 50 ms the fastest response to an input could be 30 ms, but the slowest response (with an input just missing the I/O scan and the I/O scan just missing the program scan and the programming scan just missing the I/O scan to update the output) could be 180 ms.

PLC manufacturers offer many facilities to reduce the effect of scan times. Typical are intelligent high speed independent I/O cards and the ability to sectionalise the program into areas with different scan rates.

16.3 Programming methods

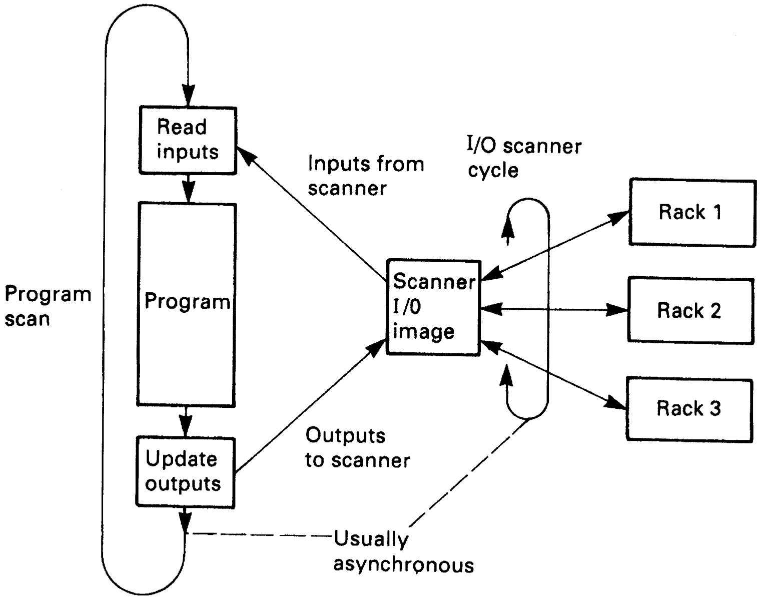
16.3.1 Introduction

The programming language of a PLC will be used by engineers, technicians and maintenance electricians. It should



Program flow
D causes output
C causes D
B causes C
A causes B
Input causes A
(a)
Program flow
Input causes A
A causes B
B causes C
C causes D
D causes output
(b)

Figure 16.14 Compounding of program scan delays: (a) logic flows against the scan, five scan times from input to output; (b) logic flows with program scan, output occurs in same program scan as input



Read inputs
Program
Update outputs
Program scan
Inputs from scanner
I/O scanner cycle
Scanner I/O image
Outputs to scanner
Rack 1
Rack 2
Rack 3
Usually asynchronous

Figure 16.15 The effect of remote input/output scan times. The remote I/O scan usually free-runs and is not synchronised with the program scan

therefore be based on techniques used in industry rather than techniques used in computer programming. In this section we shall look at the various ways of programming PLCs from different manufacturers.

16.3.2 I/O identification

The PLC program is concerned with connections to the outside plant, and these input and output devices need to be identified inside the program. Before we can examine how the program is written we will first discuss how various manufacturers treat the I/O.

The earlier *Figure 16.3* showed that a medium sized PLC system consists of several racks each containing cards, with each card interfacing generally with 8, 16 or 32 devices. I/O addressing is usually based on this rack/card/bit idea.

The Allen Bradley PLC-5 family has a range of processors which can address up to 64 racks. Its medium size 5/25 can have up to 8 racks. The rack containing the processor is automatically defined as rack 0, but the designer can allocate addresses of the other racks (in the range 1–7) by set up switches. The racks other than rack 0 connect to the processor via a remote I/O serial communications cable.

Each rack contains 16 card positions which are grouped in pairs called a 'slot'. A 16 card rack thus contains eight slots, numbered 0–7. A slot can contain one 16 way input card and one 16 way output card OR two 8 way cards usually (but not necessarily) of the same type.

The addressing for inputs is

`I:Rack Slot/Bit`

with bit being 2 digits. Allen Bradley use octal addressing for bits, so allowable numbers are 00–07 and 10–17. The address `I:27/14` is input 14 (octal remember) on slot 7 in rack 2.

Outputs are addressed in a similar manner:

`O:Rack Slot/Bit`

so `O:35/06` is output 6 in slot 5 of rack 2. Note that if 16 way cards are used an input and an output can have the same rack/slot/bit address, being distinguished only by the

I: or the O:. With 8 way cards there can be no sharing or rack/slot/bit addressing.

The digital I/O in Siemens 115 PLCs is arranged into groups of 8 bits, called a Byte. A signal is identified by its bit number (0–7) and its byte number (0–127).

Inputs are denoted

`I<byte>.<bit>,`

and outputs by

`Q<byte>.<bit>.`

`I9.4` is thus an input with bit address 4 in byte 9, and `Q63.6` is an output with bit address 6 in byte 63.

Like Allen Bradley, Siemens use card slots in one or more racks. The cards are available in 16 bit (2 byte) or 32 bit (4 byte) form. A system can be built with local racks connected via a parallel bus cable or as remote racks with a serial link.

The simplest form of addressing is fixed slot where four bytes are assigned sequentially to each slot; 0–3 to the first slot, 4–7 to the next slot and so on. Input I12.4 is thus input bit 4 on the first byte of the card in slot 3 of the first rack. If 16 bit (2 byte) cards are used with fixed (4 byte) addressing the upper 2 bytes in each slot are lost.

In all bar the simplest system the user has the ability to assign byte addresses. This is known as variable slot addressing. The first byte address and the range (2 byte for 16 bit cards or 4 byte for 32 bit cards) can be set independently for each slot by switches in the adaptor module in each rack. Although any legitimate combination can be set up, it is recommended that a logical order is used.

Siemens use different notations in different countries with multi-lingual programming terminals. A common European standard is German, where E (for Eingang or input) is used for inputs (e.g. `E4.7`) and A (for Ausgang) used for outputs (e.g. `A3.5`).

The GEM-80 again configures its I/O in terms of bits and slots within racks. The processor rack can contain 8 card positions, and additional I/O can be connected into 12 position racks local to the processor connected via ribbon cable (called Basic I/O) or remotely via a serial link.

The I/O is addressed in terms of 16 bit words, one word corresponding to one or two card positions, and the prefix A being used for inputs and B for outputs. The bit addressing runs in decimal from 0 to 15.

`A3.12` is thus input bit 12 in word 3 and

`B5.04` is output bit 4 in word 5

A word can only be an input or an output; duplication of word addresses is not allowed. I/O cards are available in 8 bit, 16 bit and 32 bit form, so one slot can be half a word, one word or two words according to the cards being used. Individual slot addresses are set by rotary switches on the back plane of each rack. The user has a more or less free choice in this allocation, but as usual it is best to use a logical sequential progression.

The ABB (originally ASEA) Master system is a more complex system than any we have discussed so far. Its organisation brings the user closer to the computer, and its language is more akin to the ideas used by programmers. If the PLCs discussed so far are taken to be represented by the home computer language BASIC, the ABB Master is analogous to PASCAL or C. This comparison is actually closer than might, at first, be thought. BASIC is quick and easy to use, but can de-generate into a web of spaghetti programming if care is not taken. PASCAL and C are more powerful but everything has to be declared and the language forces organisation and structure on the user.

The I/O cards are NOT identified by position in the rack, but by an address set on the card by a small plug with solder links. The I/O addressing does not, therefore, relate to card position, and a card can, in theory, be moved about without changing its operation.

The processor memory is arranged as *Figure 16.16*(*a*). The I/O is connected to a processor database, but unlike PLCs described earlier, the designer can specify different scan rates for different cards.

The designer also has considerable power over how the PLC program is organised. This is heavily modularised as we shall see later, and the user can also specify different scan rates for different modules of the program.

Figure 16.16(*b*) indicates the database for one input card. There are two levels of the definition, the top level relating to details of the board itself such as address and scan rate, then lower levels relating to details of each channel on the board such as its name and whether the signal is to be inverted. The database holds details for all the I/O which can then be referenced by the program either by its database identification (e.g. `DI3.1`) or by its unique name (e.g. HydPump2StartPB).

The Mitsubishi F2 range is typical of small PLCs with input/output connection, power supply and processor all contained in one unit. The smallest unit, the F2-40 M has 24 inputs and 16 outputs. (It is a characteristic of process control systems that the ratio input:outputs is generally 3:2.)

The 24 inputs are designated X400–X427 in octal notation and the 16 outputs Y430–Y447. The apparently arbitrary numbers are directly related to the storage

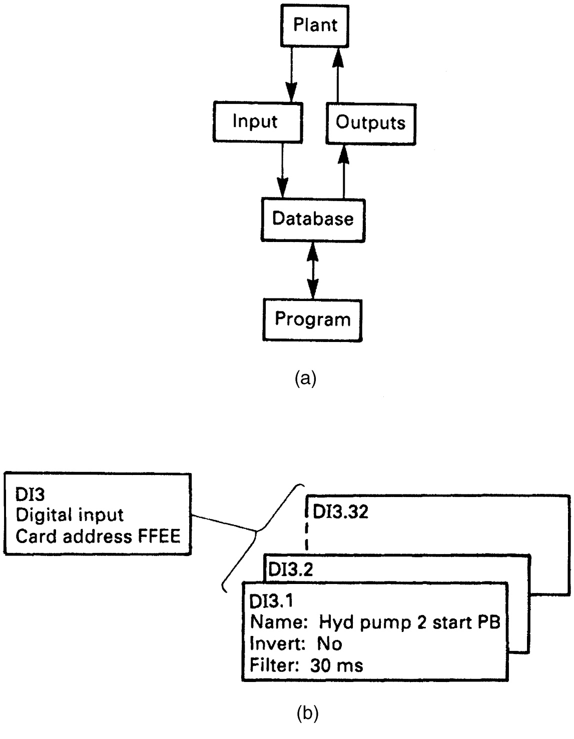


Figure 16.16 The ABB Master system: (a) organisation of the memory; (b) definition of a digital input in the database

locations used to hold the image of the inputs and output. Further addresses are used in larger PLCs in the series.

16.3.3 Ladder logic

Early PLCs, designed for the car industry, replaced relay control schemes. The symbols used in American relay drawings, `-] [-` for a normally open (NO) contact, `-] / [-` for a normally closed (NC) contact, and `-( )-` for a plant output, were the basis of the language. The earlier *Figure 16.5* showed the keyboard for a programmer for this type of PLC; the relationship to relay symbolism is obvious.

Suppose we have a hydraulic unit, and we wish to give a healthy lamp indication when

The Pump is running (sensed by an auxiliary contact on the pump starter).

There is oil in the tank (sensed by a level switch which makes for good level).

There is oil pressure (sensed by a pressure switch which makes for adequate pressure).

With conventional relays, we would wire up a circuit as *Figure 16.17*(*a*).

To use a PLC, we connect the input signals to an input card, and the lamp to an output card as *Figure 16.17*(*b*). The I/O notation used is Allen Bradley.

The program to provide the function is shown on *Figure 16.17*(*c*). The line on the left can be considered to be a supply, and the line on the right a neutral. The output is represented by a coil `-( )-` and is energised when there is a route from the left-hand rail. Output `0:22/01` will come on when signals `I:21/00`, `I:21/01` and `I:21/02` are all present.

The program is entered from a terminal with keys representing the various relay symbols. The terminal can also be used to monitor the state of the inputs and outputs, with 'energised' inputs and outputs being shown highlighted on the screen.

In *Figure 16.18*(*a*), a hydraulic cylinder can be extended or retracted by operation of two pushbuttons. The notation this time is for a GEM-80. It is undesirable to allow both solenoids to be operated together; this will almost certainly result in blown fuses in the supply to the output card, so some protection is needed. The program to achieve this is shown on *Figure 16.18*(*b*).

Normally closed contacts `-]/[-` have been used here. Output `B2.9`, the extend solenoid, will be energised when the extend pushbutton is pressed, providing the retract solenoid is not energised or the retract button pressed, and the extend limit switch has not been struck.

There are two points to note on *Figure 16.18*. Contacts can be used from outputs as well as inputs, and contacts can be used as many times as needed in the program. *Figure 16.18* also shows the origin of the name '*Ladder Program*'. A program in this form looks like a ladder, with each instruction statement forming a '*rung*' and the power rail and neutral the supports. The term 'rung' is invariably applied to the contacts leading to one output.

Let us return to the hydraulics healthy light of *Figure 16.17* and add a lamp test pushbutton (a useful feature that should be present on all panels. It not only allows lamps to be tested, but can also be used to check the PLC and the local rack are healthy). To do this we add the lamp test pushbutton to the PLC and modify the program to *Figure 16.19*.

Here we have added a branch, and the output will energise if our three plant signals are all present OR the

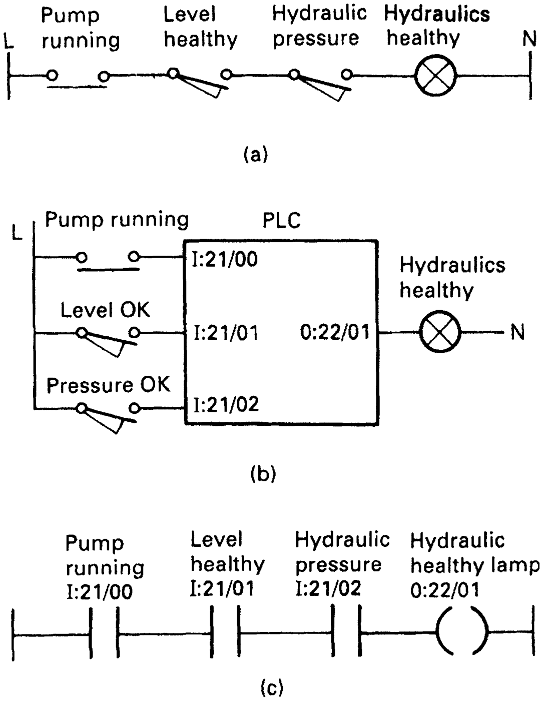

L
Pump running
Level healthy
Hydraulic pressure
Hydraulics healthy
N
(a)
L
Pump running
PLC
I:21/00
Hydraulics healthy
Level OK
I:21/01
0:22/01
N
Pressure OK
I:21/02
(b)
Pump running I:21/00
Level healthy I:21/01
Hydraulic pressure I:21/02
Hydraulic healthy lamp 0:22/01
(c)

Figure 16.17 From a relay circuit to a PLC program: (a) basic non PLC circuit; (b) wiring of I/O to a PLC; (c) the corresponding PLC program

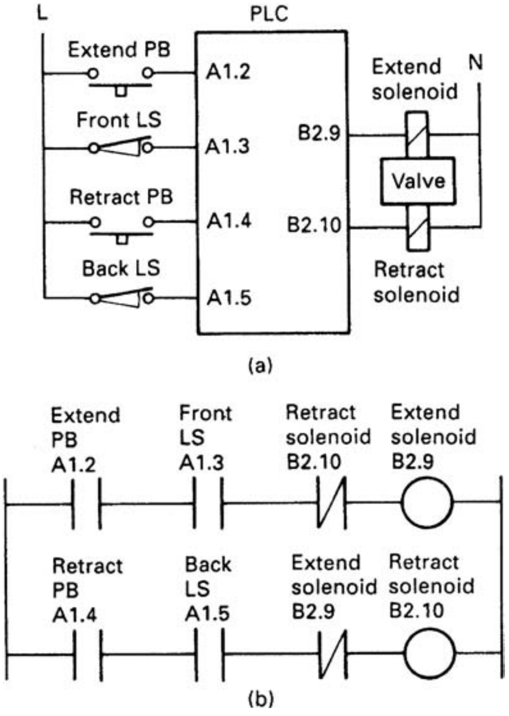

L
PLC
Extend PB
A1.2
Extend solenoid
N
Front LS
A1.3
B2.9
Valve
Retract PB
A1.4
B2.10
Back LS
A1.5
Retract solenoid
(a)
Extend PB A1.2
Front LS A1.3
Retract solenoid B2.10
Extend solenoid B2.9
Retract PB A1.4
Back LS A1.5
Extend solenoid B2.9
Retract solenoid B2.10
(b)

Figure 16.18 Ladder diagram in GEM-80 notation: (a) input/output connections; (b) GEM-80 ladder diagram

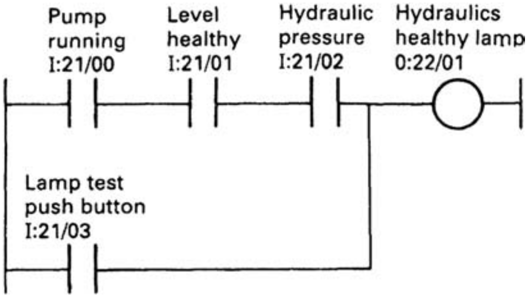

Pump running I:21/00
Level healthy I:21/01
Hydraulic pressure I:21/02
Hydraulics healthy lamp 0:22/01
Lamp test push button I:21/03

Figure 16.19 Adding a lamp test pushbutton with a branch

lamp test button is pressed. The way in which the branch is programmed need not concern us here as it varies between manufacturers. Some use start branch and end branch keys (the keypad shown earlier on *Figure 16.5* uses this method, the corresponding keys can readily be identified). Others use a branch from/to approach. All are simple to use.

A further use of a branch is shown on *Figure 16.20*. This is probably the commonest control circuit, a motor starter, shown using Siemens notation. The operation is simple, pressing the start pushbutton causes the output `Q8.2` to energise, and the contact of the output in the branch keeps the output energised until the stop button is pressed. The program, like its relay equivalent, remembers which button was last pressed.

There is, however, a very important point to note about the pushbutton wiring and the program. For safety, a normally closed stop button has been used giving an input signal on `I12.5` when the stop button is NOT pressed. A loss of supply to the button, or a cable fault, or dirt under the contacts will cause the signal to be lost making the program think the stop PB has been pressed causing the motor to stop. If a normally open stop PB has been used, the PLC program could easily be made to work, but a fault with the stop button or its circuit could leave the motor running with the only way of stopping it being to turn off the PLC or the motor supply.

This topic is discussed further in Section 16.7.4, but note the effect on the program in *Figure 16.20*. The sense of the stop button input (`I12.5`) inside the program is the opposite of what would be expected in a relay circuit. The input is really acting as '*Permit to Run*' rather than '*Stop*'.

16.3.4 Logic symbols

Logic gates are widely used in digital systems (including the boards used inside PLCs). The circuits on these boards are represented by logic symbols, and these symbols can also be used to represent the operations of a PLC program. Logic symbols are used by Siemens and ABB; initially we will use Siemens notation.

The output from an AND gate, shown on *Figure 16.21*(*a*), is TRUE if (and only if) all its inputs are TRUE. The operation of the gate of *Figure 16.21*(*a*) can be represented by the table of *Figure 16.21*(*b*). In *Figure 16.21*(*c*) we have the hydraulics healthy lamp of *Figure 16.19* programmed using logic symbols for a Siemens PLC. The output block, denoted by equals =⇐is energised when its input is true, so the lamp `Q8.2` is energised (lit) when all the inputs to the AND gate are true.

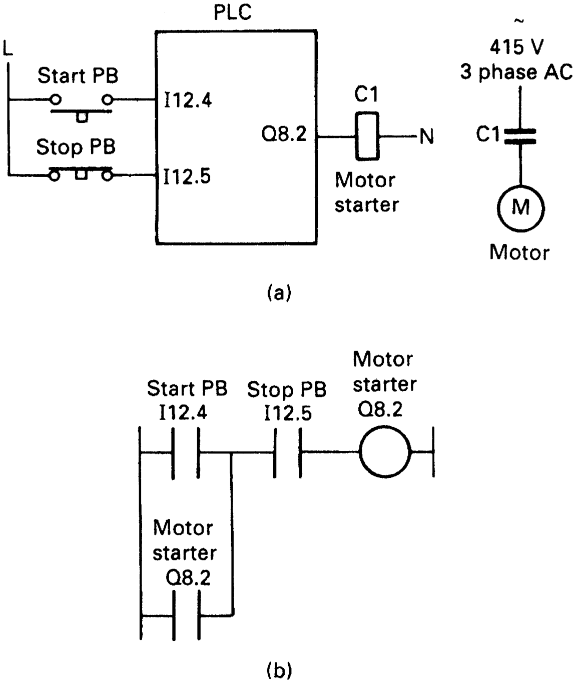
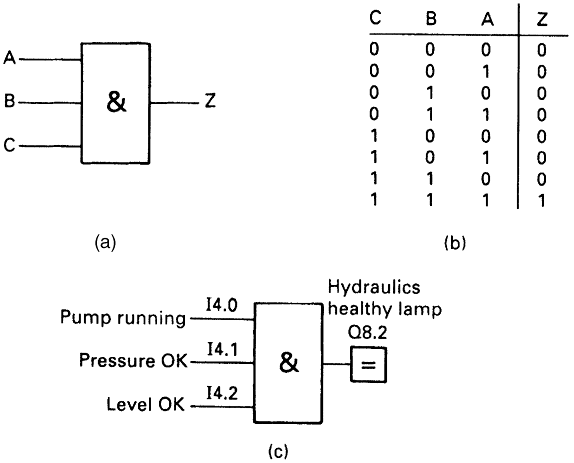


Figure 16.20 A simple motor starter in Siemens notation: (a) input/output connections; (b) the ladder diagram. Note how the stop button appears in the program

Often a test has to be made to say a signal is NOT true. This is denoted by a small circle 'o'. In the earlier *Figure 16.18* we illustrated the control of a hydraulic cylinder with a program which prevented the extend and retract solenoids from being energised simultaneously. This is shown programmed with logic symbols for a Siemens PLC in *Figure 16.22*. Note the NOT inputs on each AND gate.

The output of an OR gate, Z in *Figure 16.23*(*a*), is TRUE if any of its inputs are TRUE. The inverse of a signal can be tested, as before, with a small circle 'o'. The output Z of the



C	B	A	Z
0	0	0	0
0	0	1	0
0	1	0	0
0	1	1	0
1	0	0	0
1	0	1	0
1	1	0	0
1	1	1	1

Figure 16.21 PLC programming using logic symbols: (a) an AND gate; (b) truth table for a three input AND gate; (c) the healthy lamp of *Figure 16.17* using a logic symbol in Siemens notation

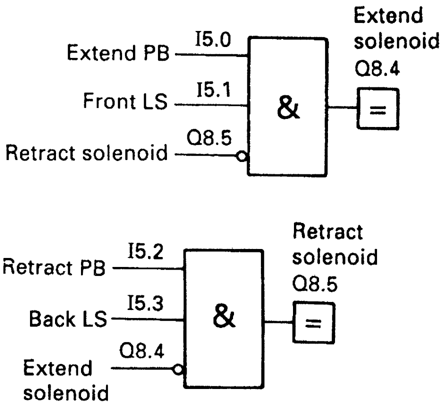

Extend PB
I5.0
Front LS
I5.1
Retract solenoid
Q8.5
&
Extend solenoid Q8.4
=
Retract PB
I5.2
Back LS
I5.3
Extend solenoid
Q8.4
&
Retract solenoid Q8.5
=

Figure 16.22 The hydraulic cylinder of *Figure 16.18* in logic notation and Siemens addressing. Note the use of inverted inputs (denoted by small circles)

gate in *Figure 16.23*(*b*) is TRUE if A is TRUE or B is FALSE or C is TRUE. In *Figure 16.23*(*c*) we have used an OR gate to add a lamp test to our hydraulic healthy lamp.

The circuit of *Figure 16.23*(*c*) is an AND/OR combination. The ABB Master has logic combination blocks as well as the basic gates. *Figure 16.24*(*a*) is the Master block corresponding to *Figure 16.23*(*c*) (with a Master program referring to the names in its database). Similarly, for an OR/AND combination the OR/AND block of *Figure 16.24*(*b*) can be used in a Master program.

16.3.5 Statement list

A statement list is a set of instructions which superficially resemble assembly language instructions for a computer. Statement lists, available on the Siemens and Mitsubishi range, are the most flexible form of programming for the experienced user but are by no means as easy to follow as ladder diagrams or logic symbols.

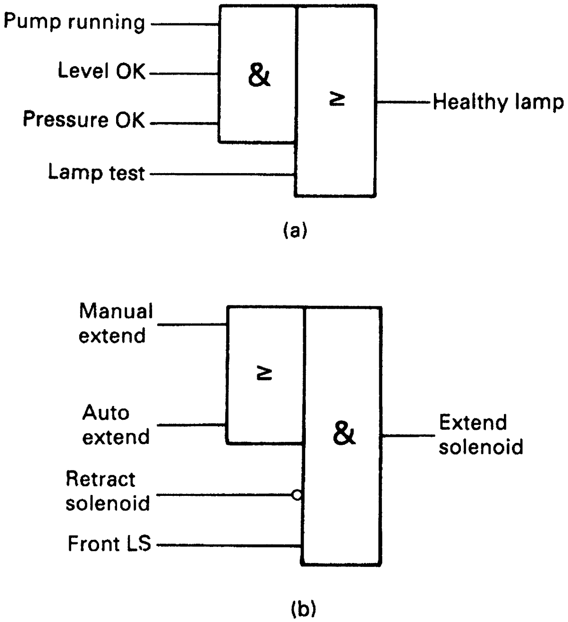

Pump running
Level OK
Pressure OK
Lamp test
&
≥
Healthy lamp
(a)
Manual extend
Auto extend
Retract solenoid
Front LS
≥
&
Extend solenoid
(b)

Figure 16.24 ABB Master composite gates: (a) AND/OR gate (equivalent to *Figure 16.23*(*c*)); (b) OR/AND gate

Figure 16.25 shows a simple operation in both ladder and logic formats for a Siemens PLC. The equivalent statement list would be:

Instruction	*Operation*	*Address number*	*Comment*
00	:A	I 3.7	Forward Pushbutton
01	:A	I 3.2	Front Limit OK
02	:AN	Q 4.2	Reverse Solenoid
03	:=⇐	Q 4.11	Output to Forward Solenoid

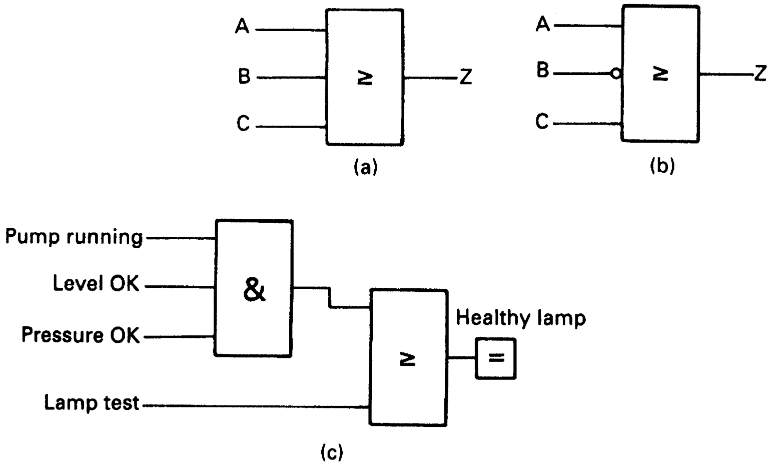

A
B
C
≥
Z
(a)
A
B
C
≥
Z
(b)
Pump running
Level OK
Pressure OK
Lamp test
&
≥
Healthy lamp
=
(c)

Figure 16.23 The OR Gate: (a) logic symbol; (b) OR gate with inverted input; (c) lamp test added to *Figure 16.21*(*c*)

```
  I3.7     I3.2     Q4.2     Q4.11
I--] [----] [----]/[----( )--I
```

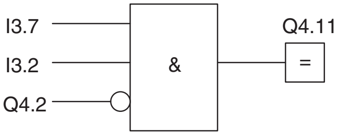


Figure 16.25 Equivalent ladder and logic statements in Siemens notation

```
  I2.7        Q4.4
+--]/[--+----( )--I
|       |
| F3.6  |
+--] [--+
|       |
| Q4.2  |
+--]/[--+
```

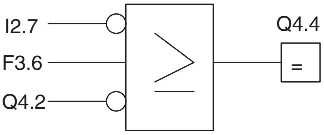


Figure 16.26 OR gate equivalence in Siemens notation

Here :A denotes AND, :AN denotes AND-NOT and : =⇐sends the result to the output address Q4.11.

An OR operation is shown on *Figure 16.26*. The equivalent statement list is:

Instruction	*Operation*	*Address number*	*Comment*
00	:ON	I 2.7	Local Pump Running
01	:O	F 3.6	Remote Pump Running
02	:ON	Q 4.2	Local Pump Starter
03	: =⇐	Q 4.4	Pump Healthy Lamp

where ON denotes OR-NOT and O denotes OR.

Where a set of statements can be anomalous, brackets can be used to define the operation precisely. This is similar to the use of brackets in conventional programming where the sequence 3 + 5/2 can be written as (3 + 5)/2 = 4⇐ or 3 + ⇐(5/2) = 5⇐.5.

Although the latter is the default assumed by a program, the brackets do make the operation clear to the reader.

Figure 16.27 shows a typical operation, as usual in both logic and ladder diagram format. The equivalent statement list is:

Instruction	*Operation*	*Address*	*Comments*
00	:A(		Open First Set of Brackets
01	:O	F 3.3	Forward from desk 1
02	:O	F 3.4	Forward from desk 2
03	:)		Result of first set of brackets
04	:A(		AND Result with second set of brackets
05	:A	I 2.0	Motor 1 Selected
06	:A	I 2.1	Motor 2 Selected
07	:)		Now at point X
08	:A	I 4.1	Front Limit Switch Healthy
09	:AN	Q 5.5	Reverse Starter
10	: =⇐	Q 5.6	Output to Forward Starter

Computer programmers will recognise this as being similar to the operation of a stack with the brackets pushing data down, or lifting data up, the stack.

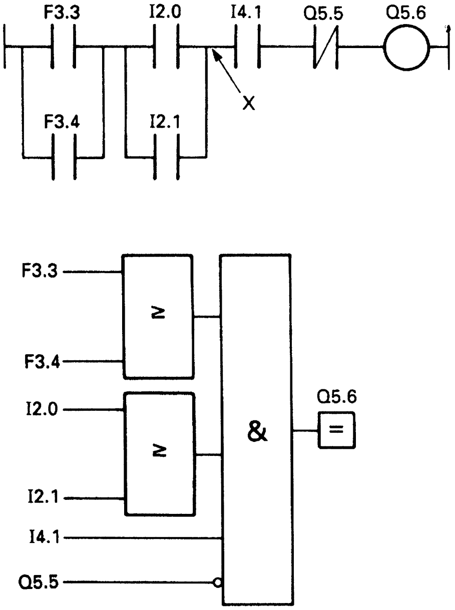


Figure 16.27 More complex statements in ladder and logic notations

The Mitsubishi PLC also uses statement lists, although the manual recommends the designer to construct a ladder diagram first then translate it into a statement list. The PLC system shown in *Figure 16.28* with Mitsubishi notation becomes the statement list:

Instruction	*Operation*	*Address*	*Comments*
0	LD	X401	LD starts rung or branch
1	AND	X402	Xnnn are inputs
2	ANI	X403	ANI is And-Not
3	LD	Y430	LD starts a new branch leg
4	AN	M100	Mnnn are internal storage
5	ORB		OR the two branch legs
6	AND	M101	
7	OUT	Y430	End of Rung

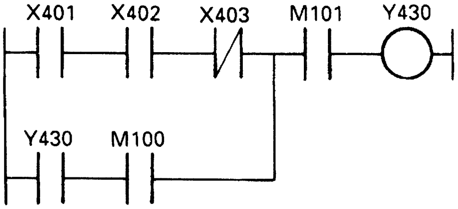


Figure 16.28 A rung in a Mitsubishi ladder program

16.3.6 Bit storage

As well as inputs and outputs, the PLC will need to hold internal signals for data such as 'Standby Pump Running', 'System Healthy', 'Lubrication Fault' and so on. It would be very wasteful to allocate real outputs to these signals, so all PLCs provide some form of internal bit storage. These are known variously as Auxiliary Relays, (Mitsubishi), Flags (Siemens), General Workspace (GEM-80) and Bit Storage (Allen Bradley). The notation used within the programs vary, of course, from manufacturer to manufacturer.

Mitsubishi use Mnnn with nnn representing numbers within the predefined area M100 to M377 octal. Like most small PLCs the memory layout is fixed and cannot be defined by the user. In the other, larger, PLCs we discuss, the user can define how many storage bits are needed.

The Siemens notation is F <Byte>.<Bit>ς(e.g. `F27.06`).

The GEM-80 has a variety of general work space. The commonest is called the G table, and appears in programs as G<Word>.<Bit>ς (e.g. `G52.14`). The G table is cleared when the PLC goes from a stopped state to a run state. Storage in the R table (e.g. `R12.03`) retains its state with the processor halted or with power removed.

Bit storage in the PLC-5 is denoted by B3/*n* where *n* denotes the signal (e.g. `B3/192`). The B denotes bit storage and the 3 is mandatory and arises out of the way the PLC-5 holds data in files. Bit storage is file 3; timers are file 4 (T4) and counters file 5 (C5) as we shall see later.

The ABB Master programming language does not really require internal storage bits, the function being provided by elements and connections within its database and the programming language.

Some form of memory circuit is needed in practically every PLC program. Typical examples are catching a fleeting alarm and the motor starter of the earlier *Figure 16.20* where the rung remembers which button (start or stop) has been last pressed. These are known, for obvious reasons, as storage circuits.

The commonest form is shown in ladder and logic form in *Figure 16.29*(*a*). Here output C is energised when input A is energised, and stays energised until input B is de-energised.

The operation is summarised on *Figure 16.29*(*b*). As can be seen input B overrides input A, the action required of a start/stop circuit. In some circuits, however, the start is required to override the stop. We all have a typical example in our motor cars; the windscreen wipers run when we switch them on, but continue to run to the park position when we turn them off. The PLC equivalent is *Figure 16.29*(*c*), where A would be the run switch, B the park limit switch and C the wiper motor. B has again been shown energised to allow running. The operation is summarised on *Figure 16.29*(*d*).

Storage is provided in digital systems by a device called a flip flop shown on *Figure 16.30*(*a*). This has two inputs, S (for Set) and R (for Reset). The device remembers which input was last energised. If both inputs occur together, the top (S) input wins. Such a circuit is called an SR flip flop. If the device is drawn with the R input at the top, as *Figure 16.30*(*b*), the reset input will override the set input if both are present together.

The flip flop is used in logic symbol PLC programming. A motor starter using a Siemens PLC is shown in *Figure 16.31*. Note that the RS version has been used to ensure the stop logic overrides the run logic, and the stop signal acts as a permit to run.

The ABB Master uses an almost identical symbol for the flip flop, with the addition that there are five versions. The first of these is the simple SR type shown earlier in *Figure 16.30*. The other versions are based on the fact that flip flops are invariably preceded by AND/OR combination of which *Figure 16.31* is typical. The additional flip flops are one unit blocks consisting of a flip flop with built in AND/OR gates of user defined size. *Figure 16.32* for example, is an ABB SRAO with an AND gate on the set input and an OR gate on the reset inputs. Other units are SRAA (AND/AND), SROA and SROO.

In Allen Bradley ladder diagrams, program clarity can be improved by the use of latch and unlatch outputs shown on *Figure 16.33*(*a*). These work on the same bit, setting the bit when the latch `-(L)-` is energised and resetting the bit when the `-(U)-` is energised. When both latch and unlatch are de-energised the bit holds its last state.

The Mitsubishi F2 uses a similar idea, but calls them S and R outputs as *Figure 16.33*(*b*). This would be coded into a statement list:

```
0       LD      X400
1       OR      X401
2       S       Y432        Set Output
3       LDI     X402
4       ORI     X403
5       R       Y432        Reset Output
```

With both the Allen Bradley latch/unlatch, and the Mitsibushi set/reset, the priority goes to which ever is last in the program because of the program scan. Both the examples of *Figure 16.33* correctly give priority to the stop signals.

Power failure or halting of the PLC can cause a problem with storage. When the PLC restarts should a memory bit hold the state it was in before the PLC halted, or should the memory be cleared? This is always a question of safety and convenience. A water pump in a pump house by a river 5 km from the main site should probably be allowed to restart itself if it was running before the power fail, an automatic stamping machine should almost certainly not restart.

The PLC manufacturers therefore allow the designer to choose whether a storage bit holds its state after a power fail (called *retentive memory*) or is cleared when the PLC is first run (called *non retentive memory*).

In the Allen Bradley PLC-5, this is determined by the circuit; the simple coil of *Figure 16.29* is non retentive, the latch/unlatch of *Figure 16.33*(*a*) is retentive.

Other PLCs use the bit address. On a Siemens 115, flag addresses `F0.0-F127.7` can be made retentive. On the Mitsubishi PLC, auxiliary relays `M100-277` are non retentive, and `M300-M377` are retentive. In the GEM-80, the general bit storage G Table is non retentive, a similar R Table is retentive, so a circuit similar to *Figure 16.29* constructed with R3.4 as the coil and retaining contact would hold its state after a power failure.

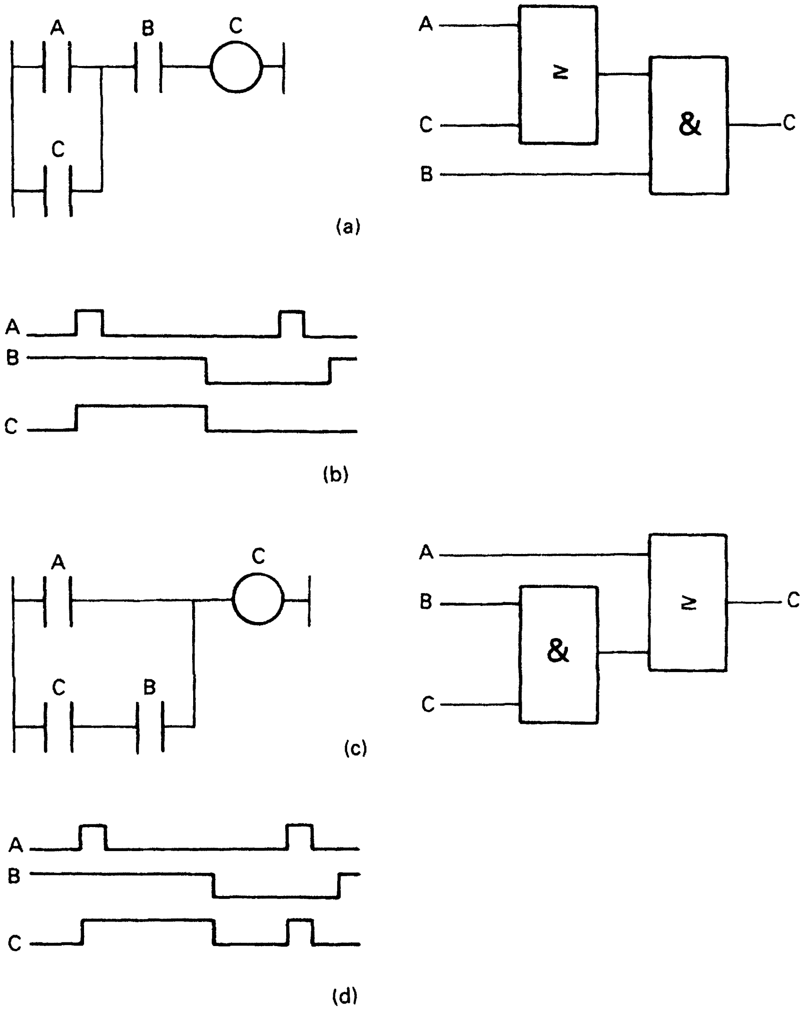

A
B
C
C
(a)
A
C
B
≥
&
C
A
B
C
(b)
A
C
C
B
(c)
A
B
C
&
≥
C
A
B
C
(d)

Figure 16.29 Bit storage programs: (a) commonest storage program, stop B overrides start A; (b) operation of (a), (c) Program where start A overrides stop B; (d) operation of (c)

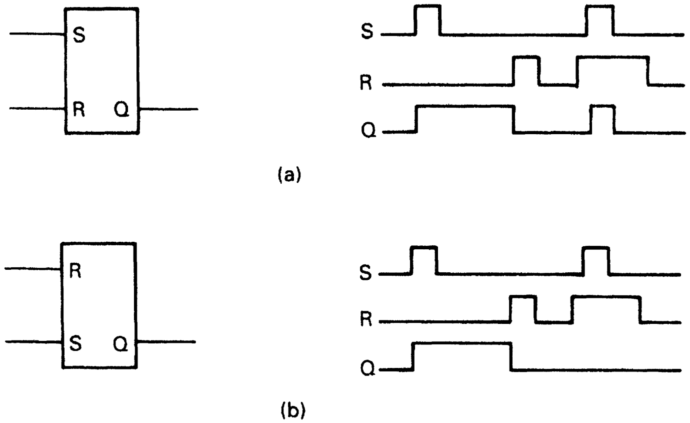

S
R
Q
S
R
Q
(a)
R
S
Q
S
R
Q
(b)

Figure 16.30 The two types of flip flop storage: (a) the SR flip flop, Set overrides R; (b) the RS flip flop, reset overrides set

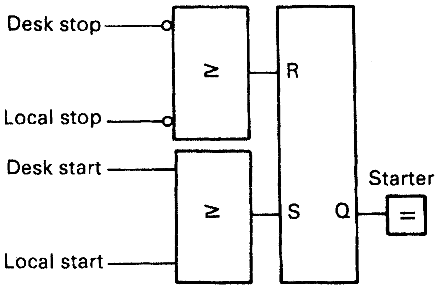


Figure 16.31 Flip flop storage is commonly preceded by logic gates. Here either stop button will reset the flip flop. Note the circles on the stop button inputs denoting inverted inputs. These are necessary because the stop buttons give a signal in the not pressed state

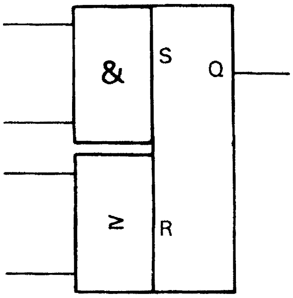


Figure 16.32 An ABB master SRAO composite flip flop

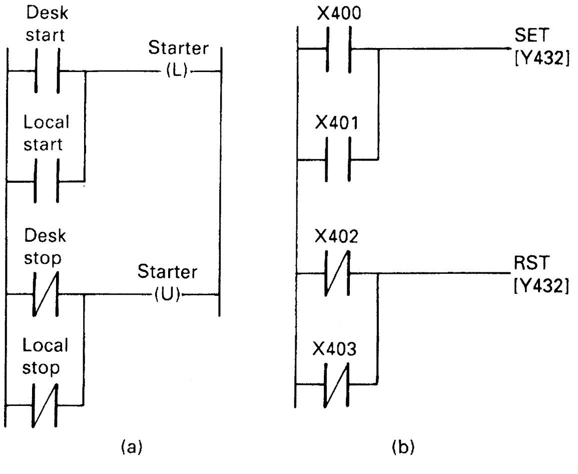


Figure 16.33 Other forms of storage: (a) the Allen Bradley latch/unlatch; (b) the Mitsubishi set/reset

The ABB Master uses a very structured PLC language, and forces a disciplined style on the programmer. The nature of sub elements such as memories and their behaviour when the PLC is first run is defined when the program elements are first declared.

Retentive storage can be very hazardous as plants can unexpectedly leap into life after a power fail. The designer should take care that the design does not accidentally introduce retentive features by an inadvertent selection of bit addresses.

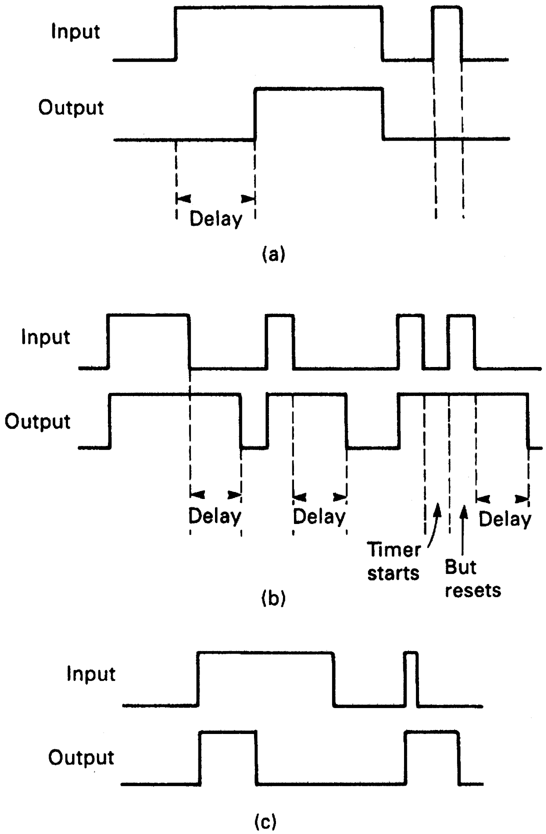


Figure 16.34 Different forms of timer: (a) the on-delay. This is the commonest timer and is often the only type available in many smaller PLCs; (b) the off-delay; (c) the fixed width pulse, often called a monostable

16.3.7 Timers

Time is nearly always a part of a control system. A typical example is: '*Lift Parking Brake, wait 0.5 seconds for brake to lift, drive to forward limit and stop drive, wait 1 second and apply parking brake*'. A PLC system must therefore include timers as part of its programming language. There are many types of timer, some of which are shown on *Figure 16.34*

By far the commonest is the on-delay of *Figure 16.34*(*a*). All the other timer blocks can be built with this block and a bit of thought. A 0 to 1 transition is delayed for a preset time T, but a 1 to 0 transition is not delayed at all. An input signal shorter then T is ignored. The GEM-80 has only this type of timer, calling it a *delay*.

The off-delay of *Figure 16.34*(*b*) passes a 0 to 1 transition instantly but delays the 1 to 0 transition. A common use of the off-delay is to remove contact bounce or noise from an input signal. An off-delay can be obtained from an on-delay by using the inverse of the input signal and taking the inverse of the timer output signal (although the resulting program lacks some clarity).

Figure 16.34(*c*) is an edge triggered pulse timer, this gives a fixed width pulse for every 0–1 transition at the timer input. The PLC-5 has a Onescan pulse timer which produces a pulse lasting one (and only one) program scan. Pulses are useful for resetting counters or gating some information from one location to another.

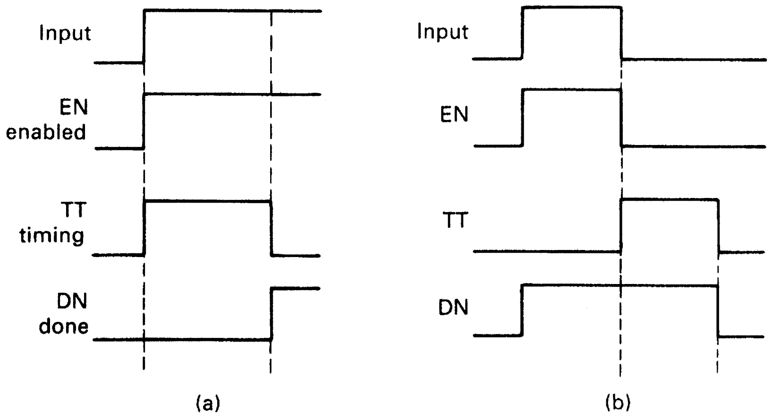


Figure 16.35 Allen Bradley timer notations: (a) EN, TT and DN for an on-delay (TON) timer; (b) EN, TT and DN for an off-delay (TOF) timer

A timer of whatever type has some values that need to be set by the user. The first of these is the basic unit of time (i.e. what units the time is measured in). Common units are 10 ms, 100 ms, 1 s, 10 s, and 100 s. The base unit does not affect the accuracy of the timer; normally the accuracy is similar to the program scan.

Next the timer duration (often called the *preset*) is defined. This is normally set in terms of the time base; a timer with a preset of 15 and a time base of 100 ms will last 1.5 s for example. In small PLCs this preset can only be set by the programmer, in the larger PLCs the duration can be changed from within the program itself. A delay off timer used to apply a parking brake, for example, could have different preset times dependent on whether the drive concerned is travelling at low speed or high speed.

When a timer is used there are several signals that may be available. *Figure 16.35* shows the signals given for a PLC-5 delay on timer (called a TON) and a delay off timer (called a TOF).

EN (for enable) is a mimic of the timer input.
TT (for timer timing) is energised whilst the time is running.
DN (for done) says the timer has finished.

In larger PLCs the elapsed time (often called the *Accumulated Time*) may be accessed by the program for use elsewhere (a program may be required to record how long a certain operation takes).

PLC manufacturers differ on how a timer is programmed. Some, such as the GEM-80, treat the timer as a delay block similar to the earlier *Figure 16.34*(*a*) with the preset being stored in a VALUE block.

Siemens use a similar idea, but have different types of timer. The PLC-5, however, uses the timer as a terminator for a rung, with the timer signals being available as contacts for use elsewhere.

Figure 16.36 shows a typical application programmed for a PLC-5 and a GEM-80 in ladder logic and a Siemens 115-U using logic symbols. The program controls a motor starter which is started and stopped via push buttons. The motor starter has an auxiliary contact which makes when the starter is energised, effectively saying the motor is running. If the drive trips because of an overload, or because an emergency stop is pressed, or there is a supply fault, the auxiliary contact signal will be lost. The contact cannot, however, be checked until 0.5 s after the starter has been energised to allow time for the contact to pull in. The program in each case checks the auxiliary contact and signals a drive fault if there is a problem. Note the difference in the way the timer is used and the fault signal is stored.

The accumulated time in the timers discussed so far goes back to zero each time the input goes to a zero. This is known as a *non retentive* timer. Most PLC timers are of this form. Occasionally it is useful to have a timer which holds its current value even though the input signal has gone. When the input occurs again the timer continues from where it stopped. This, not surprisingly, is known as a *retentive* timer. A separate signal must be used to reset the timer to zero. If a retentive timer is not available on a particular PLC, the same function can be provided with a counter, a topic discussed in the next section.

A typical timer can count up to 32 767 base time units (corresponding to 15 binary bits). Some older PLCs working in BCD can only count to 999. With a one second time base the maximum time will be just over 546 minutes or about 9 hours. Where longer times are needed, (or times with a resolution better than one second) timers and counters can be used together as described in the next section.

16.3.8 Counters

Counting is a fundamental part of many PLC programs. The PLC may be required to count the number of items in a batch, or record the number of times some event occurs. With large motors, for example, the number of starts have to be logged. Not surprisingly all PLCs include some form of counting element.

A counter can be represented by *Figure 16.37*(*a*), although not all PLCs will have all the facilities we will describe. There will be two numbers associated with the counter. The first is the count itself (often called the *accumulated value*) which will be incremented when a 0->1 transition is applied to the count up input, or decremented when a 0->1 transition is applied to the count down input. The accumulated value (i.e. the count) can be reset to zero by applying a 1 to the reset input. Like the elapsed time in a timer, the value of the count can be read and used by other parts of the program.

The second number is the *preset* which can be considered as the target for the counter. If the count value reaches the preset value, a *count complete* or *count done* signal is given. The preset can be changed by the program, a batching sequence, for example, may require the operator to change the number of items in a batch by a keypad or VDU entry.

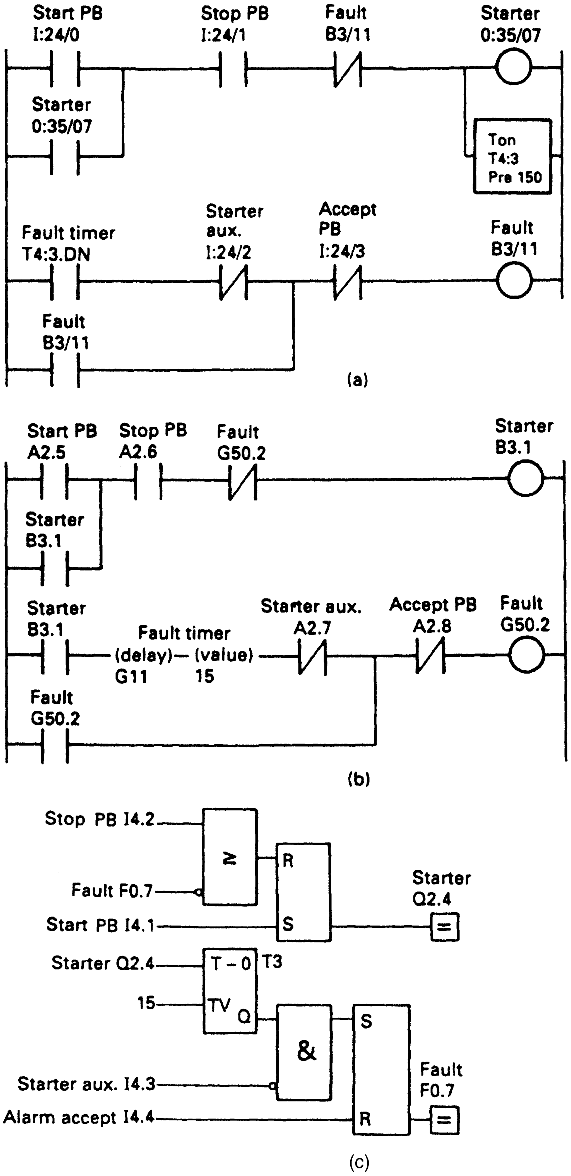


Figure 16.36 The same timer based application programmed on three different machines: (a) Allen Bradley PLC-5 TON Timer; (b) GEM-80 delay block; (c) Siemens S5 in logic notation

Similarly a signal *zero count* is sometimes available. The operation can be summarised as *Figure 16.37*(*b*).

PLC manufacturers handle counters, like timers in slightly different ways. The PLC-5 and the Mitsubishi use count up (CTU) count down (CTD) and reset (RES) as rung terminators with the count done signal (e.g. `C5:4.DN`) available for use as a contact.

The Siemens S5, ABB Master and the GEM-80 treat a counter as an intermediate block in a logic diagram or rung from which the required output signals can be used.

Figure 16.38 shows a simple count application performed by a PLC-5, a Siemens S5 and a GEM-80. Items passing along a conveyor are detected by a photocell and counted. When a batch is complete, the conveyor is stopped and a batch complete light is lit for the operator to remove the batch. When he does this, a restart button sets the sequence running again.

As we saw with timers, most PLCs allow a counter to count up to 32 767. Where larger counts are needed, counters can be cascaded with the complete (or *done*) signal from the first counter being used to step the second counter and reset the first. *Figure 16.39* is a variation on the same idea used to give a very long timer. It is shown for a PLC-5, but the same idea could be used on any PLC.

The first rung generates a free running one scan pulse with inter pulse period set by the timer period. (When the timer has not timed out, the DN signal is not present and the timer is running. When it reaches the preset, the DN signal occurs, resetting and restarting the timer.) The resulting one second pulse is counted by successive counters to give accumulated seconds/minutes/hours/days. As each counter reaches its preset it steps the next counter and resets itself.

Long duration timers built from counters are normally retentive (i.e. they hold their value when the controlling event is not present). They can be made non retentive by resetting the counters when the controlling event is not present, but this is rarely required.

16.3.9 Combinational logic

Any control system based on digital signals can be represented by *Figure 16.40*(*a*), where a system has a set of outputs Z, Y, X, W, etc. whose state is determined by inputs A, B, C, D, etc. The control scheme can operate in a combination of two basic manners.

The simplest of these is *combinational logic* where the scheme can be broken down into smaller blocks as *Figure 16.40*(*b*) with one output per block, with each output state being determined solely by the corresponding input states. The loading valve for a hydraulic pump, for example, is to be energised when

```
The pump is running
AND(Raise is selected AND top limit SW
is not struck)
OR (Lower is selected AND bottom limit
SW is not struck)
```

The operation of this loading valve can be implemented with the simple ladder or logic program of *Figure 16.41*, but it is worth developing a standard way of producing a combinational logic program.

The first stage is to break the control system down into a series of small blocks, each with one output and several inputs. For each output we now draw up a so called *truth table* in which we record all the possible input states and the required output state. In *Figure 16.42*(*a*) we have an output Z controlled by four inputs A, B, C, D. There are sixteen possible input states, and Z is energised for four of these. This can be translated directly into the ladder diagram of *Figure 16.42*(*b*) or the logic circuit of *Figure 16.42*(*c*), with each rung branch and or gate corresponding to one row in the truth table. The use of a truth table method for the design of combinational logic circuits leads directly to an AND/OR arrangement called, technically, a *Sum of Products* (S of P) circuit.

16.3.10 Event driven logic and SFCs

The states of outputs in combinational logic are determined solely by the input signals. In event driven logic (also known

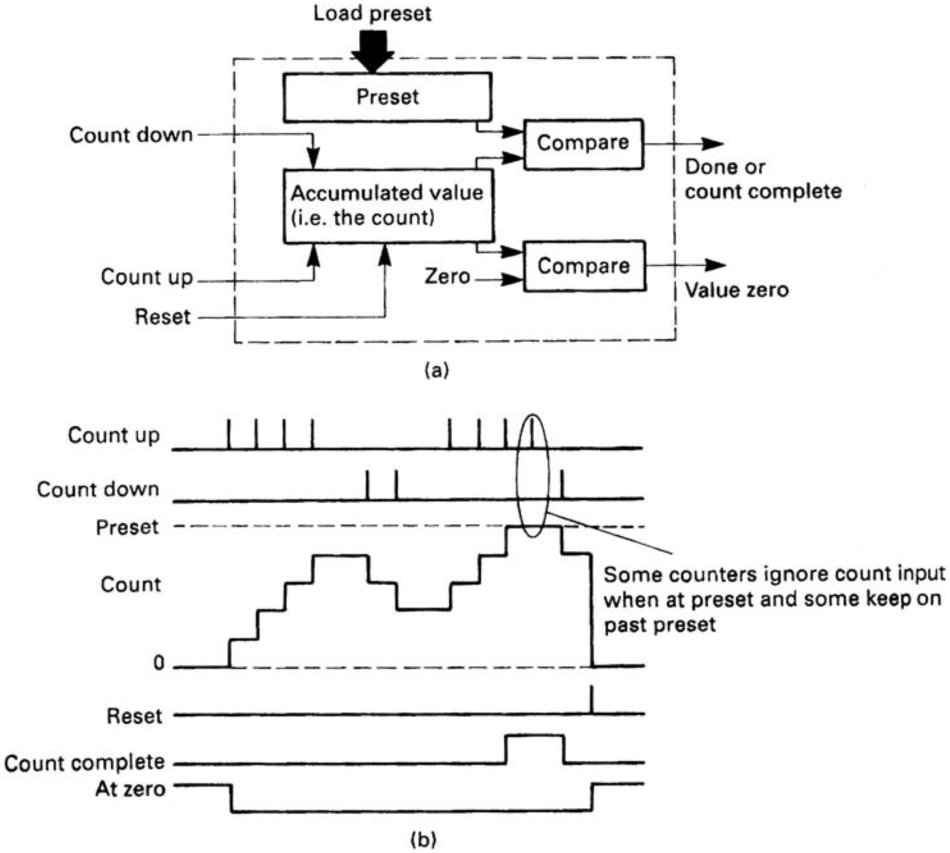


Figure 16.37 The up/down counter: (a) counter diagram; (b) counter operation

as a sequencer) the state of an output depends not only on the state of the inputs, but also on what was occurring previously. It is not therefore possible to draw a truth table from which the required logic can be deduced.

Consider, for example, the simple motor starter circuit of *Figure 16.43*(*a*). With neither button pressed, the motor could be running or stopped depending on what occurred last. The operation can be described by *Figure 16.43*(*b*) which is known as a *state transistion diagram*, (often shortened to *state diagram*).

The square boxes are the states the system can be in; the motor can be running or stopped, and the arrows are the transitions that cause the system to change states. If the motor is running, pressing the stop button will cause the motor to stop. A bar above a signal (e.g. above stop PB OK) means signal not present; note the wiring of the stop PB and the signal sense. It is a useful convention to label states with numbers and transitions with letters.

State transition diagrams can be constructed from storage elements, with one less storage element than there are states, and the one default state being inferred by the absence of others. It therefore requires just one storage element (latch, SR flip flop or whatever) to implement the motor starter of *Figure 16.43*.

Figure 16.44 is a more complex example (based on a real silo). A preset weight of material is fed into a weigh hopper ready for the next discharge, which is initiated by a Discharge pushbutton. A hood then lowers (to reduce dust emissions) and the material discharges. After the discharge, the hood retracts and the weighhopper re-fills. An abort pushbutton stops a discharge, and a feed permit switch stops the feed.

There are two fault conditions; failure to get the batch weight in a given time (probably caused by material jamming in the feeder) and failure to get zero weight from the discharge (again in a given time and again probably caused by a material jam). Both of these trip the system from automatic to manual operation to allow the cause of the fault to be determined.

We can now draw the state diagram of *Figure 16.44*(*b*). The default state is the state that the system will enter from manual, and care needs to be taken in its selection. Here feed is the sensible choice; if the hopper if already full the system will immediately pass to state 1 (ready), if not, the hopper will be filled. The choice of any other state as default could lead to a wasted cycle through all the states with no material in the weigh hopper.

We can now construct a table linking the outputs to the states. This is straightforward and is given on *Figure 16.44*(*c*).

The next stage is to translate this state diagram into a PLC program. The programming method relies very much on the idea of the program scan, described earlier in Section 16.2.4. By breaking down the program for our state diagram into four areas as *Figure 16.45* we can control the

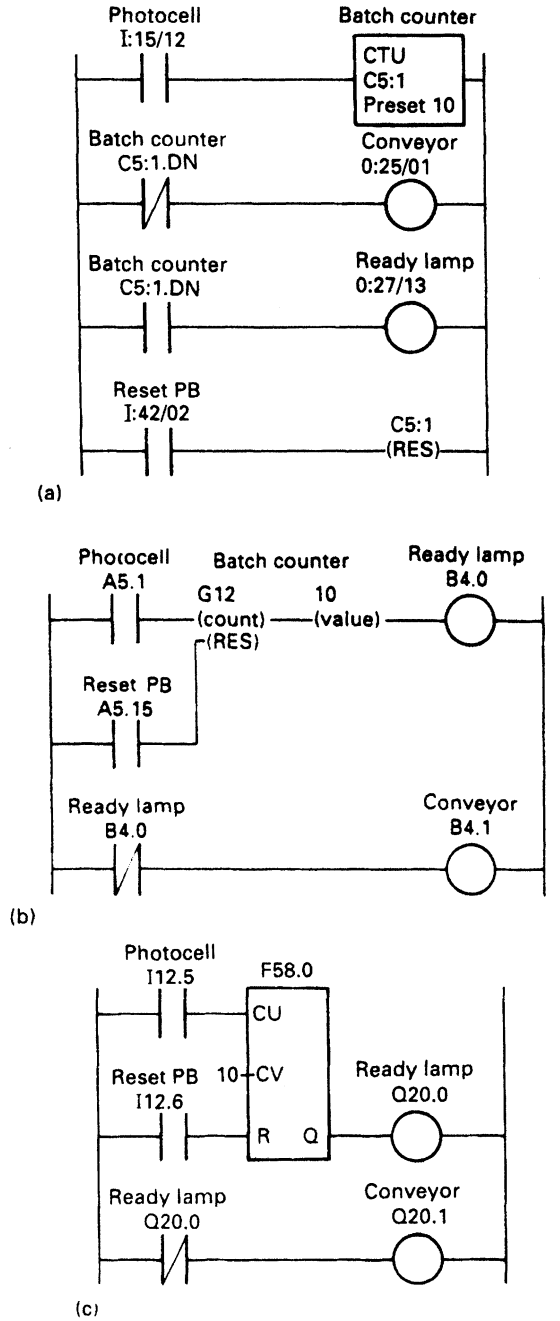

Photocell
I:15/12
Batch counter
CTU
C5:1
Preset 10
Batch counter
C5:1.DN
Conveyor
0:25/01
Batch counter
C5:1.DN
Ready lamp
0:27/13
Reset PB
I:42/02
C5:1
(RES)
(a)
Photocell
A5.1
Batch counter
G12
(count)
10
(value)
(RES)
Ready lamp
B4.0
Reset PB
A5.15
Ready lamp
B4.0
Conveyor
B4.1
(b)
Photocell
I12.5
F58.0
CU
Reset PB
I12.6
10
CV
R
Q
Ready lamp
Q20.0
Ready lamp
Q20.0
Conveyor
Q20.1
(c)

Figure 16.38 A simple batch counter programmed on three different machines: (a) Allen Bradley PLC-5; (b) GEM-80; (c) Siemens S5 in logic notation

order in which each stage operates. The actual layout is not critical, but it is essential for transitions and states to be kept separate and not mixed.

Automatic/manual selection comes first, this is achieved with the simple rung of *Figure 16.46*. Automatic mode is only allowed if there are no faults and the hood is raised.

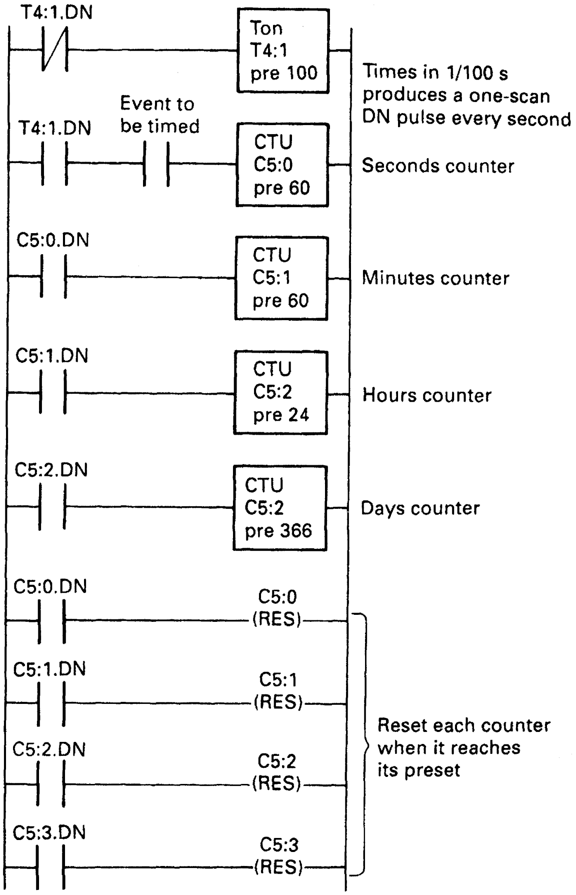

T4:1.DN
Ton
T4:1
pre 100
Times in 1/100 s produces a one-scan DN pulse every second
T4:1.DN
Event to be timed
CTU
C5:0
pre 60
Seconds counter
C5:0.DN
CTU
C5:1
pre 60
Minutes counter
C5:1.DN
CTU
C5:2
pre 24
Hours counter
C5:2.DN
CTU
C5:2
pre 366
Days counter
C5:0.DN
C5:0
(RES)
C5:1.DN
C5:1
(RES)
C5:2.DN
C5:2
(RES)
C5:3.DN
C5:3
(RES)
Reset each counter when it reaches its preset

Figure 16.39 Cascading counters to give a long delay. Allen Bradley PLC-5 notation has been used

Next come the transitions, some of which are shown on *Figure 16.47*. These are straightforward and need little comment. Note that the first contact in each rung is a state, so inputs are only examined at the correct point in a sequence.

Some of the states themselves are given in *Figure 16.48*. With the exception of state 0, simple latches have been used throughout for the states and the auto/man selection so that after a power failure the system will resume in manual mode. Note that these are set and reset by the transitions.

Finally we have the outputs themselves on *Figure 16.49*. An output is energised during the corresponding state(s) in automatic or from the manual maintenance push button in manual.

The state diagram technique is very powerful, but it can lead to confusion if the basic philosophy is not understood. The often quoted argument is it takes more rungs or logic elements than a direct approach programmed around the outputs.

This is true, but programming around the outputs can lead to very twisted and difficult to understand programs. *Figure 16.50* is one rung roughly corresponding to state 2 of our state diagram. It mixes manual and automatic operation and its action is by no means clear (known colloquially as *Spaghetti programming*). Problems can arise where transitions go against the program scan as transition E on the earlier *Figure 16.44*(*b*). If care is not taken, a sequence based

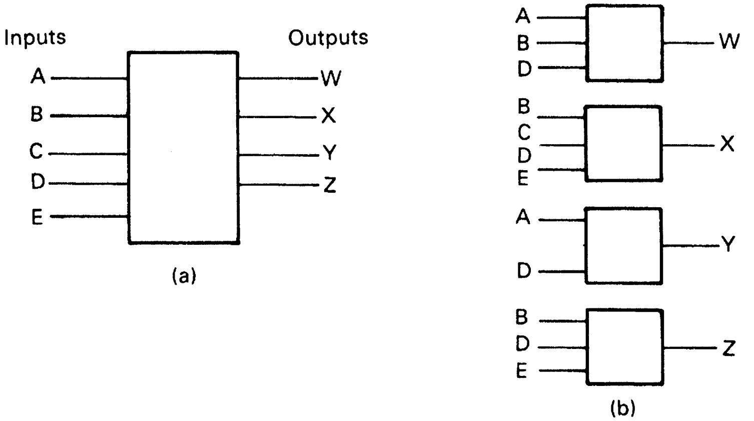


Figure 16.40 Combinational Logic: (a) top level view; (b) broken down into smaller blocks, each with one output, for programming

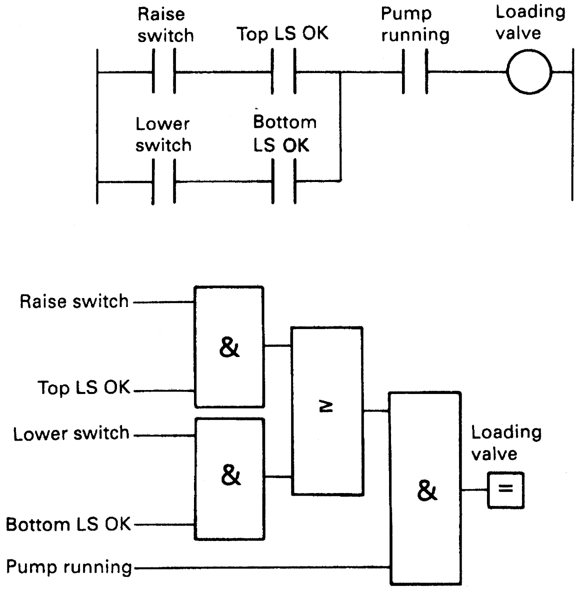


Figure 16.41 Combinational logic in ladder and logic notations. Both perform the same function

purely on outputs can easily end up doing two things at once, or nothing at all because of the way the program scan operates. Modifications are also tricky with a direct approach, but simple with a state diagram.

State diagrams are being formalised by the International Electrotechnical Commission and the British Standards Institute, and already exist with the French Standard Grafset. These are basically identical to the approach outlined above, but introduce the idea of parallel routes which can be operated at the same time. *Figure 16.51*(*a*) is called a *divergence*, state 0 can lead to state 1 for condition *s* OR to state 2 for condition *t* with transitions *s* and *t* mutually exclusive. This is the form of the state diagrams described so far.

Figure 16.51(*b*) is a *simultaneous divergence*, where state 0 will lead to state 1 AND state 2 simultaneously for transition *u*. States 1 and 2 can now run further sequences in parallel.

Figure 16.51(*c*) again corresponds to the state diagrams described earlier, and is known as a *convergence*. The sequence can go from state 5 to state 7 if transition *v* is true OR from state 6 to state 7 if transition *w* is true.

Figure 16.51(*d*) is called a *simultaneous convergence* (note again the double horizontal line) state 7 will be entered if the left-hand branch is in state 5 AND the right-hand branch is in state 6 AND transition *x* is true.

The state diagram is so powerful that most medium size PLCs include it in their programming language in one form or another. Telemecanique give it the name Grafcet (with a ‘c’), others use the name Sequential Function Chart (SFC) (Allen Bradley) or Function Block (Siemens). We will return to these in the next chapter.

Even the simple Mitsubishi F2 supports state diagrams with its STL (Stepladder) instruction. These have the prefix S and can range from S600 to S647. They have the characteristic that when one or more are set, any others energised are automatically reset. A RET instruction ends the sequence. The state diagram of *Figure 16.52*(*a*) thus becomes the ladder diagram of *Figure 16.52*(*b*).

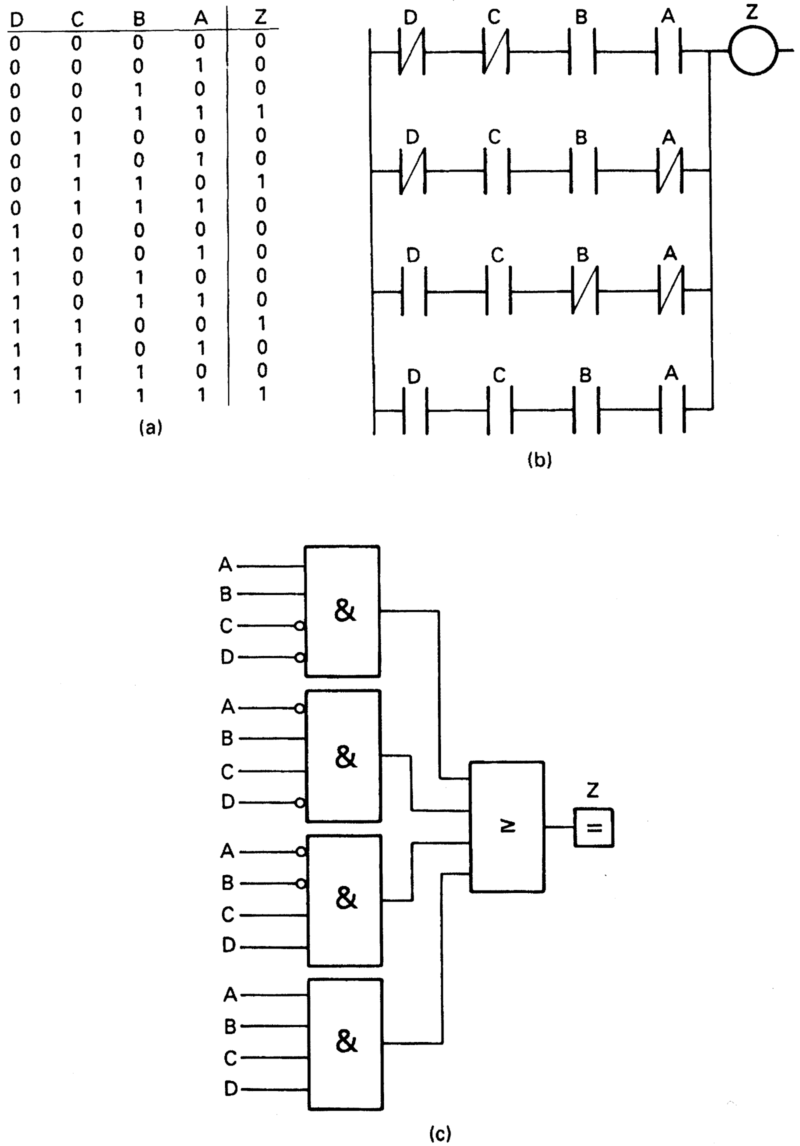
Where there are no branches and the sequence is a simple ring (operating rather like a uniselector) a sequence can be driven by a counter which selects the required step. The counter is stepped when the transitions for the current step are met. The GEM-80 has a SEQR (sequence) instruction which acts as a sixteen step uniselector.

16.3.11 IEC 1131

We have seen that PLCs can be programmed in several different ways. In recent years the International Electrotechnical Commission (IEC) have been working towards defining standard architectures and programming methods for PLCs. The result is IEC 1131, a standardised approach which will help at the specification stage and assist the final user who will not have to undergo a mind-shift when moving between different machines.

The earliest, and probably still the commonest, programming method described is the *Ladder Diagram* (or LD in IEC 1131).

Function Block Diagrams (FBDs) use logic gates (AND/OR etc.) for digital signals and numeric function blocks (arithmetic, filters, controllers, etc.) for numeric signals. FBDs are similar to PLC programs for the ABB Master and Siemens SIMATIC families. There is a slight tendency for digital programming to be done in LD, and analog programming in FBD.



D	C	B	A	Z
0	0	0	0	0
0	0	0	1	0
0	0	1	0	0
0	0	1	1	1
0	1	0	0	0
0	1	0	1	0
0	1	1	0	1
0	1	1	1	0
1	0	0	0	0
1	0	0	1	0
1	0	1	0	0
1	0	1	1	0
1	1	0	0	1
1	1	0	1	0
1	1	1	0	0
1	1	1	1	1

(a)

D
C
B
A
Z
(b)
A
B
C
D
&
≥
=
Z
(c)

Figure 16.42 Building combinational logic from a truth table: (a) truth table; (b) Direct conversion to a ladder program. Each row in the truth table which makes Z = 1 is represented by one level on the branch; (c) Direct conversion to a logic diagram. Each row in the truth table which makes Z = 1 is represented by an AND gate. The AND gate outputs are then OR'd together

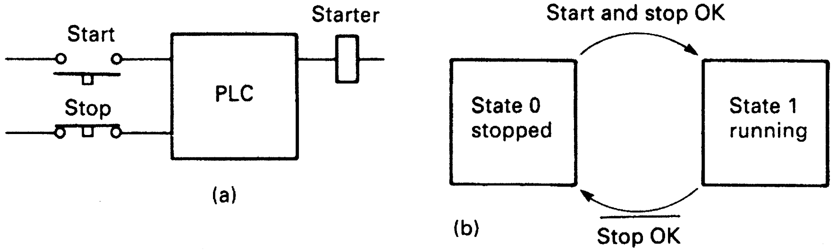
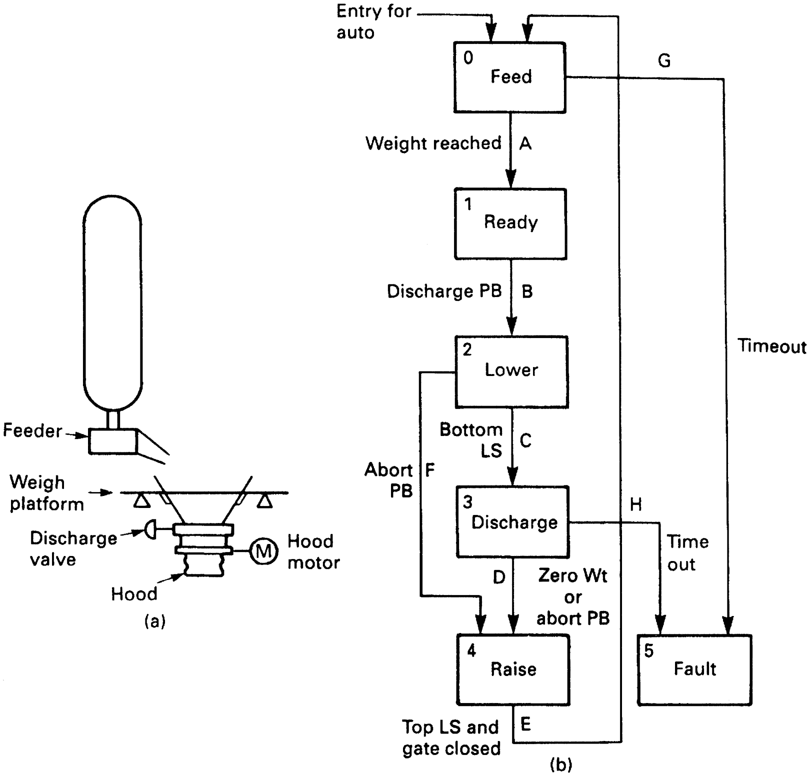

Start
Stop
PLC
Starter
(a)
Start and stop OK
State 0
stopped
State 1
running
$\overline{\text{Stop OK}}$
(b)

Figure 16.43 A simple state transition diagram; (a) A motor starter; (b) State transition diagram. Note that with no buttons pressed the system can be in either state


Feeder
Weigh platform
Discharge valve
Hood
Hood motor
M
(a)
Entry for auto
0 Feed
G
Weight reached
A
1 Ready
Discharge PB
B
Timeout
2 Lower
Bottom LS
C
Abort PB
F
3 Discharge
H
Time out
D
Zero Wt or abort PB
4 Raise
5 Fault
Top LS and gate closed
E
(b)

	State	Feed	Lower	Raise	Disch	Auto Permit	Ready Lamp	Fault Lamp
Feed	0	1	0	0	0	1	0	0
Ready	1	0	0	0	0	1	1	0
Lower	2	0	1	0	0	1	0	0
Discharge	3	0	0	0	1	1	0	0
Raise	4	0	0	1	0	1	0	0
Fault	5	0	0	0	0	0	0	1

(c)

Figure 16.44 A more complex state transition diagram for a real plant: (a) physical layout; (b) state transition diagram; (c) output table

Many control systems are built around state transition diagrams, and IEC 1131-3 calls these *Sequential Function Charts* (SFCs). The standard is based on the French Grafcet standard shown earlier on *Figure 16.51.*

Finally are text based languages. *Structured Text* (ST) is a structured high level language with similarities to Pascal and C. *Instruction List* (IL) contains simple mnemonics such as LD, AND, ADD etc. IL is very close to the programming method used on small PLCs where the user draws a program up in ladder form on paper, then enters it as a series of simple instructions.

Figure 16.53 illustrates all of these programming methods.

A given project does not have to stick with one method, they can be intermixed. A top level, for example, could be an SFC, with the states and transitions written in ladder rungs or function blocks as appropriate.

It will be interesting to see the effect of IEC 1131-3. Most attempts at standardisation fail for reasons of national and commercial pride. MAP, and latterly fieldbus, have all had problems in gaining wide acceptance. A standard will be useful at the design stage, and could be accepted by the end user if programming terminals presented a common face regardless of the connected machine.

16.4 Numerics

16.4.1 Numerical applications

So far we have been primarily discussing single bit operations. Numbers are also often part of a control scheme; a PLC might need to calculate a production rate in units per

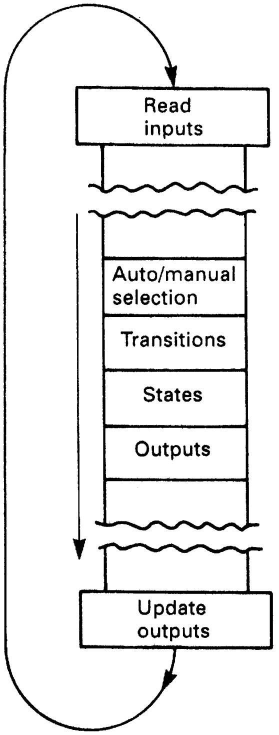

Read inputs
Auto/manual selection
Transitions
States
Outputs
Update outputs

Figure 16.45 The program scan and the layout of a state transition diagram program

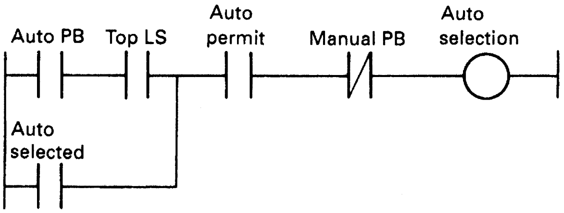

Auto PB
Top LS
Auto permit
Manual PB
Auto selection
Auto selected

Figure 16.46 Auto/manual selection

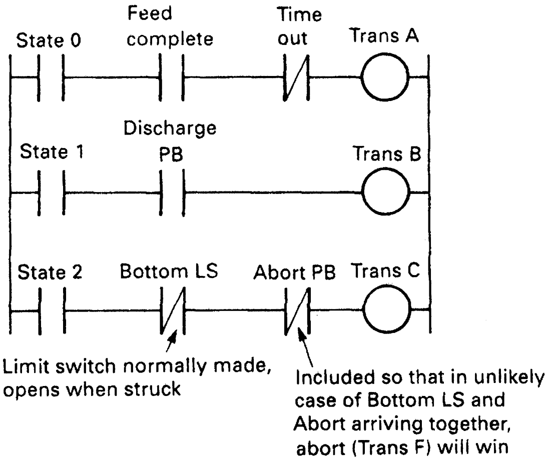

State 0
Feed complete
Time out
Trans A
State 1
Discharge PB
Trans B
State 2
Bottom LS
Abort PB
Trans C
Limit switch normally made, opens when struck
Included so that in unlikely case of Bottom LS and Abort arriving together, abort (Trans F) will win

Figure 16.47 The first three transitions

hour averaged over a day, or give the amount of liquid in a storage tank. Such operations require the ability to handle numeric data.

16.4.2 Numeric representations

Most PLCs work with a 16 bit word, allowing a positive number in the range 0 to +65 535 to be represented, or a signed (positive or negative) number in the range −32 768 to +32 767. In the latter case, known as *two's complement*, the most significant bit represents the sign, being 1 for negative numbers and 0 for positive numbers.

Numbers such as these are known as integers, and can only represent whole numbers in the above range. Where larger whole numbers are required, two sixteen words can be used allowing a range −2 147 483 648 to +2 147 483 647. This type of integer is available in the ABB Master (where it is known as a *long integer*) and the 135-U and 155-U in the Siemens family (where the term *double word integer* is used).

Where decimal fractions are needed (to deal with a temperature of 45.6 °C for example) a number form similar to that found on a calculator may be used. These are known as *real* or *floating point* numbers, and generally consist of two sixteen bit words which contain the *mantissa* (the numerical portion) and the *exponent*. In base ten, for example, the number 74 057 would have a mantissa of 7.4057 and an exponent of 4 representing 10^4. PLCs, of course, work in binary and represent mantissa and exponent in two's complement form.

Real numbers are very useful but their limitations should be clearly understood. There are two common problems. The first occurs when large numbers and small numbers are used together. Suppose we had a system operating to base ten with four significant figures, and we wish to add 857 800 (stored as 8.578E5) and 96 (stored as 9.600E1). Because the smaller number is outside the range (four significant figures) of the larger, it will be ignored giving the result 857 800 + 96 = 857 800.

The second problem occurs when tests for equality are made on real numbers. The conversion of decimal numbers to binary numbers can only be made to the resolution of the floating point format. If real numbers must be used for comparison, a simple equates (=) is very risky. The composites > =, (greater than or equals), and < =, (less than or equals), are safer, but it is generally better practise to use integers for tests if at all possible.

The final representation, BCD for *Binary Coded Decimal*, is used for connection to outside world devices such as digital displays or thumbwheel switches. Such devices are arranged in a decimal format, with 4 binary bits per decade, for example

```
1001 0101
```

can be interpreted in BCD as 95.

This representation is wasteful, as six 'numbers' are not used per four bits (10 to 15 inclusive). It is, however, a convenient form to use with external wiring. Most PLCs therefore have instructions which convert BCD to the internal binary format of the PLC, and binary back to BCD.

The types of numbers available in each PLC range vary considerably according to the model (and obviously the price). The Mitsubishi F2, for example, purely allows movement, comparison and output of numerical data from counters or timers, making it essentially a bit operation machine.

In the Siemens range, the popular 115-U uses only 16 bit integer numbers but the next model in the range, the 135-U,

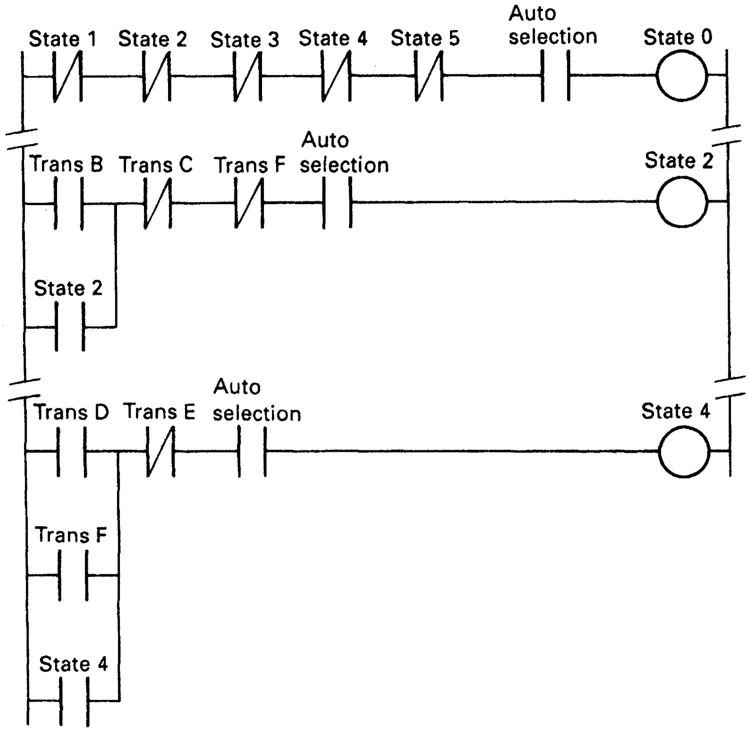

State 1
State 2
State 3
State 4
State 5
Auto selection
State 0
Trans B
Trans C
Trans F
Auto selection
State 2
State 2
Trans D
Trans E
Auto selection
State 4
Trans F
State 4

Figure 16.48 Three of the six states

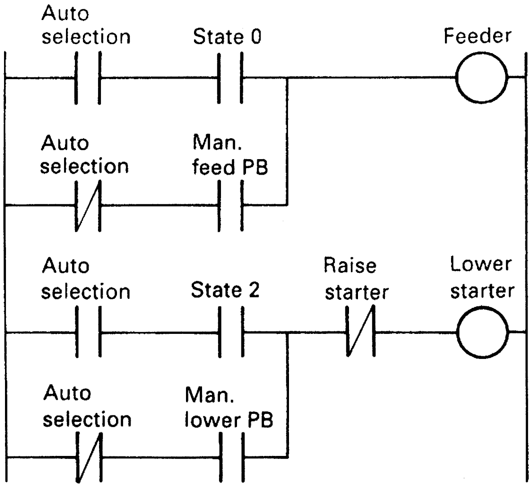

Auto selection
State 0
Feeder
Auto selection
Man. feed PB
Auto selection
State 2
Raise starter
Lower starter
Auto selection
Man. lower PB

Figure 16.49 Two of the plant outputs

can handle 16 bit and 32 bit integers and floating point numbers. A similar spread of capabilities will be found amongst the Allen Bradley, GEM-80 and ABB families.

16.4.3 Data movement

Numbers are often moved from one location to another; a timer preset may be required to be changed according to plant conditions, a counter value may need to be sent to an output card for indication on a digital display or the result of some calculations may be used in another part of a program.

The Allen Bradley PLC-5 uses one rung per move operation, and is possibly the simplest to explain first. Its simplicity of one rung per operation is continued in all the arithmetic functions we shall consider, but it can lead to more rungs being used for a given operation than in other machines.

Figure 16.54(a) shows the form of the rung. It starts with some binary conditions; if these are all made the output MOV (for MOVE) is obeyed, transferring data from the source to the destination. The source and destination can be any location where numerical data can occur, for example

Integer number (e.g. `N7:26`)
Floating point number (e.g. `F8:33`)
Counter or timer preset (e.g. `C5:17.PRE` or `T4:52.PRE`)
Counter or timer accumulated value (e.g. `C5:22.ACC` or `T4:6.ACC`)
I/O word data (e.g. `I:23` which is all 16 bits from inputs on card 3 in rack 2)

If data is transferred between integer and floating point forms, the conversion is performed automatically however care must be taken transferring floating point numbers to integers as an error can occur if the floating point number is outside the integer range. Finally, as a source only, a constant (such as 3, 17 or 4057) can be used.

The example of *Figure 16.54(a)* thus moves the number held in `N7:34` to the preset of timer `T4:6` when the rung conditions are met.

Siemens and GEC use a slightly different approach which leads to more compact programs. Both treat a data movement

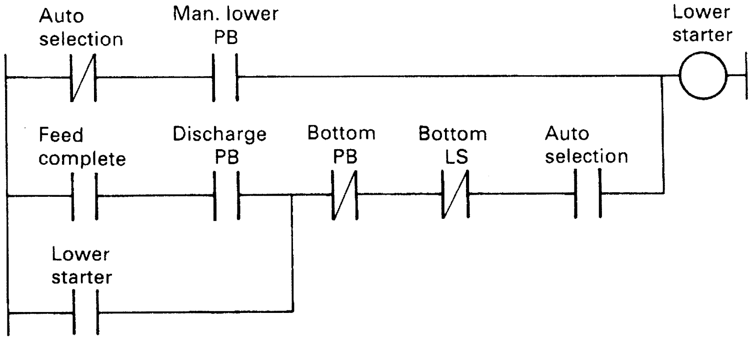

Auto selection
Man. lower PB
Lower starter
Feed complete
Discharge PB
Bottom PB
Bottom LS
Auto selection
Lower starter

Figure 16.50 An example of spaghetti programming approximating to state 2

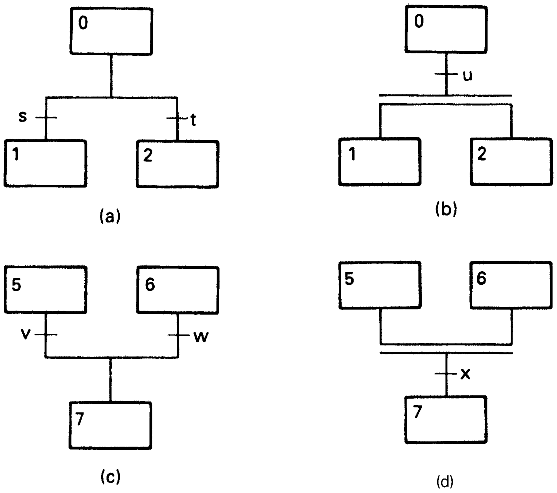

0
s
t
1
2
(a)
0
u
1
2
(b)
5
6
v
w
7
(c)
5
6
x
7
(d)

Figure 16.51 Grafset symbols: (a) divergence; (b) simultaneous divergence; (c) convergence (d) simultaneous convergence

as two separate instructions via a separate accumulator (a single word storage location). Siemens use the instructions load to move data from a source to the accumulator, and transfer to move data from the accumulator to the destination as *Figure 16.54*(*b*). The data can come from (or go to) any data storage area, some of which are

IW	a 16 bit input word
QW	a 16 bit output word
T	a timer word
C	a counter word
DW	a 16 bit data storage word

Figure 16.54(*b*) would thus be programmed as

```
:L  T113  (timer value to accumulator)
:T  DW45  (accumulator to data word 45)
```

The use of the accumulator is not obvious in the GEM-80. The−<AND>− instruction puts the binary number from the specified location (again internal storage or

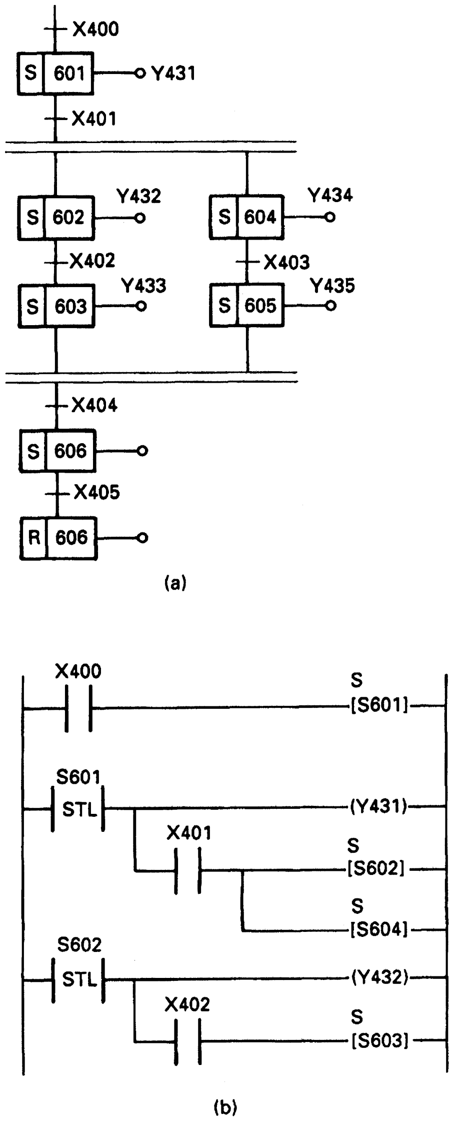
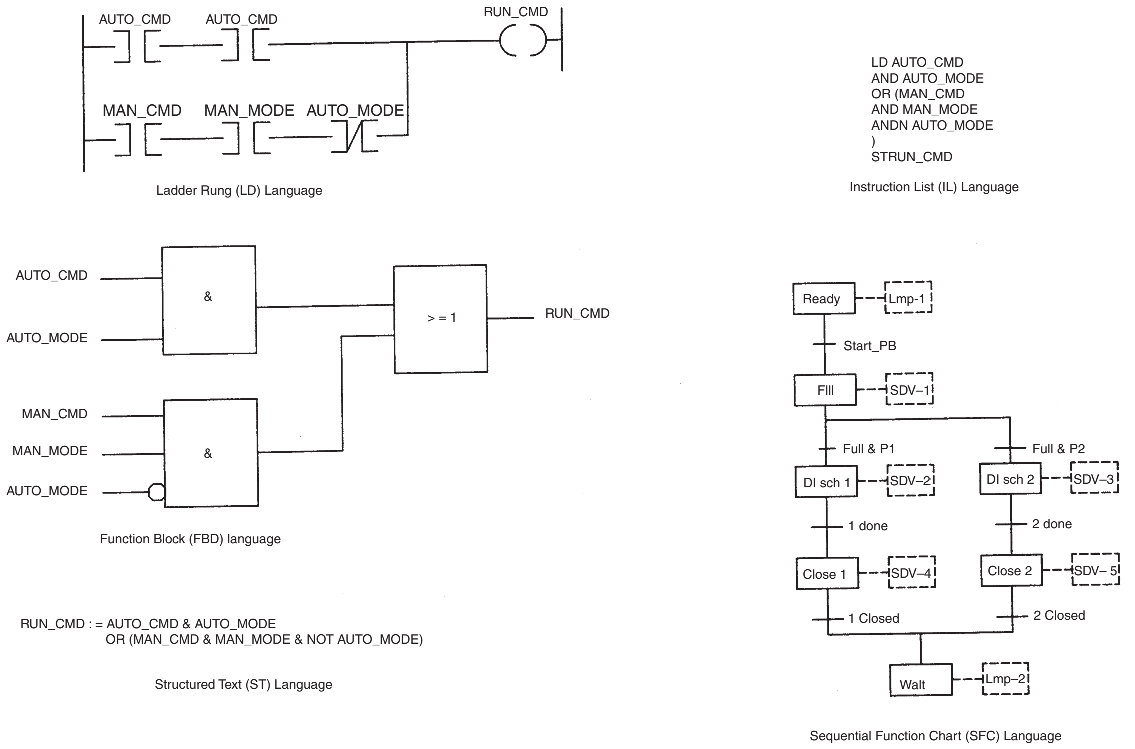

X400
S 601
Y431
X401
S 602
Y432
S 604
Y434
X402
X403
S 603
Y433
S 605
Y435
X404
S 606
X405
R 606
(a)
X400
S
[S601]
S601
STL
(Y431)
X401
S
[S602]
S
[S604]
S602
STL
(Y432)
X402
S
[S603]
(b)

Figure 16.52 State diagrams on the Mitsubishi F2: (a) state diagram; (b) part of the ladder diagram corresponding to (a)


AUTO_CMD
AUTO_CMD
RUN_CMD
MAN_CMD
MAN_MODE
AUTO_MODE
AUTO_CMD
AUTO_MODE
&
> = 1
RUN_CMD
MAN_CMD
MAN_MODE
&
AUTO_MODE
Ready
Lmp-1
Start_PB
Fill
SDV–1
Full & P1
Full & P2
Dl sch 1
SDV–2
Dl sch 2
SDV–3
1 done
2 done
Close 1
SDV–4
Close 2
SDV– 5
1 Closed
2 Closed
Walt
Lmp–2

Ladder Rung (LD) Language

```
LD AUTO_CMD
AND AUTO_MODE
OR (MAN_CMD
AND MAN_MODE
ANDN AUTO_MODE
)
STRUN_CMD
```

Instruction List (IL) Language

Function Block (FBD) language

```
RUN_CMD : = AUTO_CMD & AUTO_MODE
        OR (MAN_CMD & MAN_MODE & NOT AUTO_MODE)
```

Structured Text (ST) Language

Sequential Function Chart (SFC) Language

Figure 16.53 The five programming methods defined in IEC 1131-3

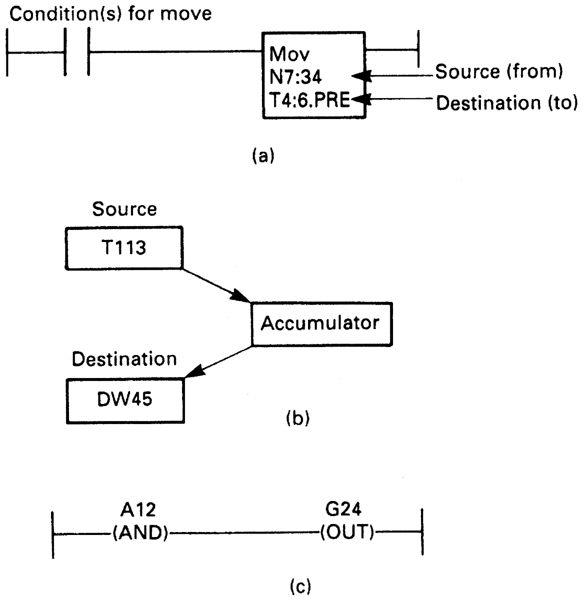

Condition(s) for move
Mov
N7:34
T4:6.PRE
Source (from)
Destination (to)
(a)
Source
T113
Accumulator
Destination
DW45
(b)
A12
(AND)
G24
(OUT)
(c)

Figure 16.54 Data Movement: (a) Allen Bradley PLC-5; (b) Siemens S5; (c) GEM-80

I/O) into the rung, and the `-<OUT>-` instruction puts the value from the rung to the specified address. In *Figure 16.54(c)* the (binary) value from 16 bit input word A12 is placed into 16 bit storage word G24.

BCD/binary conversion is available with `-<BCDIN>-` and `-<BCDOUT>-` instructions, the direction of the conversion being obvious.

In the ABB Master, the points between which data is to be transferred are simply linked on the logic diagram.

16.4.4 Data comparison

Numerical values often need to be compared in PLC programs; typical examples are a batch counter saying the required number of items have been delivered, or alarm circuits indicating, say, a temperature has gone above some safety level.

These comparisons are performed by elements which have the generalised form of *Figure 16.55*, with two numerical inputs corresponding to the values to be compared, and a binary (on/off) output which is true if the specified condition is met.

Many comparisons are possible; most PLCs provide

A Greater Than B	(A>B)
A Greater Than or Equal to B	(A>=B)
A Equals B	(A=~~B~~)
A Less Than or Equal to B	(A<=B)
A Less Than B	(A<B)

where A and B are numerical data. With real (floating point) numbers the equal test should be avoided for the reasons given in the previous section. There are many other possible comparisons; a PLC-5, for example has a limit instruction which tests for A lying between B and C and the GEM and Siemens have a not equal test.

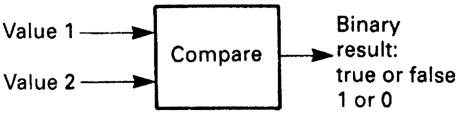


Figure 16.55 Basic idea of data comparison

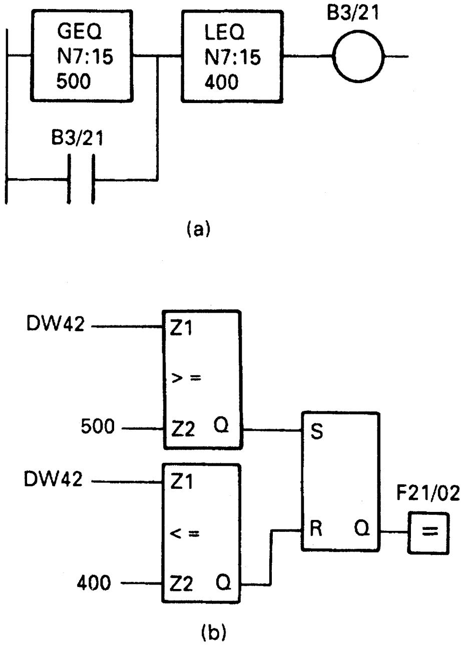


Figure 16.56 Use of data comparison for a high temperature alarm: (a) Allen Bradley PLC-5; (b) Siemens S5 in logic notation

Figure 16.56 shows the setting and resetting of an alarm flag `B3/21` (for a PLC-5 ladder diagram) and `F21/02` (for Siemens logic symbols). The alarm bit is set if temperature (read from an analog input card in format NN.N °C and held in `N7:15` in the PLC-5 or `DW42` in the Siemens 115-U) goes above 50.0 °C. Once set, the alarm is stored until the temperature goes below 40.0 °C.

16.4.5 Arithmetical operations

Numerical data implies the ability to do arithmetical operations, and all PLCs we are considering (apart from the simple F2) provide the ability to do at least four function maths (add, subtract, multiply and divide).

In Section 16.4.2 we discussed integer and floating point numbers. Care needs to be taken with integer operations. The range of a 16 bit two's complement number is −32 768 to +32 767. If an arithmetical operation goes outside this range, the number will overspill, for example

```
 26732

+ 8647

 -----

-30157            in 16 bit two's complement
```

which is not quite the expected result. The PLCs have an overspill flag which can be examined and used to flag an alarm, or set the result to, say, zero with a move instruction. Similar precautions need to be taken with subtraction and multiplication (the latter being particularly vulnerable to giving an overspill; for example 200 × 200 = 40 000, well over-range.)

Even greater care needs to be taken with division. A fault condition on external plant or a PLC input card or a programming error can lead to a divide by zero error. This will stop many PLCs dead in their tracks with a 'Program Fault'. It is therefore good practice to proceed any vulnerable divide instruction with a limit check to ensure it will only be obeyed when a sensible result is obtained.

Each PLC manufacturer handles arithmetic in a slightly different way with varying degrees of ease and readability. None are as simple as a high level language such as BASIC or Pascal, and the facilities are generally limited to four function maths plus square root in all bar the most expensive machines.

A PLC-5 uses maths blocks such as `ADD, SUB, MULT, DIV`, giving a simple, if somewhat lengthy, program. *Figure 16.57* shows how a simple calculation could be performed for a self correcting length cutting program. More powerful PLC-5s (such as the 5–40) have a block compute instruction (CPT) which allows a mathematical expression to be evaluated in a single instruction.

The 115-U only evaluates arithmetic instruction in STL (statement list) format. It will be remembered from our discussion of the accumulator that the load, (L) and transfer (T) instructions use an internal accumulator. There are, in fact, two accumulators, and a load instruction moves the contents of accumulator 1 to accumulator 2 then moves the contents of the source to accumulator 1, shown in *Figure 16.58*(*a*). An arithmetic instruction (add, subtract, etc.) works on the contents of both accumulators. *Figure 16.58*(*b*), thus adds two numbers and transfers the result to storage.

The Siemens equivalent of *Figure 16.57* would be

```
L DW30   (required length)
L DW31   (measured length)
SUB      (leaving error in Acc 1)
L DW32   (gain)
MULT     (leaving correction)
L DW40   (the old cut length)
ADD      (add change to give new length)
T DW40   (put back to store)
```

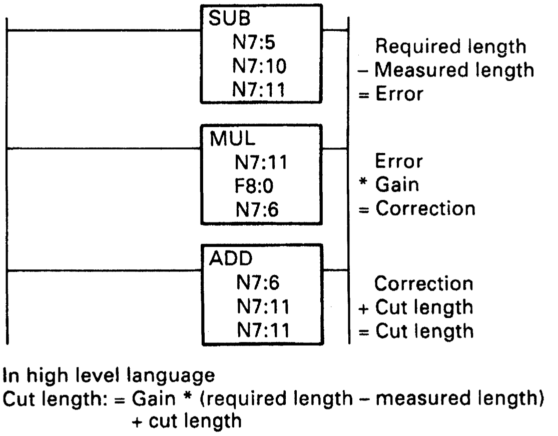


Figure 16.57 Arithmetic in the PLC-5

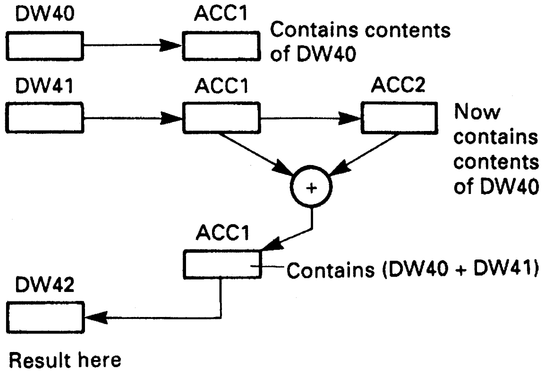


Figure 16.58 Arithmetic in a Siemens S5

The most understandable form of representation is possibly the GEM-80 ladder and the ABB Master formats shown in *Figure 16.59*(*a*) and (*b*) respectively.

All maths operations, particularly those involving floating point numbers, are time consuming, and it is good programming practice to only obey instructions when they are needed, and not waste time repetitively obeying them on every PLC scan.

16.4.6 Analog signals

So far we have considered signals that are essentially digital (on/off) in nature plus simple numerical data from timers and counters. Often, though, a PLC will be required to measure, or control, plant signals which can assume any value in some predetermined range. Typical signals of this type are temperatures, flows, pressure, speeds etc. These are known as analog signals. In a similar way a PLC may have to produce analog output signal to drive meters and proportional valves or provide a speed reference for a motor drive controller.

To meet these requirements a PLC needs analog input and output cards. These have somewhat different characteristics to the simple digital cards we have discussed so far. This section considers analog signals and the way they are handled.

An analog input card converts a continuously varying analog signal to a digital form that can be used inside a PLC program. The analog signal is generally represented initially, at least, as an integer number.

This analog to digital conversion (usually known by the initials ADC) is inherently accompanied by a loss of resolution which depends on the number of bits used. An 8 bit byte for example, can represent an integer in the range 0–255. If this was used to represent an analog signal measuring a flow with a span (range) from 0–1800 l/min, one bit will represent approximately 7 l/min (given by 1800/255). Any control strategy in the program based on finer resolution is meaningless (and particular care should be taken with comparisons, as some values can never be obtained; a flow of 138 l/min, for example, would never be given by our 8 bit system, it would jump from 134 l/min to 141 l/min. Comparisons should therefore always be based on (greater than or equal to) or (less than or equal to)).

A commoner resolution is 12 bits. This gives a representation as an integer from 0–4095. With our flow of 0–1800 l/min, one bit would represent just under 0.5 l/min (1800/4095 = 0.44).

This 'coarseness' is not the problem it might at first appear. Although an analog transducer can give any value

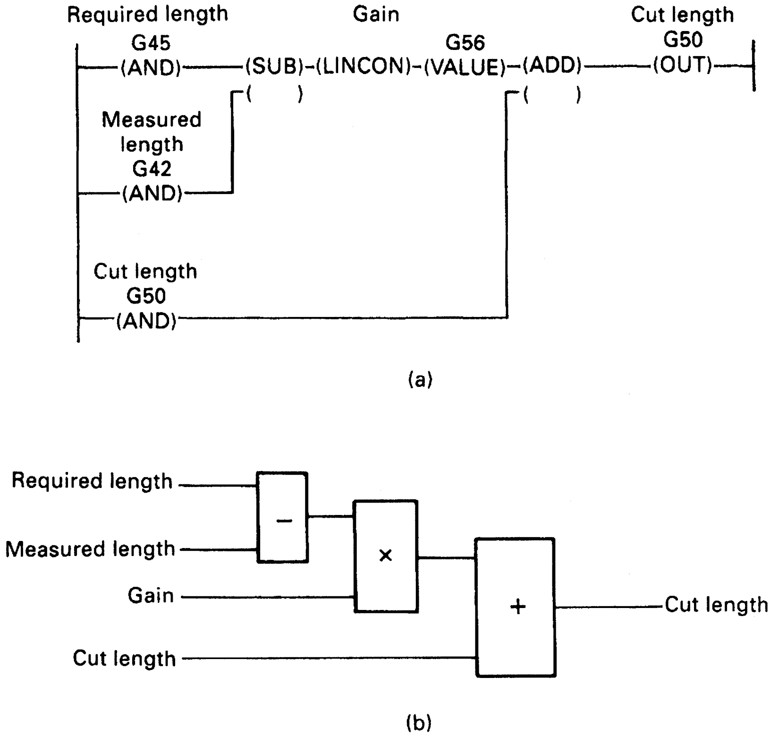


Figure 16.59 The same mathematical function in a GEM-80 and ABB Master: (a) GEM-80 program. LINCON is an arithmetic function used to avoid truncation errors with integer arithmetic; (b) ABB master program using function blocks. Variables are accessed by database names

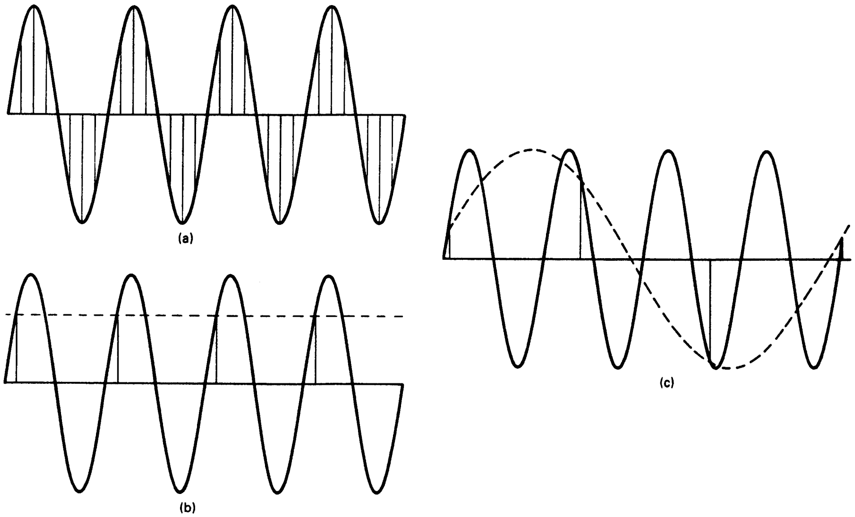


Figure 16.60 The effect of the sampling rate

in its span, it will have inherent errors. Many first line transducers are only 2% accurate. If our flow transducer had 2% accuracy, its measurement could be in error by 36 l/min. Alongside this error, the 7 l/min resolution from an eight bit card is probably quite reasonable.

It is therefore useful to think of the resolution in terms of an error which is to be added to the error from the transducer itself

No of bits	Range	Error
8	0-255	0.5%
10	0-1023	0.1%
12	0-4095	0.025%

Few industrial transducers have an accuracy better than 0.1%, and a 12 bit conversion will add little error in most applications.

The conversion from an analog signal to a digital representation is not instantaneous. Typically signals are read ten times per second. An analog input card thus takes regular 'snapshots' of each analog signal. In *Figure 16.60*(*a*) this causes no problems, in *Figure 16.60*(*b*) information is starting to be lost and in *Figure 16.60*(*c*) a totally false view of the signal is being given. This latter effect is known as '*aliasing*'.

It is therefore very important to have a sufficiently fast conversion time. Every analog signal will have a maximum frequency at which it can change, and can be represented by a gain/frequency plot as *Figure 16.61* from which the bandwidth and the critical frequency f_c can be observed. To get a true series of 'snapshots' we must sample the signal at least twice the rate of f_c. If a certain analog signal has a maximum frequency of 2 Hz, we must at least sample it at 4 Hz, or once every 250 ms. This, somewhat simplified, is known as *Shannon's sampling theorem*. In real systems, f_c is rarely known precisely and a scan rate of 4 f_c to 10 f_c is normally chosen to give a reasonable safety margin. For our 2 Hz signal, an 8 Hz sampling rate or 12.5 ms conversion time, would be needed. It is good practice to pass the signal through a low pass filter before the ADC to ensure frequencies above f_c are removed. This is known as an *anti-aliasing filter*.

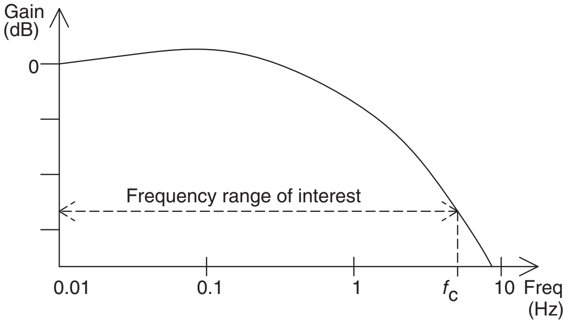


Figure 16.61 Gain/frequency response for an industrial process

Surprisingly this rarely gives problems. Practical industrial systems, dealing with real plant signals concerned with materials with significant mass, rarely have bandwidths greater than 0.5 Hz, and any frequency higher than this can be considered to be extraneous noise and filtered out. Temperature loops, for example, can often be sampled as slowly as once every few minutes without introducing any errors.

A typical analog input card can read eight 12 bit signals, each ranging from 0–4095 in their 'raw' form. Generally these will need to be accessed via the PLC program and converted to engineering units such as °C, or psi, or l/min.

A common method of handling these signals, is shown in *Figure 16.62*. A block of storage locations in the PLC store is directly associated with the analog input card. The card 'free runs', writing digitised values into the store from where they can be read by the rest of the program. In Siemens PLCs with fixed slot addressing, for example, the store addresses are determined directly by the analog card position in the rack; a card in slot 2 of the first rack will write its values to a block of stores starting at location 192.

Conversion from a raw 12 bit signal to engineering units can have subtle traps for the unwary. In theory the conversion is simple. If N is the raw signal, H_R the high range

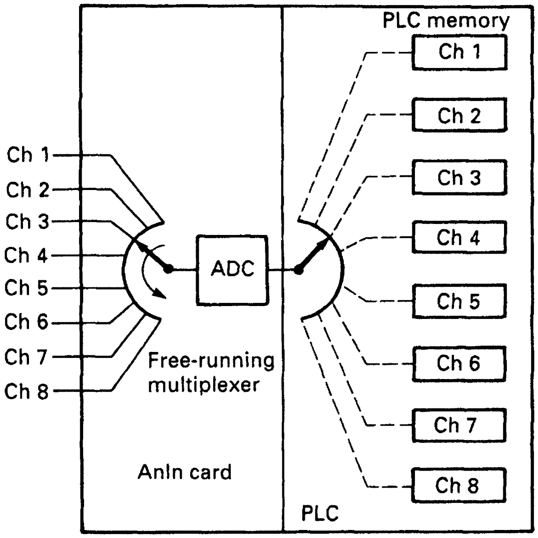


Figure 16.62 Linking channels on an analog input card to a PLC's memory

signal (corresponding to 4095) and L_R the low range (corresponding to zero) then the measured value, M_V is simply

$$M_V = \frac{N \times (H_R - L_R)}{4095} + L_R \qquad (16.1)$$

If the calculation is done with real (floating point) numbers there should be no problem, and Equation 16.1 can be used directly.

If, however, integer numbers have to be used, great care must be taken. If the multiplication $N \times (H_R - L_R)$ is performed first, arithmetic overspill is likely unless 32 bit results can be accommodated. If the division $N/4095$ is performed first, the equation will not work as N is always less than 4095 giving an integer result of zero (and an M_V of L_R). Wherever possible real numbers should be used if Equation 16.1 has to be performed.

To avoid this problem, the different manufacturers have devised methods to read analog input signals. In the ABB Master for example, the database for each signal defines H_R, L_R, the sample rate and a name by which the signal will be referred to in the program. There are, obviously, detail differences, so by way of example we will look at the way analog signals are read by an Allen Bradley PLC-5.

The Allen Bradley PLC-5 reads analog signals with an analog input card (1771-IFE) which can in its simplest form, read 8 analog inputs. The PLC communicates with the card via instructions called block transfers which transfer data to (or from) a block of store locations. Data transfers from the PLC to a card are called block transfer writes (`BTW`) and, not surprisingly, transfers from a card to the store are block transfer reads (`BTR`). For each type of instruction, somewhat simplified, the programmer states:

(a) The direction of transfer (`BTW` or `BTR`).
(b) The card address (rack, slot and slot half, left or right).
(c) The store location start address where the data is to be received.
(d) The number of 16 bit words to be transferred.

The analog input card uses both `BTW` and `BTR` instructions, the `BTW` being used once, after power up, to configure the module and the `BTR`s subsequently to read the data as summarised in *Figure 16.63*.

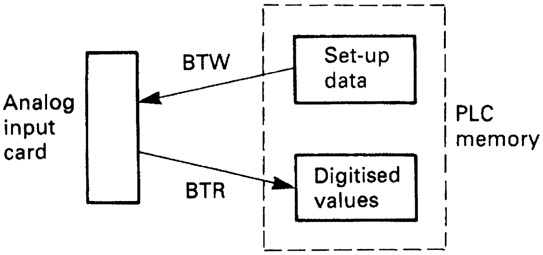


Figure 16.63 The PLC-5 block transfer write (BTW) and block transfer read (BTR) instructions

The post power up `BTW` sets how the module is to behave; whether it gives data in binary or BCD and the minimum and maximum values for the input range (H_R and L_R in Equation 16.1) on each channel. The card uses these to return readings in engineering units (in 12 bit binary integer or two's complement format or 12 bit BCD).

Once set up, values can be read at the required time intervals with a `BTR`. This gives signal values in the specified store locations along with over-range and similar alarms. The values can then be used elsewhere in the program.

PLCs are often required to provide analog output signals. Like analog inputs, these signals have standard voltage ranges of 1–5 V or 0–10 V or the current range of 4–20 mA.

A typical analog output card, for example, is the Allen Bradley 1771-OFE which has four output channels, each turning a 12 bit (0–4095) digital signal into an analog output. Isolation amplifiers are used on the outputs to reduce the effects of noise and allow the signals to connect into external devices fed from different electrical supplies. The digital signals come from storage locations inside the PLC as shown on *Figure 16.64*. This conversion is known as *Digital to Analog conversion,* or DAC.

For best resolution the PLC should use the full 0–4095 range, but this is frequently impossible. If the PLC, for example, is setting the speed range of a motor from 0–1350 rpm, it will need to convert 0–1350 into the range 4–20 mA. Equation 16.1 can be re-arranged as

$$X = \frac{4095(N - L_R)}{H_R - L_R} \qquad (16.2)$$

where X is the value passed to the DAC (in the range 0–4095). N is the output number from the PLC in engineering units, and H_R/L_R are the high and low range values. As before, great care must be taken with Equation 16.2 to avoid overspill or loss of resolution.

The PLC-5 communicates with the 1771-OFE with the `BTW` instruction described previously. The programmer sets up a block of twelve words, the first four of which contain the values, and the balance the set up data such as H_R and L_R. The block of data is then written to the card with a `BTW`. *Figure 16.65* shows a typical example where an analog speed reference can be raised or lowered by operator controlled pushbuttons. Note the use of greater than (`GTR`) and less than (`LES`) instructions to confine the counter value within the allowed range of 0–1350 rpm.

Ranging as above allows engineering units to be used inside the program, the counter in *Figure 16.65*, for example, holds the required speed directly in rpm, but this is accompanied by a loss of resolution as explained earlier. For the range 0–1350 rpm, we have a resolution of about 0.1%, compared with the theoretical 0.025% resolution available from the card.

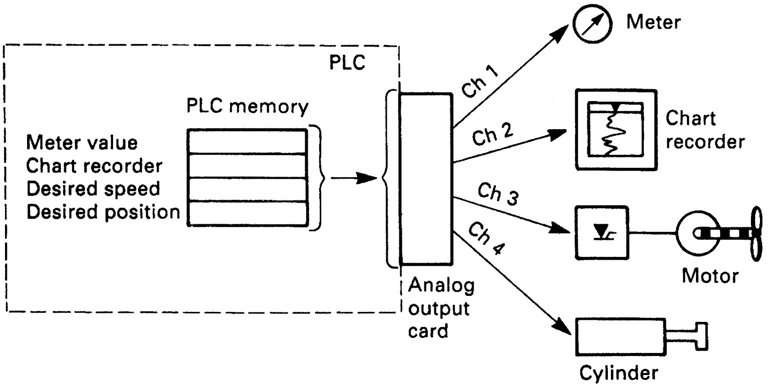

PLC
PLC memory
Meter value
Chart recorder
Desired speed
Desired position
Ch 1
Ch 2
Ch 3
Ch 4
Meter
Chart recorder
Motor
Analog output card
Cylinder

Figure 16.64 Analog output signals

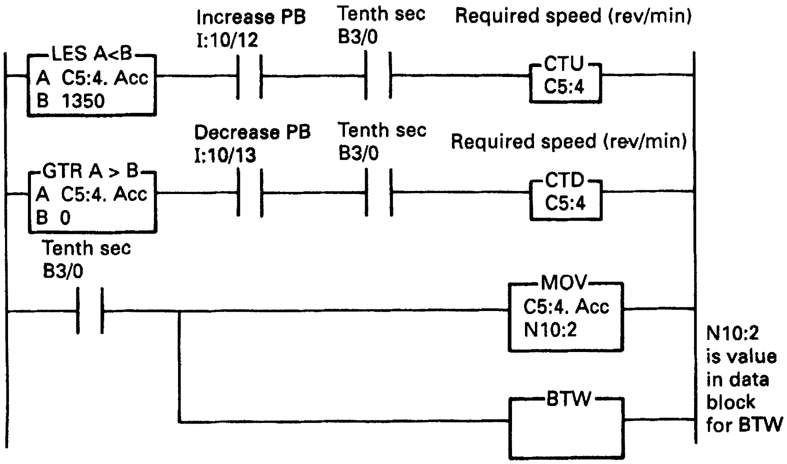

Increase PB
I:10/12
Tenth sec
B3/0
Required speed (rev/min)
LES A<B
A C5:4. Acc
B 1350
CTU
C5:4
Decrease PB
I:10/13
Tenth sec
B3/0
Required speed (rev/min)
GTR A > B
A C5:4. Acc
B 0
CTD
C5:4
Tenth sec
B3/0
MOV
C5:4. Acc
N10:2
BTW
N10:2 is value in data block for BTW

Figure 16.65 Setting the speed for a motor with a counter and an analog output card

There are other operations that can be performed on analog signals. A typical list, for the GEM-80, is

SQRT	Square root, mainly used with signals from orifice plates.
LINCON	Performs $X*(A/B)+C$ with limiting.
FGEN	Multipoint straight-line function generator used for linearisation as *Figure 16.66(a)*.
LIMIT	Performs limiting of signals as shown on *Figure 16.66(b)*.
RAMP	Rate limiting (with different rise and fall rates).
DEDBAND	Deadband functions as *Figure 16.66(c)*. Useful for preventing 'dither' in closed loop control when PV and SP are close.
ANALAG	First Order lag. Used for filtering.

A simple first order filter can be produced by any PLC which supports floating point numbers using the procedure shown on *Figure 16.67(a)*. This procedure uses just three rungs or three function blocks and is obeyed for one program scan at regular time intervals Δt. V_i is the raw input signal and V_f is the filtered output signal. $V_{f(n-1)}$ is the filtered value obtained on the previous execution Δt seconds ago. The error between V_i and $V_{f(n-1)}$ is calculated (V_e) then multiplied by a gain K to give a change V_c. This is added to $V_{f(n-1)}$ to give the new filtered value V_f. *Figure 16.67(b)* shows the response for a step change in V_i with K set at 0.25. At each execution of the routine V_f moves 25% of the difference between V_i and $V_{f(n-1)}$. The gain, K, determines the apparent time constant and must be in the range $0 < K < 1$. The gain K should be set to Δt/T where T is the required time constant.

16.4.7 Closed loop control

A closed loop system based on PLCs will be similar to *Figure 16.68*. The plant variable, P_V, is read by an analog input card, and the output O_P provided by analog output cards. The setpoint, SP, is provided by the operator or by some program sequence. The PID algorithm is then provided by the program. Chapter 13 gives more detail on the theory of closed loop control and an explanation of the PID algorithm.

It is possible to write PID algorithms with four function (+ – */) mathematics, but it needs great care. The program scan time must be known for the integral and derivative routines, and protection against output actuator saturation must be built in to overcome an effect called integral wind-up.

PLCs, are becoming increasingly powerful, and most medium range PLCs now provide a three term PID function in

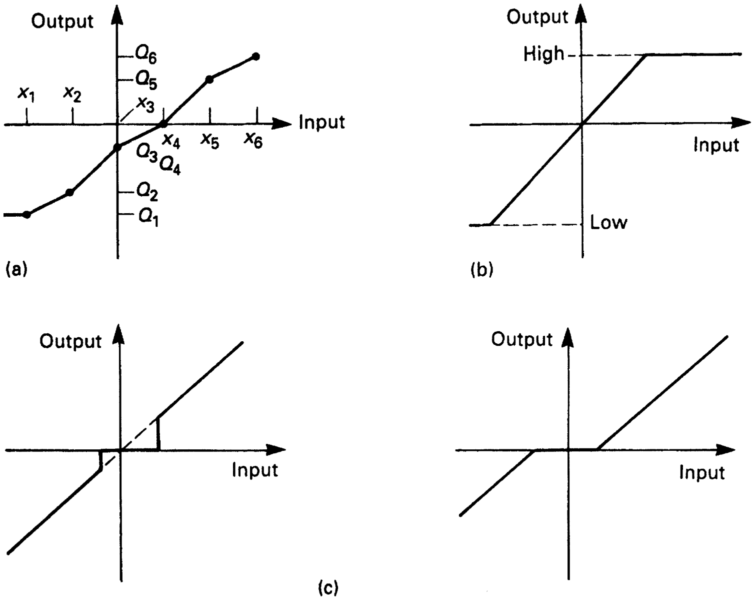


Figure 16.66 GEM-80 special functions for use with analog signals: (a) FGEN with *N* points at equal intervals *x*; (b) LIMIT, high and low limits can be different; (c) DEDBAND without and with offset

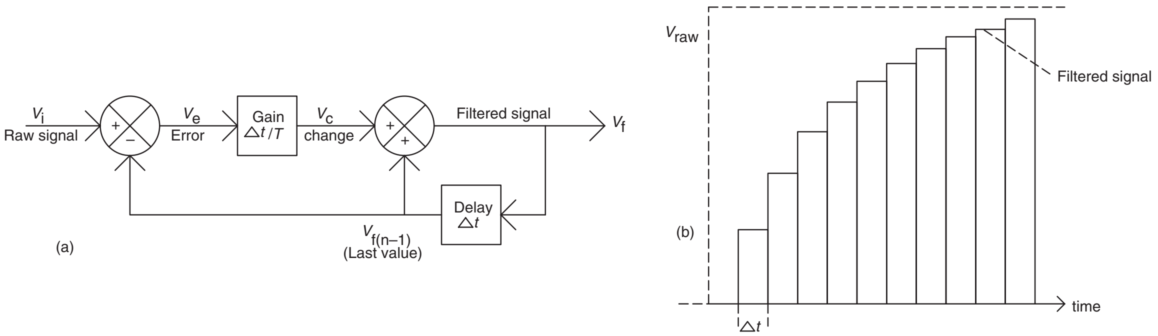


Figure 16.67 Programming a first order filter: (a) schematic diagram; (b) response to step input

their instruction set. *Figure 16.69* shows a ratio temperature control program written for an Allen Bradley PLC-5 processor.

These three rungs are controlling the temperature in a furnace, with the temperature PID block controlling the air valve. The air flow is measured, multiplied by the required ratio and used as the setpoint for the gas PID block.

The control blocks in each PID instruction hold the data and working areas for the PID function; things like auto/man status and the sum for the integral action. The setpoint is written directly into the third word of the control block. The process variable is the feedback signal from the variable being controlled, usually obtained from an analog input card. Settings for gain, T_i and T_d are also contained in the control block data.

A three term control algorithm can suffer from integral wind-up in saturation or manual operation. The tieback variable is used to give the current value corresponding to the driven actuator output (possibly after auto/manual changeover) and is used to prevent wind-up and to give bumpless transfer. The control variable is the signal from the PID algorithm, usually sent to an analog output card via auto/manual changeover logic.

The three rungs of *Figure 16.69* mask, to some extent, the work that must be done elsewhere in the program. Data from the outside world must be obtained with analog input cards, and the controller output(s) must be written to the actuators with analog output cards. The timing of these reads and writes must be regular and linked to the PID instructions.

Auto/manual changeover logic will also be required, linked into the PID instructions with the tieback variable and the auto/man status flag (which makes the integral term track the tieback in manual).

The operator will also require a link to the control, so pushbuttons, displays and alarms must be provided. All of this is in addition to the basic PID control.

16.4.8 Intelligent modules

We have so far considered analog input and output modules, which are semi-intelligent (compared to 'dumb' digital

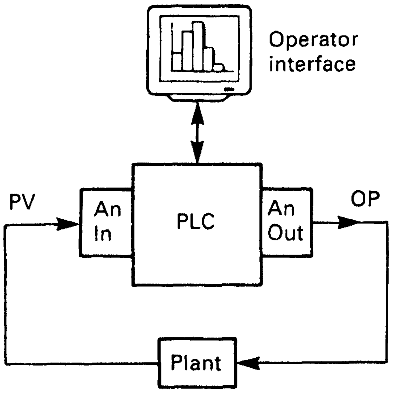


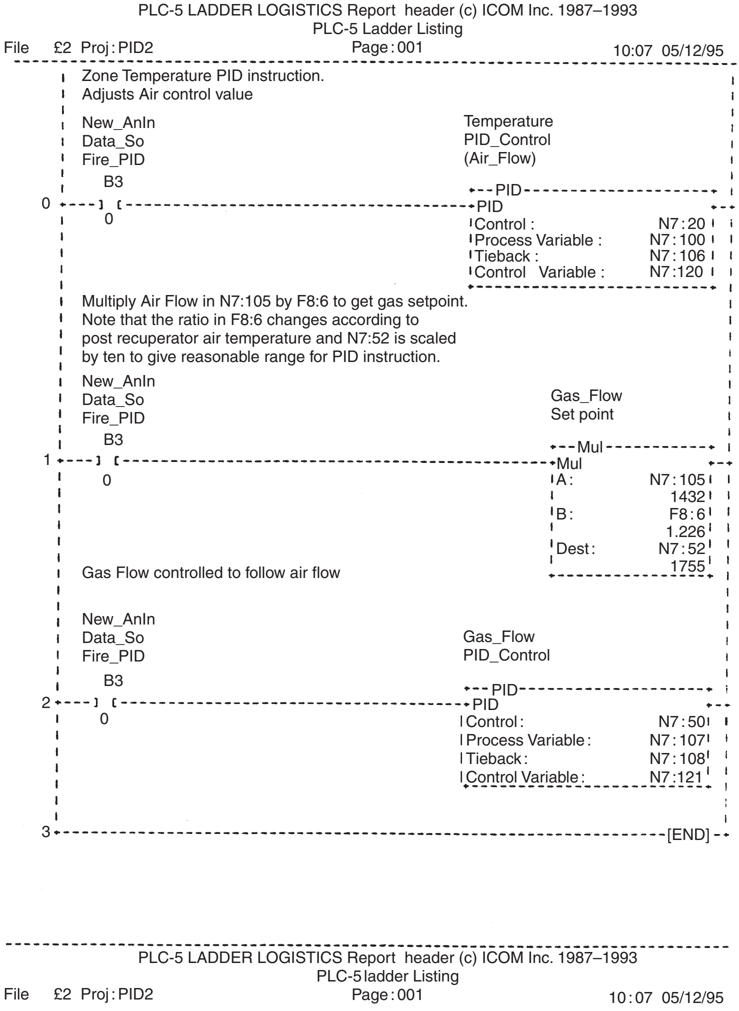
Figure 16.68 Closed loop control with a PLC

input and output cards). These are examples of a more general range of intelligent modules which most manufacturers offer to simplify the designers task.

A typical example is a high speed counter. We saw earlier in Section 16.2.4 that the scan time limits the maximum count rate of a PLC to about 10 Hz. High speed counter cards are available for use where higher count speeds are needed, or the program scan time introduces an unacceptable random error.

In these, the card contains a bi-directional counter which can be directly driven by a pulse encoder. The counter value can be loaded from the PLC, and read back when needed. The PLC can also download a preset value, allowing the counter card to directly drive outputs according to the relationship between the count and the preset.

Other common intelligent modules are bar code readers (for stock tracking), stepper motor controllers (for position control systems) and vision modules (for quality control applications).



```
                PLC-5 LADDER LOGISTICS Report  header (c) ICOM Inc. 1987–1993
                                    PLC-5 Ladder Listing
File   £2  Proj:PID2                      Page:001                         10:07  05/12/95
-------------------------------------------------------------------------------------------
      | Zone Temperature PID instruction.                                               |
      | Adjusts Air control value                                                       |
      |                                                                                 |
      | New_AnIn                                  Temperature                           |
      | Data_So                                   PID_Control                           |
      | Fire_PID                                  (Air_Flow)                            |
      |   B3                                                                            |
      |                                           +--PID--------------------+           |
    0 +---] [-------------------------------------+PID                      +--+
      |   0                                       |Control :           N7:20 |          |
      |                                           |Process Variable :  N7:100 |         |
      |                                           |Tieback :           N7:106 |         |
      |                                           |Control  Variable : N7:120 |         |
      |                                           +--------------------------+          |
      | Multiply Air Flow in N7:105 by F8:6 to get gas setpoint.                        |
      | Note that the ratio in F8:6 changes according to                                |
      | post recuperator air temperature and N7:52 is scaled                            |
      | by ten to give reasonable range for PID instruction.                            |
      | New_AnIn                                                                        |
      | Data_So                                              Gas_Flow                   |
      | Fire_PID                                             Set point                  |
      |   B3                                                                            |
      |                                                      +--Mul-----------+         |
    1 +---] [------------------------------------------------+Mul             +--+
      |   0                                                  |A:        N7:105 |        |
      |                                                      |            1432 |        |
      |                                                      |B:          F8:6 |        |
      |                                                      |           1.226 |        |
      |                                                      |Dest:       N7:52 |       |
      | Gas Flow controlled to follow air flow               |            1755 |        |
      |                                                      +-----------------+        |
      |                                                                                 |
      | New_AnIn                                                                        |
      | Data_So                                   Gas_Flow                              |
      | Fire_PID                                  PID_Control                           |
      |   B3                                                                            |
      |                                           +--PID--------------------+           |
    2 +---] [-------------------------------------+PID                      +--+
      |   0                                       |Control:            N7:50 |          |
      |                                           |Process Variable:   N7:107 |         |
      |                                           |Tieback:            N7:108 |         |
      |                                           |Control Variable:   N7:121 |         |
      |                                           +--------------------------+          |
      |                                                                                 |
    3 +--------------------------------------------------------------------[END]--+



-------------------------------------------------------------------------------------------
                PLC-5 LADDER LOGISTICS Report  header (c) ICOM Inc. 1987–1993
                                    PLC-5 ladder Listing
File   £2  Proj:PID2                      Page:001                         10:07  05/12/95
```

Figure 16.69 PID control on an Allen Bradley PLC-5

16.5 Distributed systems and fieldbus

16.5.1 Introduction

For a true distributed control system we need a method where several PLCs or computers can be linked together to allow communication to freely take place between any member of the system.

To achieve this we need to establish a connection topology, some way of sharing the common network that prevents time wasting contention and an address system that allows messages to be sent from one member to another. Such systems are known as *Local Area Networks* (LAN) or *Wide Area Networks* (WAN) dependent on the size of the area and the number of stations.

16.5.2 Transmission lines

Any network will be based, to some extent, on cable, and at the high speeds used there are aspects of transmission line theory that need to be considered. Consider the simple circuit of *Figure 16.70(a)*. At the instance that the switch closes, the source voltage does not know the value of the load at the far end of the line. The initial current step, *i*, is therefore determined not by the load, but by the characteristics of the cable (dependent on the inductance and capacitance per unit length). A line therefore has a *characteristic impedance*, typically 75 Ω or 50 Ω for coax, and 120 to 150 Ω for biaxial or screened twisted pair. The initial current step will therefore be V/Z where Z is the characteristic impedance.

After a finite time, this current step reaches the load R, and produces a voltage step $i \times R$. If R is not the same as Z, this voltage step will not be the same as V, and a reflection will result. Typical results are shown on *Figure 16.70(b)*.

This effect occurs on all cables and is normally of no concern as the reflections only persist for a short time. If, however, the propagation delay down the line is similar to the maximum frequency rate of the signal, the reflections can cause problems. It follows that a transmission line should be terminated by a resistance equal to the characteristic impedance of the line. Normally, devices for connecting onto a transmission line have a high input impedance to allow them to tap in anywhere, with terminating resistors being used at the ends of the line.

A side effect of this is that T connections, or spurs, are not allowed (unless the length of the spur is short). In *Figure 16.70(c)* a T has been formed. To the signal, coming from the left, the two legs appear in parallel giving an apparent impedance of $Z/2$ and a reflection.

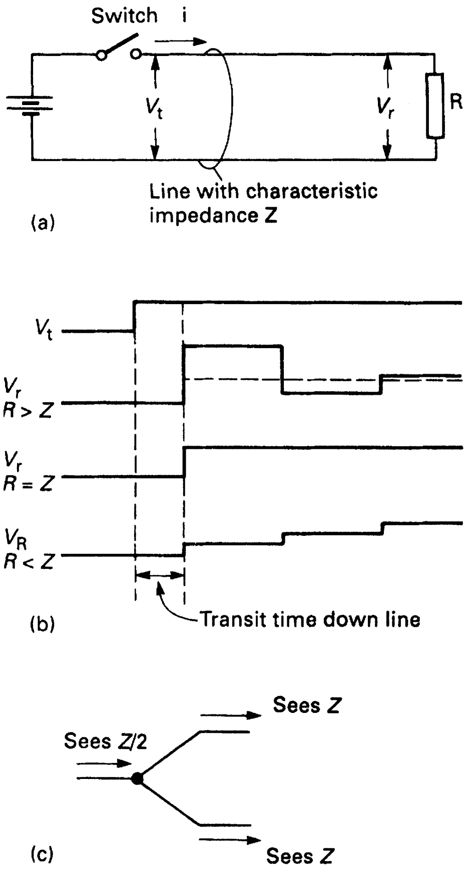


Figure 16.70 Transmission lines and the characteristic impedance: (a) a transmission line; (b) the effect of the terminating resistor; (c) The effect of a 'T' in the line

16.5.3 Network topologies

From the previous section it should be apparent that any network can sensibly only be based on a ring (which needs no terminating resistors) or a line (with a terminating resistor at each end). *Figure 16.71* is a master/slave system where a common master wishes to receive or send data from/to slave devices, but the slaves never wish to talk to each other. All the slaves have addresses, which allows the master to issue commands such as '*Station 3; give me the value of analogue input 4*' or '*Station 14; your setpoint is 751.2*'. Such systems are often based on RS422 to provide improved noise immunity and allow longer lengths of line.

The Star network of *Figure 16.72* is again based on a master with a point to point link to individual stations. This arrangement is commonly used for high level computer

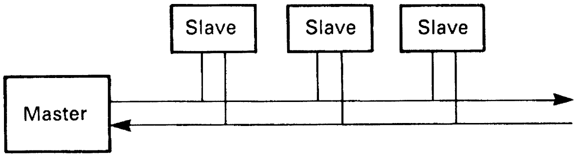


Figure 16.71 A Master/slave network

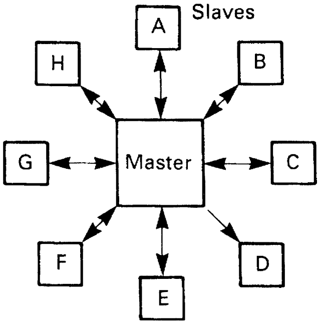


Figure 16.72 A Star network

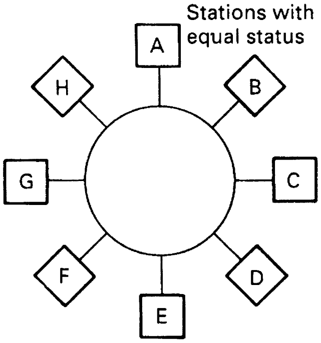


Figure 16.73 A masterless peer to peer or ring network

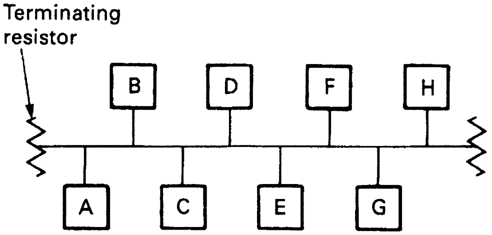


Figure 16.74 A peer to peer link arranged as a single line bus with terminating resistors

systems. Communication control is performed by the master station. Station to station communication is possible via, and with the co-operation of, the master.

In *Figure 16.73* all the stations have been connected in a ring. There is no master, and all stations can talk to any other station and all have equal right of access. The term *peer to peer link* is often used for this arrangement. With *Figures 16.71* and *16.72* control was firmly in the hands of the master. With the ring, some technique is needed to avoid clashes when two stations wish to use the line at the same time. We will discuss this in the following section.

Figure 16.74 is probably the commonest type of network used by PLCs. It is a single line with terminating resistors and, like the ring, is a peer to peer link where all stations have equal standing.

16.5.4 Network sharing

A peer to peer link allows many stations to use the same network. Inevitably two stations will want to communicate at the same time. If no precautions are taken, the result will be chaos. Various methods are used to govern access to the network.

One idea is to allocate time slots into which each station can put its messages. This is known as *Time Division Multiplexing*, or TDM. Whilst it prevents clashes, it can be inefficient as a station will have to wait for its time slot even if no other station has a message to send. To some extent a mismatch between the frequency of messages from different stations can be overcome by giving more slots to hardworking stations. This is sometimes known as *Statistical TDM*.

The *empty time slot* of *Figure 16.75* uses a packet which continuously circulates around the ring. When a station wishes to send a message it waits for the empty slot to come round, when it adds its message. In *Figure 16.75*, station A wishes to send a message to station D. It waits until the empty packet comes round. Then it puts its message onto the network along with the destination address D. Stations B and C pass the message but ignore it because it is not for their address. Station D matches the address, reads the contents (and appends that it has received the message). Stations E–H ignore it, but pass it on. Station A receives the message back again, sees the acknowledgement and removes its message leaving the empty packet circulating the ring again. A similar idea is a *token passing*, where a ‘Permit to Send’ token circulates round the network. A station can only transmit when it is in possession of the token, which is released when the acknowledgement that the message arrived is received.

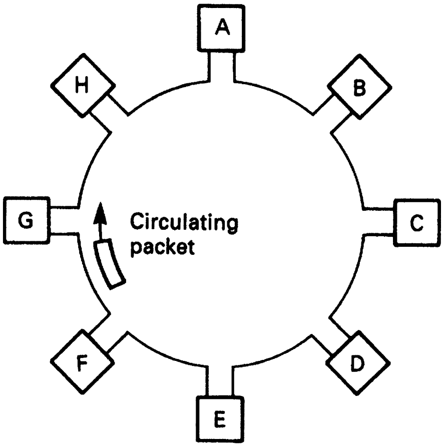


Figure 16.75 Empty slot and token passing network

Bus systems usually employ a method where a station wishing to send a message listens to the network to see if it is in use. If it is, the station waits. If the network is free, the station sends its message (thereby locking out any other station until the message ends). This is known as *Carrier Sense Multiple Access* (CSMA).

Situations can still arise, however, where two stations simultaneously start to send a message, and a collision (and garbage) results. This situation can easily be detected, and both stations then stop and wait for a random time before trying again. A random time is used to stop the two stations clashing again. This is known as *Carrier Sense Multiple Access with Collision Detection* (or much more simply as CSMA/CD).

There is a fundamental difference between TDM, empty slot, and token passing as one group and CSMA. With the former there is a certain amount of time wasting, but every station is guaranteed access within a specified time. With CSMA there is little time wasting, but a station can, in theory, suffer repeated collisions and never get access at all.

A useful analogy is to consider motor car traffic control. TDM/token passing approximates to traffic lights, CSMA to roundabouts. In heavy traffic the best solution is traffic lights; everyone gets through and the waiting is shared evenly. Roundabouts can ‘lock out’ one road when the traffic flow is heavy and uneven from one direction. In light traffic, however, roundabouts keep the traffic flowing smoothly.

16.5.5 A communication hierarchy

Early process control systems tended to be based on a single large computer or PLC. The advent of cheap PLCs with good

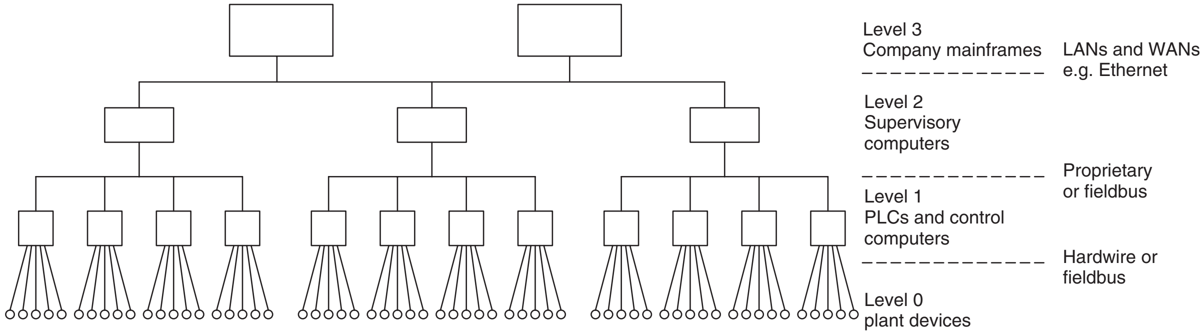


Figure 16.76 A simple communication hierarchy

communications has led to the development of a hierarchy of machines which split the tasks between them. Such an arrangment is called a *Distributed Control System* or DCS.

Such a system is generally arranged as *Figure 16.76* with a hierarchy split into four levels.

Level 0 is the actual plant, with devices linking to the next level by direct wiring or simple RS232/422 serial links.

Level 1 is the level the majority of this chapter is concerned with, consisting of PLCs and small computers directly controlling the plant.

Level 2 is supervisory computers for large areas of plants.

Level 3 is the large company mainframe.

Usually the layout is not as clear cut as this implies. There are also differences between different companies, some number the layers from top to bottom and some ignore level 0.

There are many advantages to distributed systems. The resulting tree is conceptually simple, and as such is easy to design, commission, maintain and modify. A correctly designed system will be, for short periods, fault tolerant and can cope in a limited mode with the failure of individual stations. A distributed system can also bring about an increase in performance as lower level machines take the work off higher level machines.

16.5.6 Proprietary systems

In this section we will look at a typical proprietary system used to link PLCs from the same manufacturer. Typical examples are the GEM-80s Coronet and ESP, Siemen's Sinec and Modicon's Modbus. For reasons of space we shall consider how machine to machine links are achieved with Allen Bradley PLC-5s which communicate with each other on a peer to peer (no master) token passing highway based on twinaxial cable and operating at 57.6 Kbaud. Their trade name is Data Highway Plus. The PLC stations' addresses are set on switches in each PLC, and up to 64 stations can exist on one line with octal addresses 0–77.

Communication is established with a single message (MSG) instruction. This can be set up to read or write a block of data, the programmer specifying:

(1) The start address at the local end;
(2) The start address at the target end;
(3) The length of the block to be transferred (in words); and
(4) The station address at the remote end.

The MSG instruction appears in a program as *Figure 16.77(a)*, the transfer being initiated every time the rung goes true. The ENable bit goes true when the transfer is started, and the DoNe bit goes true when it has been successfully completed. The ERRor flag goes true when an error occurs. Common errors are a line fault, a non existent address at the far end or the PLC at the far end shutdown. The cause of the fault is given in flags set in the message control word. Link statistics (e.g. number of retries) are kept in the processor for diagnostic purposes.

The details of the MSG instruction are set up by the programmer via the screen of *Figure 16.77(b)*. These are mostly self explanatory, with the possible exception of the remote link which is concerned with sending data via a gateway module to a different highway, possibly of a different type.

The data highway is also used by the programming terminal, so a programmer can connect anywhere onto the data highway and link into any machine on the network.

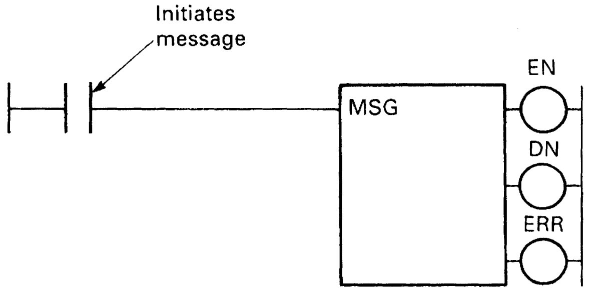
16.5.7 Ethernet

Ethernet is a very popular bus based LAN originated by DEC, Xerox and Intel and is commonly used to link the computers at level 2 in *Figure 16.76*. It uses 50 Ω coaxial cable, with a maximum cable length of 500 m (although this can be extended with repeaters). Up to 1024 stations can be accommodated, although in practical systems the number is far lower. Baseband (i.e. non modulated) signalling is used with CSMA/CD access control. The raw data rate is 10 Mbaud, giving very fast response at loading levels up to about 20–30% of the theoretical maximum. Beyond this, collisions start to occur.

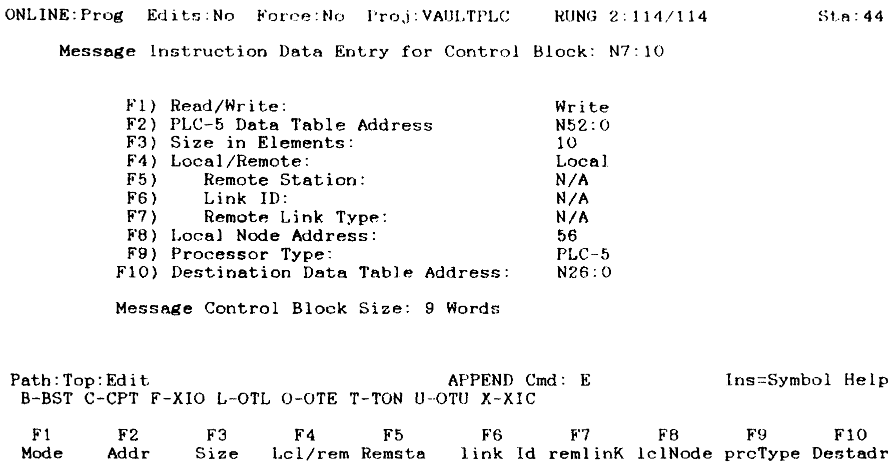
Because Ethernet uses CSMA/CD the successful transmission of a message cannot be guaranteed. It is possible, (but unlikely) for a given message to continually suffer from crashes and never get access to the network. In the jargon ethernet is '*non determinstic*'. In practice, if the network loading is kept below 30% of its theoretical capacity this is not a problem. Many PLCs (such as the PLC-5/40E) now can provide direct connection to an Ethernet network.

16.5.8 Towards standardisation

We have already discussed the difficulties of linking different equipment. There is normally little problem linking PLC networks to higher level computers. PLC manufacturers publish their message format and protocols, and interfacing software (called 'drivers') has been written for all common computers and PLCs. The difficulty comes when you want to link two machines from different manufacturers at level 1 in *Figure 16.76*. In many cases, the only economical solution is to do it through the computers and the higher level link.



(a)



```
ONLINE:Prog   Edits:No  Force:No  Proj:VAULTPLC      RUNG 2:114/114            Sta:44

      Message Instruction Data Entry for Control Block: N7:10

              F1) Read/Write:                           Write
              F2) PLC-5 Data Table Address              N52:0
              F3) Size in Elements:                     10
              F4) Local/Remote:                         Local
              F5)    Remote Station:                    N/A
              F6)    Link ID:                           N/A
              F7)    Remote Link Type:                  N/A
              F8) Local Node Address:                   56
              F9) Processor Type:                       PLC-5
             F10) Destination Data Table Address:       N26:0

             Message Control Block Size: 9 Words


Path:Top:Edit                            APPEND Cmd: E               Ins=Symbol Help
 B-BST C-CPT F-XIO L-OTL O-OTE T-TON U-OTU X-XIC

  F1      F2      F3      F4      F5       F6       F7       F8       F9      F10
 Mode    Addr    Size   Lcl/rem Remsta   link Id  remlinK  lclNode prcType Destadr
```

(b)

Figure 16.77 The PLC-5 Message (MSG) Instruction: (a) as written in the ladder diagram; (b) as seen in detail on the programming terminal

General Motors (GM) in the USA were faced with this problem and attempted to specify a LAN for industrial control. This was called MAP (Manufacturing Automation Protocol). A similar office based LAN called TOP (Technical Office Protocol) was conceived at the same time. With GM's purchasing muscle, it involved several automation equipment manufacturers. A firm commitment to the OSI model was made, and the network based on broadband token bus as specified in IEEE 802.4 was chosen as it is deterministic.

MAP, however, has not been widely adopted. There appears to be several reasons for this distinct lack of enthusiasm. The first is a bureaucratic organisation and a changing specification. The second reason is cost; MAP links often cost more than the PLC to which they are connected. The third reason is speed; by using token passing MAP is slow by comparison with other standards. The final, and perhaps most crucial, fact is that MAP seems to have settled at a level where it is in direct competition with established LANs such as Ethernet rather than the proprietary systems at level 1 of *Figure 16.76.*

A standardised fieldbus system would allow PLCs, sensors and actuators to be connected and communicate with minimal cost. Unfortunately at the time of writing (early 2002) a standard seems as far away as ever with progress being slowed by commercial and national infighting.

Profibus is one the more common fieldbus contenders, largely because it has been adopted by Siemens and many other German electrical companies. There are three versions of Profibus designed for three different application areas. All use token passing.

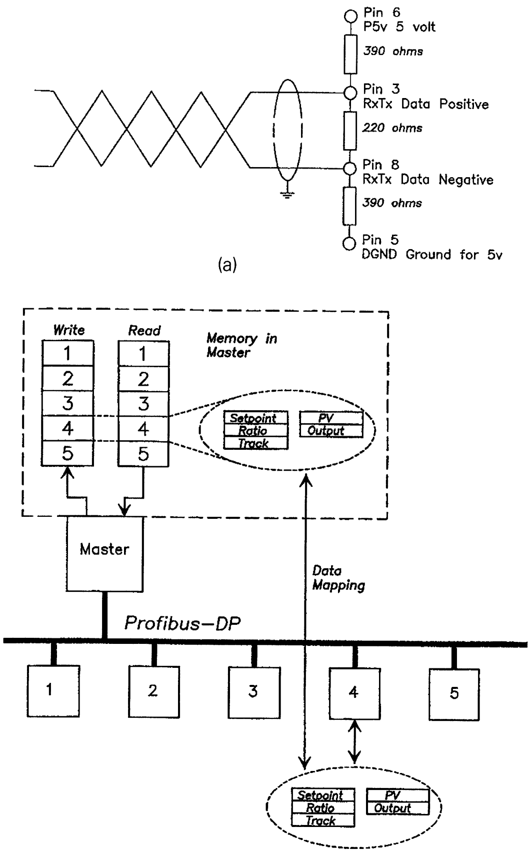
The first, called Profibus-DP, for decentralised periphery, is by far the commonest and is designed to link intelligent masters (e.g. a PLC), to slave devices such as sensors, drives or actuators. Profibus only uses levels 1 and 2 of the ISO/OSI model. Twisted pair RS485 or fibre optics are used for transmission.

The second, Profibus-FMS, for field message specification, is designed for the higher level with multiple masters and allowing peer to peer communication. Levels 1, 2 and 7 of the ISO/OSI model are used and RS485 or fibre optics for transmission.

Both DP and FMS share the same transmission standards and can consequently work together on the same network.

The final form, designed for process automation in hazardous areas, is Profibus-PA which permits the construction of an intrinsically safe network. Profibus-PA uses slightly different standards to DP and FMS, but can be linked by a segment coupler device.

All are a linear bus system, i.e. a straight line. Transmission speeds from 9.6 kbit/s (up to 1200 m) to 12 Mbit/s (up to 100 m) can be used. Screened twisted pair is used, with terminating resistors at each end of the bus. Up to 32 stations can be used in each segment, each with a unique station address. Segments can be coupled with segment repeaters, allowing a total of 127 stations to be



(a)

(b)

Figure 16.78 Profibus-DP network: (a) connection at a Profibus device. Non terminating devices use only pins 3 and 8; (b) mapping between a Profibus device and memory in the network master

addressed. Addresses are assigned for global or group data reducing the number of messages and time lag problems when data for several devices are to be changed together.

Connections to masters or slaves are made via standard 9 pin D-type connectors, as shown on *Figure 16.78(a)*. Terminating resistors are either switched in internally at the end stations or connected inside the final plugs. Note that the terminating resistors require power, this normally comes from the end stations themselves.

The manufacturer of each device on the network, e.g. a VF drive, provides a disc file, called the GSD, which is a description of the data exchange the device can support (e.g. accepting speed reference and run command and providing load current and drive state etc.) plus operating parameters such as supported transmission speeds. Included in the GSD file is a unique identification number assigned by the Profibus User Organisation. The GSD files for all the devices on the network are used along with the station addresses to build a network description which is held in the master.

Because Profibus-DP only uses levels 1 and 2, the data exchange maps onto pre-determined areas in the master controller (usually a PLC) as shown on *Figure 16.78(b)*. To change the speed of the drive, the user simply writes the new speed into the mapped area, and the data is transferred with no further action. In a similar manner, slave data and status is automatically read from the mapped area. A Profibus-DP network is thus totally transparent to the user.

A typical example of the problems that any attempts to specify a standardised fieldbus system may encounter is the continual introduction of new ideas. All the communications systems described so far are based on what is called the *source/destination model.* If station A has information for station B, a message is sent with the format:

`Source A | Destination B | Data`

If this information is to be sent to several stations, each will need their own message. In applications where multiple setpoints have to be sent to multiple controllers, the delay caused by the time shift between the messages can cause problems, although this can be overcome to some extent by the use of group or global addresses as used by Profibus.

In addition, if station A needs information from station B, (the state of an interlock for example), station A must perform a read on each occasion the data is required.

A recent development, called the *producer/consumer model*, uses a different approach. Here data is placed onto the network with no indication as to who it is for. The format is now simply:

`Identifier | Data`

All stations using this data accept it at the same time, eliminating the need for multiple messages. This significantly reduces the number of messages and hence increases the network speed.

The placement of data onto the network can be done in two ways. The first, and fastest, is '*notify on change*'. Here a station only places information on the network when a new value is different than the old. Stations with an interest in this data assume that the status or value remains the same until notified otherwise. There are obvious dangers in this, and a regular pre-defined 'heartbeat' is included to say a station is active on the network. The second approach updates on a time basis, each data item having its own, or a global, update time.

At the time of writing, Foundation Fieldbus is the only producer/consumer fieldbus network, and Rockwell (Allen Bradley) have also adopted the method for their proprietary ControlNet. The latter is interesting as it combines the ideas of their remote I/O and Data Highway onto one system and allows PLC racks, (and their data), to be shared equally amongst several processors and not dedicated to one as before.

16.6 Graphics

Operator controls are being increasingly provided by computer graphic screens. These can be a display device designed specifically for a particular range of PLCs (for example the Allen Bradley Panelview and the CEGELEC Imagem) and general purpose graphic display devices (such as ABB/ASEA's excellent Tesselator) or graphics software running on conventional personal computers. *Figure 16.79* shows some typical examples.

The major advantages are simplicity of installation and flexibility. A graphics terminal has just two connections to the outside world, a serial link connection and a power supply. If it is used to replace a desk full of switches and indicators there are obvious cost savings.

The designer of desks or control stations often has to deal with changes and modifications. Constructing a desk is

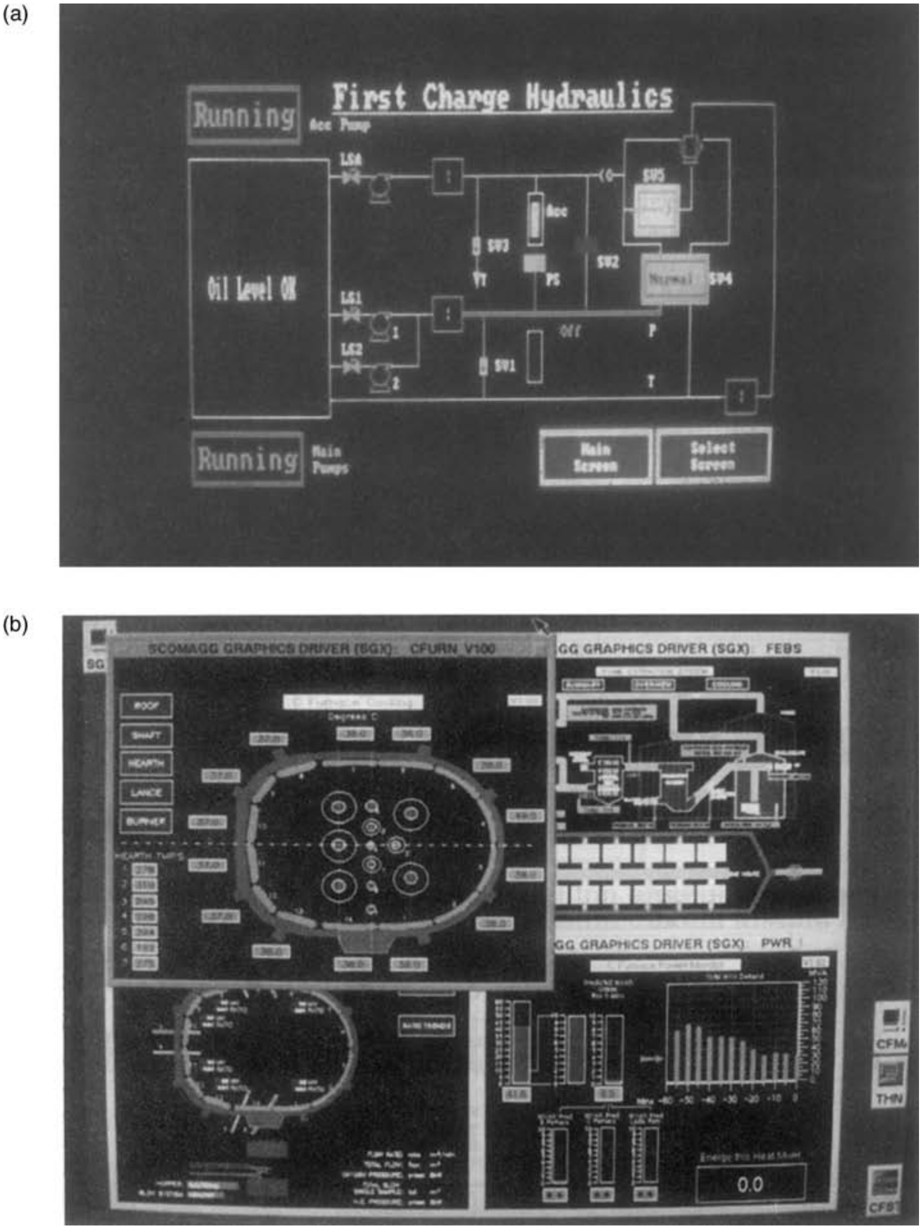

(a)
First Charge Hydraulics
Running
Acc Pump
LSA
SV5
Acc
SV3
SV2
Oil Level OK
YT
PS
Normal
SV4
LS1
1
Off
P
LS2
2
SV1
T
Running
Main Pumps
Main Screen
Select Screen
(b)
SCOMAGG GRAPHICS DRIVER (SGX) CFURN V100
GG GRAPHICS DRIVER (SGX) FEBS
ROOF
SHAFT
HEARTH
LANCE
BURNER
HEARTH TMPS
GG GRAPHICS DRIVER (SGX) PWR
0.0
SG
CFM
THN
CFS

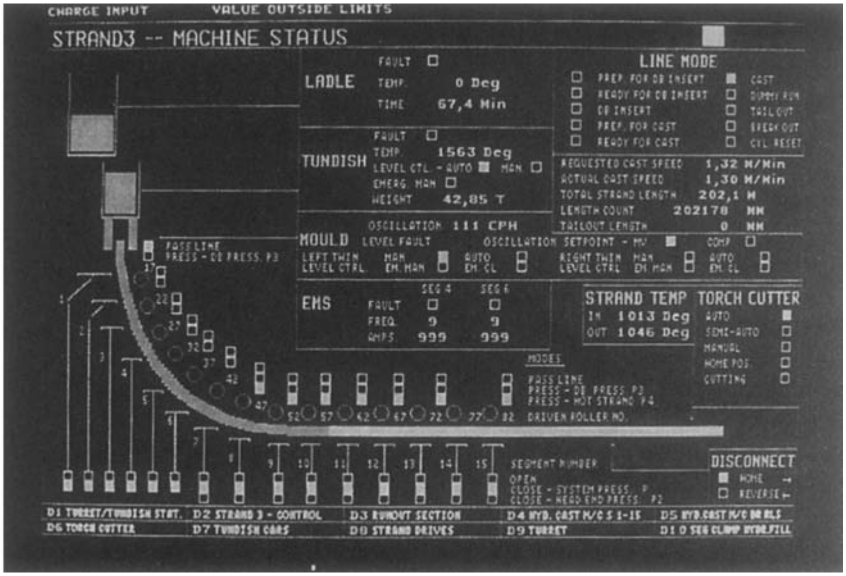
Figure 16.79 Various graphic displays: (a) Allen Bradley touchscreen Panelview using block graphics; (b) high resolution Scada system; (c) the ABB Tesselator. *Photos courtesy of Co-Steel Sheerness, Scomagg and ABB*

always a fine balance of time, choosing between waiting until all the requirements are clear, and the minimum time needed to make it. Modifications at the commissioning stage rarely look neat. The displays on a graphical terminal can be modified relatively easily, and, more importantly, the modifications leave no scars. If the design of a normal desk can only start when the desk contents are 95% finalised (which is about right) a graphic screen can be started at 75% finalised. This flexibility is of great assistance as no job is ever right first time.

There are disadvantages, though. The most important of these is the limited amount of information that can be displayed on a single screen. It is very easy to overcrowd a screen (giving a screen similar to a page full of text on a word processor) making it difficult for the operator to identify critical items. A useful rule of thumb is not to use more than 25–30% of the screen.

The effect of this is often a need to build up a hierachy of screens; the top screen showing an overview, lower screens showing more and more detail. The problem with this is the time delay needed to shift through the screens. Direct screen to screen movement is possible by calling for a page number (which needs a good human operator memory, or a directory piece of paper, or wasted screen space) or by making all

(c)



CHARGE INPUT VALUE OUTSIDE LIMITS

STRAND3 -- MACHINE STATUS

LADLE
FAULT ☐
TEMP. 0 Deg
TIME 67,4 Min

TUNDISH
FAULT ☐
TEMP. 1563 Deg
LEVEL CTL. - AUTO ■ MAN ☐
EMERG. MAN ☐
WEIGHT 42,85 T

MOULD
OSCILLATION 111 CPH
LEVEL FAULT
OSCILLATION SETPOINT - MV ■ COMP ☐

EMS
SEG 4 SEG 6
FAULT ☐ ☐
FREQ. 9 9
AMPS 999 999

LINE MODE
☐ PREP. FOR DB INSERT ■ CAST
☐ READY FOR DB INSERT ☐ DUMMY RUN
☐ DB INSERT ☐ TAIL OUT
☐ PREP. FOR CAST ☐ BREAK OUT
☐ READY FOR CAST ☐ CYL RESET

REQUESTED CAST SPEED 1,32 M/Min
ACTUAL CAST SPEED 1,30 M/Min
TOTAL STRAND LENGTH 202,1 M
LENGTH COUNT 202178 MM
TAILOUT LENGTH 0 MM

STRAND TEMP
IN 1013 Deg
OUT 1046 Deg

TORCH CUTTER
AUTO ■
SEMI-AUTO ☐
MANUAL ☐
HOME POS. ☐
CUTTING ☐

DISCONNECT
■ HOME →
☐ REVERSE ←

Figure 16.79 (*continued*)

screen changes via an intermediate directory page (with additional delay). These time delays are small (less than a second typically) but the cumulative annoyance is large.

The time taken to update screen data can also be problematical, particularly where a machine to machine link is involved. Again a response time of around one second is typical, but several seconds is by no means uncommon. The use of a graphic terminal for fault finding on a fast moving plant is not really feasible.

There are generally two types of graphic terminal. The simplest, known as block graphics, has one store location for each character position on the screen and approximates to the old CGA standard on a personal computer.

The second type of display deals not with individual characters, but with individual points on the screen called 'pixels'. A typical medium resolution screen will have 640 (horizontal) by 480 (vertical) pixels, a total of 307 200 points. Each of these can be accessed individually, allowing lines to be drawn at any angle, fill patterns of any type to be used and trend graphs of plant variables to be displayed. Each individual pixel can have its own colour (from over 256 possible colours in some displays) and intensity. The result is an almost photographic resolution. There are additional costs, the most obvious of which is a large store requirement. The system hardware and software is more complex (and hence more expensive) but, perhaps surprisingly, this is not apparent to the designer. Programming for these screens is surprisingly simple with instructions using keywords like

`DRAW FROM < > to < >.`

or pick and place functions similar to a good commercial graphics package.

Supervisory Control and Data Acquisition (SCADA) systems are based around graphical objects which are linked to variables in the control systems. The state of the objects on the screen (e.g. size, colour, rotation, etc.) can be changed according to the the values of the variables in the control systems giving a very visual image of the operation.

The environment around a display needs to be carefully considered. Most screens are mounted angled up, and are prone to annoying reflections from overhead lights and windows. Bright lighting (and above all direct sunlight) can make a display impossible to read. Displays are also adversely affected by magnetic fields. Close proximity to electric motors, transformers or high current cables will cause a picture to wobble and the colours to change. The effect can be overcome by screening the monitor with a mu-metal cage). Flat screen TFT or LCD displays do not suffer from this effect.

The size and weight of the monitors are often overlooked making them difficult to mount neatly, and even more difficult to change. Access should be made as easy as possible; trying to hold a 25 kg display in place with one hand whilst undoing interminably long mounting screws is not much fun.

Displays fail, and the implication of this needs to be considered in the design. If all the plant control is performed by screens, what will happen during the ten or so minutes it will take to locate a spare and change the faulty unit? Often dual displays are used to overcome this problem.

The operator will obviously need to input data and initiate actions. Keyboards are one approach, but many people are nervous of them and the cable connecting the keyboard always seems prone to damage. In dirty environments keys can become blocked with dirt and membrane keypads with tactile (feel) feedback should be used.

If the operator has to access points anywhere on the screen, a tracker ball is a useful device. Rather like an upside down mouse it controls the movement of a cursor on the screen. All normal actions can be performed with three buttons on the trackerball and a numerical keypad. Trackerballs work surprisingly well in dirty environments as they are open underneath and dirt seems to fall straight through. Mice perform a similar function but are vulnerable to damage and dirt and seem more suited to an office environment.

A final consideration is security. Most modern graphics systems are based on good quality personal computers. These have value outside of industry and are vulnerable to theft. Often is it is not the PCs or screens which are stolen, but the internal motherboards and memory cards. Suitable security methods should be used if a PC based system is to be left unattended for a period of time (e.g. during a Christmas shutdown). Needless to say backups should not be stored on the same PC as the original system.

16.7 Software engineering

Any project goes through six stages during its life. The first of these, *analysis*, is studying the application to understand what is required. This is by far the most difficult stage as the project requirements are usually unclear. Most projects that come unstuck do so because this first stage has been cut short or overlooked.

Next comes *specification*, which is documenting the analysis so everyone concerned can agree what is to be done and what the end result should be. If you can't produce a specification, how can you sensibly design it? Never say 'we'll sort that out later' because later becomes 3 a.m. as the plant starts up. The final testing procedures must also be defined at the specification stage; again if you don't know how you will test it, how will you know if it's working properly? Defining testing procedures in the cold light of day several months before the final frantic rush to meet a deadline also helps the poor commissioning engineer to resist the pleas for a premature start up.

The importance of these two first stages cannot be overemphasised, too often the users do not know, or do not say, what they want, but once the project is complete they are sure it wasn't that. With these first two difficult stages over, the rest of the project becomes much easier!

The *design* stage can now start, (simple with a good specification) followed by *installation*. Next comes *commissioning*. These can also be difficult times, as in any project the control engineer ends up collecting everybody's delays and comes under pressure to 'get the plant away'. It is here that the advantage of the test schedule from the specification stage will be invaluable.

It is not generally understood that commissioning involves both *positive* and *negative* testing. Positive testing is obvious; it is ensuring that when the firkling button is pressed the plant firkles. Practically everyone sees the need for this. Negative testing is less obvious; it is ensuring that the control system deals correctly with all the unlikely circumstances and fault conditions. Negative testing takes far far longer, because there are many more fault modes than healthy modes. It is very common for people to say '*it works, let's go*' when only the positive testing has been done. Try to resist this pressure, at best it can lead to damaged plant a few years hence, at worse some safety features could be overlooked.

Finally the plant is handed over to the maintenance department. In commercial software it is generally thought that over 50% of the effort goes into *maintenance* as changes are made to meet new requirements or correct the inevitable bugs. For easy maintenance all the documentation must be complete and up to date.

16.8 Safety

Most industrial processes are hazardous, and the safety of all personnel must be of prime importance. This section is a personal view and can only give a simple discussion of safety considerations. The topic of safety is covered by both criminal and civil law. The designer and user of any system must therefore consult the relevant legislation and codes of practice to ensure compliance.

Every single person has a safety responsibility. Employers have a '*duty of care*' for their employees and the public and must ensure the plant is kept in a safe condition, safe working procedures are devised for all conceivable activities, and training in these procedures provided for all relevant employees. Suppliers must ensure their equipment meets safety criteria, and draw the attention of purchasers to unavoidable hazards (protection and labelling of parts which are live during normal operation for example). Employees must follow safety procedures and not expose themselves (or others) to danger. These responsibilities are covered by the Health and Safety at Work Act 1974 (HASWA) which makes the universal responsibility for safety absolutely clear.

More recently in 1992 a block of EEC Health & Safety Regulations (commonly known as the '*six pack*') introduced the idea of *risk assessment*. This recognises the need to balance the cost and complexity of the safety system against both the likelihood and severity of injury. The procedures outlined use common terms with specific definitions:

Hazard	The potential to cause harm.
Risk	A function of the likelihood of the hazard occurring and the severity.
Danger	The risk of injury.

and outlines procedures to achieve acceptable safety standards. Risk assessment is a legal requirement under most modern legislation, and is covered in detail in standard prEN1050 '*Principles of Risk Assessment*'.

A Health & Safety Executive study of safety in control systems ('*Out of Control*' HSE books 1995 ISBN 0717608476) makes worrying reading. It suggests that more than 60% of safety related failures are introduced into a system before it is taken into service for the first time. Approximately 44% of safety incidents come from specification errors, 15% from design errors and 6% from poorly thought out changes during commissioning.

The inclusion of a programmable controller brings additional hazards (and solutions) which must be recognised. A PLC can introduce potentially dangerous situations in different ways. The first (and probably commonest) route is via logical errors in the program. These can be the result of oversight, or misunderstanding, on behalf of the original designer who did not appreciate that this set of actions could be dangerous, or by later modifications by people who deliberately (or accidentally) removed some protection to overcome a failure in the middle of the night. '*Midnight programming*' is particularly worrying as usually the only person who knows it has been done is the offending person, and the danger may not be apparent until a considerable time passes and the hazardous condition occurs.

The second possible cause is failure of the input and output modules; in particular the components connected directly to the plant which will be exposed to high voltage interference (and possibly direct connected high voltages in the not unlikely event of cable damage). Output modules can also suffer high currents in the (again not unlikely) event of a short circuit. Typical output devices are triacs, thyristors or transistors. The failure mode of these cannot be predicted; all can fail short circuit or open circuit. In these failure conditions the PLC would be unable to control the outputs. Similarly an input signal card can fail in either the 'on' or 'off' state, leaving the PLC misinterpreting a possibly important signal.

The next failure mode is the PLC itself. This can be further divided into hardware, software and environmental failures. A hardware failure is concerned with the machine itself; its power supply, its processor, the memory (which contains the 'personality' of the PLC, the user's program, and the data storage). Some of these failures will have predictable effects; a power supply failure will cause all outputs to de-energise, and the PLC supplier will have included memory checks in the design. Environmental effects arise from peculiarities in the installation such as dust, humidity, temperature (and rapid temperature changes) possible water ingress and vibration, and these can result in unexpected operation of output devices.

The final cause is electrical interference (usually called noise). Internally almost all PLCs work with 5 V signals, but are surrounded by high voltage high current devices. Noise can cause input signals to be misread by the PLC, and in extreme cases can corrupt the PLC's internal memory. PLCs generally have internal protection against memory corruption and noise on remote I/O serial lines, so the usual effect of noise is to cause a PLC to stop (and outputs to de-energise). This cannot, however, be relied upon.

Figure 16.80 shows a normal motor starter circuit built without a PLC. Safety precautions here are:

- Isolation switch at the MCC removes the supply for maintenance work.
- Normally closed contacts on the stop and emergency stop buttons. A broken wire will look like a stop button being pressed, as will loss of the control supply.
- If the emergency stop is pressed and released, the motor does not restart.
- Isolation and emergency stop have priority over start.

It is still possible to identify dangerous failure modes in this system. The button head of the emergency stop button could unscrew and fall off, or the contacts of the contactor could weld made, or a short could occur between the cores to the stop button but these failure modes are exceedingly rare, and *Figure 16.80* would be generally accepted as safe for use in normal circumstances.

In *Figure 16.81*(*a*) the same functions have been provided by an *unsafe* PLC system. To save costs the MCC door isolator has been replaced by a simple switch which makes to say 'Isolate'. Similarly normally open contacts have been used for stop and emergency stop. This is controlled by the unsafe program of *Figure 16.81*(*b*).

It is important to realise that to the casual user, *Figures 16.80* and *16.81* behave in an identical manner. The differences (and dangers) come in fault, or unusual, conditions. In particular:

- A person using a programming terminal can force inputs or outputs and over-ride the isolation. Although it is unlikely that anyone would do this deliberately, it is easy to confuse similar addresses and swop digits by mistake (forcing 0:23/01 instead of 0:32/01 for example).
- A loss of the input control supply during running will mean the motor cannot be stopped by any means other than totally removing the supply to the system.
- The system is very vulnerable to input and output card faults.

None of these are apparent to the user until an emergency occurs.

A prime rule, therefore, for using PLCs is:

'*The system should be at least as safe as a conventional system*'

Figure 16.82(*a*) is a revised PLC version of *Figure 16.80*. The isolator has been re-instated with an auxiliary contact as PLC inputs, and normally closed contacts used for the stop and emergency stop buttons. An auxiliary contact has been added to the starter, and this is used to latch the PLC program of *Figure 16.82*(*b*). The emergency stop is hard-wired into the output and is independent of the PLC, and on release the motor will not restart (because the latching auxiliary contact in the program will have been lost). On loss of control supply the program will think the stop button has been pressed, and the motor will stop. *Figure 16.82* thus behaves in failure as *Figure 16.80*.

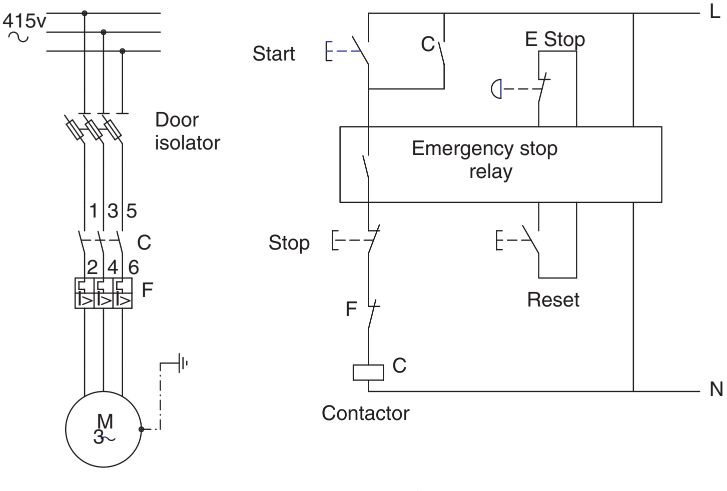
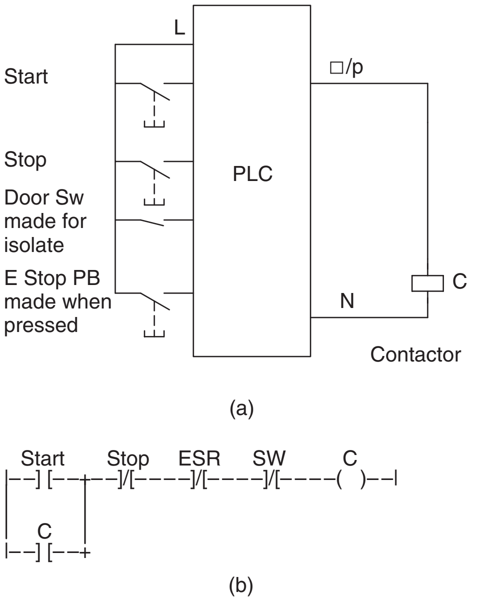


Figure 16.80 A standard hardwired motor starter for a low risk application. This would normally be considered to be safe. In higher risk applications there would probably be dual connections on the emergency stop button, dual contactors and the state of the contactors would be monitored by the emergency stop relay



(a)

(b)

Figure 16.81 An unsafe PLC based system totally reliant on software. The dangers of this system only become apparent when failures occur

Although this example is simple, it illustrates the necessary analysis and considerations that must be applied in more complex systems.

Complex electronic systems can bring increased safety. Consider a thyristor drive controlling the speed of a large d.c. motor. In a typical arrangement there will be an upstream a.c. contactor to enable the drive. Hardwire connection of an emergency stop button into the a.c. contactor will obviously stop the drive, but the inertia of the motor and the load will keep it rotating for several seconds. A thyristor drive, however, can stop the load in less than one second by regeneratively braking the motor, but this requires the drive to be alive and functional. The operation of the emergency stop implies a dangerous condition in which the fastest possible stop is required. It is almost certain that at this time the drive controls are functional and there are no 'latent' faults.

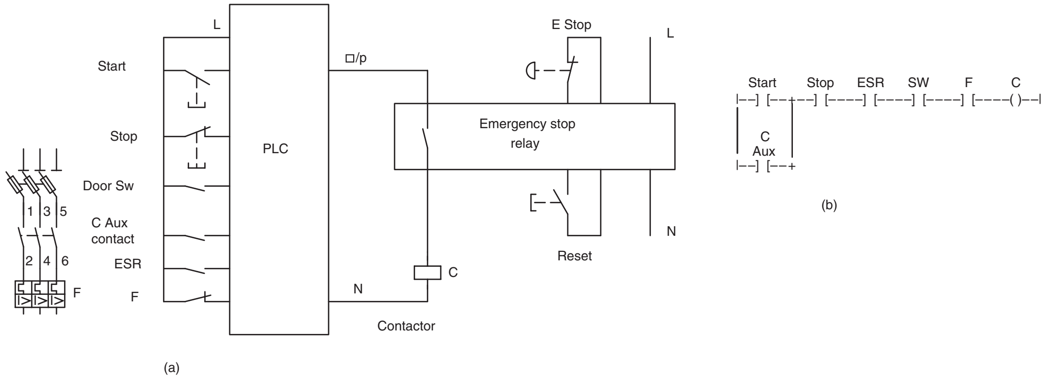
Here the emergency stop can operate in two ways. First it initiates an electronic regenerative crash stop via the control system which should stop the drive in less than one second. The emergency stop also releases a delay drop out hardwire relay set for 1.5 s which releases the a.c. contactor. This gives the safest possible reaction to the pressing of the emergency stop button. Safety considerations do not therefore, explicitly require relay based, non electronic hardware, but the designer must be prepared to justify the design decisions and the methods used.

Where complex control systems are to be used, a common method of improving safety is to duplicate sensors, control systems and actuators. This is known as *redundancy*. A typical application occurs in boilers where feed water is held in a drum. Deviations in water level are dangerous; too low and the boiler will overheat, possibly to the point of melting the boiler tubes; too high and water can be carried over to the downstream turbine with risk of catastrophic blade failure. High and low level sensors are therefore usually provided with each being duplicated. The safety system reacts to *any* fault signal, so two sensors have to fail for a dangerous condition to arise. If the probability of a sensor failure in time T is p (where $0 \leqslant p \leqslant 1$) the probability of both failing is p^2. In a typical case, p will be of the order of 10^{-4} failure per year giving p^2 of 10^{-8}.

There are two disadvantages. The first is that a sensor can fail into a permanently safe signal state, and this failure will be '*latent*', i.e. hidden from the user with the plant running on one sensor. The second problem is that the plant reliability will go down, since the number of sensors goes up and any sensor failure can result in a shutdown. Both of these effects can be reduced by using '*majority voting*' circuits, taking the vote of two out of three or three out of five signals.

Redundancy can be defeated by '*common mode*' failures. These are failures which affect all the parallel paths simultaneously. Power supplies, electrical interference on cables following the same route and identical components from the same batch from the same supplier are all prone to common mode failure. For true protection, *diverse redundancy* must be used, with differences in components, routes and implementation to reduce the possibility of simultaneous failure.

To give true redundancy it is sensible to provide duplication in the control system as well to protect against hardware and software failures in the system itself. Duplicate control schemes, though, are vulnerable to a form of common mode failure called a '*systematic failure*'. Suppose duplicated temperature sensors are compared, inside a PLC program, with an alarm temperature. Suppose both are identical devices, running the same program containing a bug which inadvertently (but rarely so it does not show up



(a)

(b)

Figure 16.82 A safe PLC system for low risk applications. As for *Figure 16.80* more features could be added if the risk was higher

during simple testing) changes the setting for the alarm temperature (from 60 °C say, to 32 053 °C). Such an effect could easily occur by a mistype in a MOVE instruction in a totally unrelated part of the program. This error will affect both control systems, and totally remove the redundancy.

If reliance is being made on redundant control systems, therefore, they should be totally different; different machines with different I/O and different programs written by different people with the machines installed running on different power supplies with different types of sensors connected by different cable routes.

The Health and Safety Executive (HSE) became concerned about the safety of direct plant control with computers, and produced an occasional paper OP2 '*Microprocessors in Industry*' in 1981. This was followed in 1987 by two booklets '*Programmable Electronics Systems in Safety Related Applications*'. Book 1 (an Introductory Guide) is a general discussion of the topic, and Book 2 (General Technical Guidelines) goes through the necessary design stages. They suggest a five stage process:

(i) Perform a hazard analysis of the plant or process;
(ii) From this, identify which parts of the control system are concerned with safety and which are concerned purely with efficient production. The latter can be ignored for the rest of the analysis;
(iii) Determine the required safety level (based on accepted attainable standards or published material);
(iv) Design safety systems to meet or exceed these standards; and
(v) Assess the achieved level (by using predicted probability of failure for individual parts of the design). Revise the design if the required level has not been achieved.

The books stress the importance of '*Quality*' in the design; quality of components, quality of the suppliers and so on.

The IEC standard *IEC 61508 Functional Safety of Electrical/Electronic/Programmable electronic safety related systems* covers similar grounds to the HSE books. This is based on the ideas of *safety functionality* (what it is designed to protect against and how the protection is achieved

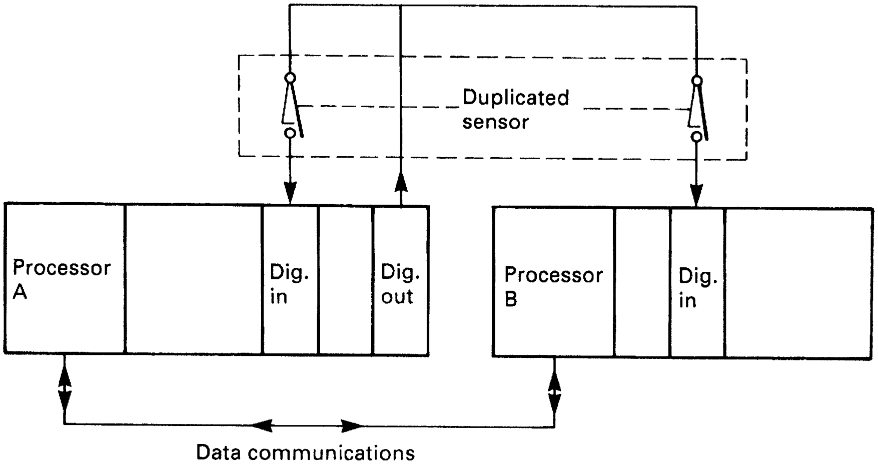


Figure 16.83 Safety critical input with the Siemens 115F PLC

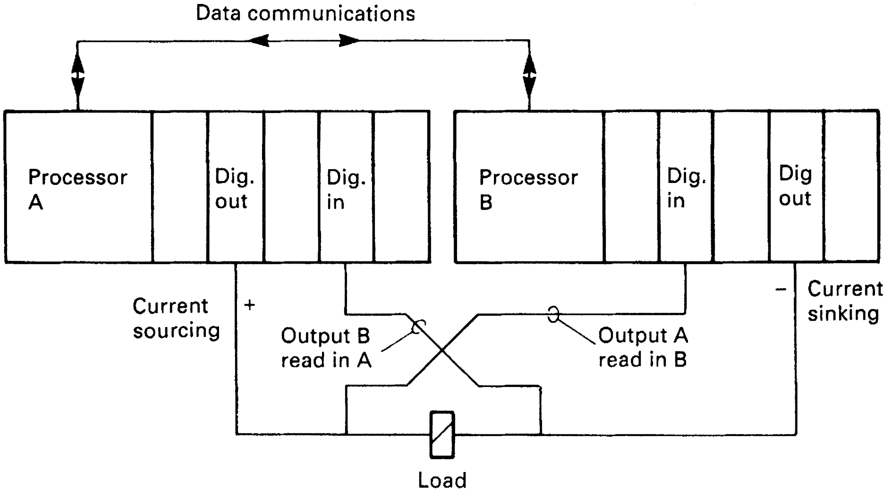


Figure 16.84 Safety critical output with the Siemens 115F PLC

e.g. '*open quench valve if temperature rises above 250° C*') and the *Safety Integrity Level* (or SIL) which, somewhat simplified, is the probability, p, that the safety system will fail to operate on demand. The SIL covers the entire safety system including sensors, control system, and actuators. Four SILs are defined from a basic SIL-1 ($10^{-1} > \varsigma p > 10^{-2}$) to SIL-4 ($10^{-4} > \varsigma p > 10^{-5}$). The required SIL is determined from a risk assessment of the system. For continuous protection on a hazardous plant the normal requirement is SIL-3 or SIL-4. Guidelines for architectural constraints (such as keeping the safety system separate from the control system, and using redundancy) are also given. It is probable that IEC 61508 will become a European standard in the near future, and the two HSE books are being re-written to incorporate ideas from IEC 61508.

Surprisingly some fieldbus systems (e.g. specialist versions of Profibus and SafetyBus from Pilz) can achieve SIL-3 which makes a fieldbus safety system attractive in hazardous applications. Extreme care must, of course, be taken.

In America, the Instrument Society of America (ISA) standard S84 follows broadly similar lines to the HSE guidelines and IEC 61508.

Because the HSE books, IEC 61508 and S84 are standards they have the legal status of guidance notes and there is no formal requirement to follow them. In the event of an incident, however, the designers and users must be prepared to justify the actions they have taken and conformance with good practice is a legal defence.

Very high safety levels can be achieved with some PLCs. Siemens market the 115F PLC which has been approved by the German TUV Bayern (Technical Inspectorate of Bavaria) for use in safety critical applications such as transport systems, underground railways, road traffic control and public elevators. The system is based on two 115 PLCs and is a model of diverse redundancy. The two machines run diverse system software and check each other's actions. There is still a responsibility on the user to ensure that no systematic faults exist in the application software.

Inputs are handled as *Figure 16.83*. Diverse (separate) sensors are fed from a pulsed output. A signal is dealt with only if the two processors agree. Obviously the choice of sense of the signal for safety is important. For an overtravel limit, for example, the sensors should be made for healthy and open for a fault.

Actuators use two outputs (of opposite sense) and two inputs to check the operation as *Figure 16.84*. Each sub-unit checks the operation of the other by brief pulsing of the outputs allowing the circuit to detect cable damage, faulty output modules and open circuit actuators. If, for example, output B fails on, both inputs A and B will go high in the Off state (but the actuator will safely de-energise.)

The operation of *Figures 16.83* and *16.84* is straightforward, but it should not be taken as an immediately acceptable way of providing a fail-safe PLC. The 115F is truly diverse redundant, even the internal integrated circuits are selected from different batches and different manufacturer, and it contains well tested diverse self checking internal software. A DIY system would not have these features, and could be prone to common mode or systematic failures.

A PLC system is an electrical system, and is subject to the same legislation as conventional electrical schemes. Apart from the Health and Safety at Work Act, the designer should also observe the Institute of Electrical Engineers Wiring Regulations, and the Electricity at Work Regulations 1990.

Section D

Power Electronics and Drives

17

Power Semiconductor Devices

R J Bassett PhD, MTech, CEng, MIEE, MInstP
GEC Alsthom Engineering Research Centre

P D Taylor PhD, BSc, CEng, MIEE, CPhys, MInstP
Chief Technical Officer, Dynex Semiconductor Ltd

Contents

17.1 Junction diodes 17/4
  17.1.1 The p–n junction 17/4
  17.1.2 The p–n junction power diode 17/4
  17.1.3 Fast-recovery p–n junction power diode 17/6
  17.1.4 Epitaxial fast-recovery p–n junction power diode 17/7
  17.1.5 The p–i–n junction power diode 17/7

17.2 Bipolar power transistors and Darlingtons 17/7
  17.2.1 Bipolar n–p–n power transistors 17/7
  17.2.2 Bipolar n–p–n power Darlington transistor 17/12

17.3 Thyristors 17/14
  17.3.1 The basic thyristor 17/14
  17.3.2 The converter thyristor 17/21
  17.3.3 The fast thyristor 17/21
  17.3.4 The asymmetric thyristor 17/23
  17.3.5 Triacs 17/23
  17.3.6 Light fired thyristors 17/23
  17.3.7 HVDC thyristors 17/24
  17.3.8 The gate turn-off thyristor 17/24

17.4 Schottky barrier diodes 17/25

17.5 MOSFET 17/27

17.6 The insulated gate bipolar transistor (IGBT) 17/32
  17.6.1 Device physics 17/32
  17.6.2 Packaging and thermal considerations 17/34
  17.6.3 On state characteristics and conduction losses 17/34
  17.6.4 Turn on and turn off 17/34
  17.6.5 Safe operating area 17/35

The development of power semiconductor devices began in the late 40s and 50s with the bipolar transistor and the power diode, this was followed a decade later by small thyristors and triacs. The thyristor has continued to develop over the years, growing in power and frequency rating, leading to high power press-pack fast switching thyristors and ultra high voltage thyristors in the 70s, and GTO Thyristors in the 80s. In turn these new devices have enabled the development of greater and greater efficiency in a wide range of power conversion systems. However the fundamental disadvantage of these devices was the complexity of the drive and control equipment they require. The first major breakthrough came with the advent of the Power MOSFET in the 70s and 80s. That enabled a major reduction in the complexity of the control systems, and pushed power conversion forward in frequency and efficiency, but the MOSFET would never challenge the thyristor in the high power arena. All this was changed however with the arrival in the 1990s of the high power insulated gate bipolar transistor (IGBT), essentially the single chip integration of a Power MOSFET and a highly efficient bipolar transistor.

The main types of power semiconductor switching device available today are shown in *Figure 17.1* illustrating the spectrum of power and switching frequency occupied by each type, and giving some of their main areas of application. Diodes are not shown on this picture as they encompass the whole range and essentially complement the switching devices. There are still on going developments of all these power switches, some of which will be covered in this chapter, and potentially new materials and new device structures will continue to re-write the story for power semiconductors.

Because of the generally superior performance that can be achieved, almost all power semiconductor devices are made today using monocrystalline silicon, and many of the highest power devices use extremely pure silicon crystals as their starting point. At the top end of the power range single-element silicon devices are available that are able to withstand voltages in excess of 6000 V with average current ratings of 1000 A and voltages of 2000 V with average current ratings of 6000 A. Typically the largest devices (diodes) have a forward volt drop of 1.2 V at 5000 A and a reverse leakage current of less than 100 mA at 5000 V, the voltage that a semiconductor device is able to withstand with a low leakage current is referred to as its 'blocking voltage'. This demonstrates the very small power loss dissipated in the device compared to the power controlled or rectified and the high efficiency of this type of component. Typically, large power devices of this type are mounted in a package that can be double-side cooled such as that shown in *Figure 17.2*(*a*). More recently the plastic power module has become a widely accepted alternative to the double-side cooled press pack, *Figure 17.2*(*b*). This module is single-side cooled through an electrically isolated base-plate and has all the power terminals on one face. It has become the preferred outline for the high power IGBT and diode module (see Section 17.6.1). At the other end of the spectrum is the Power MOSFET at a few 10s of watts but able to switch at high speed in the range of 100 MHz.

Power semiconductor devices can be divided into two groups, roughly splitting the power spectrum in two, these are minority carrier devices (bipolar), and majority carrier devices (unipolar). The bipolar device relies on the injection of both minority and majority carriers into a high-resistivity region of the device. This reduces the device on-state voltage by modulating the resistivity of its high-resistivity region and reducing the voltage drop across this region. However, these conductivity-modulated devices such as the diffused p–n junction silicon diode and the thyristor or silicon controlled rectifier (SCR) suffer the disadvantage that the injected carriers or charge must be extracted before a blocking state can be sustained. The charge also cannot be injected or extracted instantaneously into or from the high-resistivity region of the device. This takes some time to occur and during this period, before the low on-state voltage is established or the high blocking voltage can be attained, substantially higher instantaneous power is dissipated in the

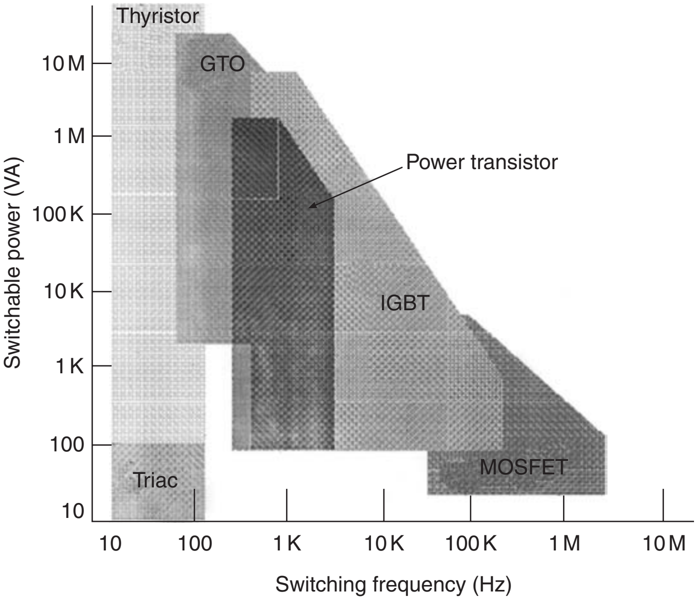


Figure 17.1 Power device power–frequency range

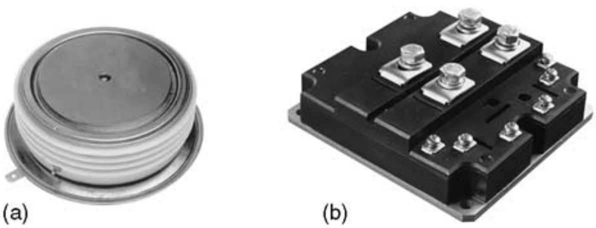


Figure 17.2 (a) Press pack device; (b) IGBT power module

device. Similarly during turn on, a substantial level of charge must be injected before the device attains its optimum low conductivity state. In the case of a switching device this can result in a high power requirement from the gate drive, and also high instantaneous power during turn on. These phenomena combine to limit the maximum permissible switching speed of these devices. This has also resulted in a number of different types of conductivity-modulated devices being developed. Some of them are specifically optimised for 50 or 60 Hz operation, which are sometimes referred to as 'converter-grade' or 'phase control' devices. Faster types of devices have also been designed for operation at higher frequencies and are often referred to as 'inverter-grade' or 'fast' devices. Generally, however, the fastest conductivity modulated devices are limited in practical applications to an operating frequency below 50 kHz.

The second group of devices is the majority-carrier or unipolar device that does not rely on conductivity modulation, for example metal oxide silicon field effect transistors (MOSFETs) and Schottky diodes. In a majority carrier device, current is carried only by the majority carriers. The on-state voltage is therefore dependent on the concentration of majority carriers in the conducting region of the device. Injection of minority charge and conductivity modulation does not occur so that there is no reverse-recovery charge to be extracted on turn-on and turn-off losses are, therefore, much lower, resulting in an optimum operation at higher switching frequencies, as high as 100 MHz in some cases. However, on-state losses will generally be higher since the number of current carriers is restricted although this will also be dependent on the voltage design of the device.

Thus the maximum useful voltage rating of the MOSFET is less than 1000 V, and more typically less than 600 V, however this is quite suitable for most domestic mains voltage equipment, and so unipolar devices are used in very high volumes in such consumer based applications.

A third group of hybrid devices includes most notably the insulated gate bipolar transistors (IGBT) that combine some aspects of conductivity-modulated and majority carrier devices and offers special advantages in many applications. Although more complicated in construction it is the IGBT that has revolutionised power electronics in the more recent years.

It can be seen that in general it is the minority carrier (bipolar) devices which dominate the low frequency high power area, the majority carrier (unipolar) types in the high frequency and low power arena, and the hybrid IGBT which is predominant in the mid range.

17.1 Junction diodes

17.1.1 The p–n junction

A p–n junction is made by increasing the concentration of electron or hole populations in certain regions of a monocrystalline silicon disc or chip. This is done by introducing impurity atoms, such as phosphorus with five electrons in its valence band or boron with three electrons in its valence band, to create either electron-rich n-type or hole-rich p-type regions within the crystal lattice, respectively. The conductivity in these regions at normal device operating temperature is wholly determined by the concentration of the impurities introduced.

A depletion region (a region which is devoid of charge carriers and therefore very highly resistive) is established at the point in the crystal where the crystal ceases to become either electron or hole rich and reverts to the opposite type. This is caused by holes diffusing from the p region where they are present in a higher concentration into the n region and alternate electrons diffusing from the n to the p region. This forms an electrical potential barrier at the junction between the p and n regions. Under an externally applied forward bias with the p region held positive to the n region, carriers are injected across this potential barrier and a high current can flow through the diode with only a small voltage appearing across the device. When the p–n junction is biased with p region held negative with respect to the n region the depletion region widens to sustain the voltage and allows only a small reverse current to flow. This is called a reverse biased p–n junction. When the polarity is reversed it is called a forward biased p–n junction and current may flow freely across the junction with only a small resultant voltage drop. A more complete description of the physical behaviour of the p–n junction can be found in reference 1.

17.1.2 The p–n junction power diode

Construction of the power diffused p–n junction diode normally begins with a high-purity defect-free silicon wafer with a very uniform low doping concentration of n-type impurity such as phosphorus. The n-type silicon is preferred as the starting material because it offers a superior reverse-recovery characteristic as explained below. The silicon is then doped by the diffusion of a p-type impurity such as boron, aluminium or gallium at a temperature above 1000°C into one side of the wafer. The p-type diffusion takes place in a diffusion furnace under very carefully controlled conditions. This is followed by a similar high concentration n-type diffusion to the opposite side of the silicon disc to form the impurity concentration profile shown in *Figure 17.3*. The conductivity and final thickness of the central n-type region together with the final surface area of the silicon device ultimately determines the voltage capability and current rating of the device.

Under high forward bias conditions a high concentration of charge is injected in the central n region resulting in the injected carrier distribution (also shown in *Figure 17.3*). The injected electron and holes are present in the n region in equal quantities, which maintains charge balance locally

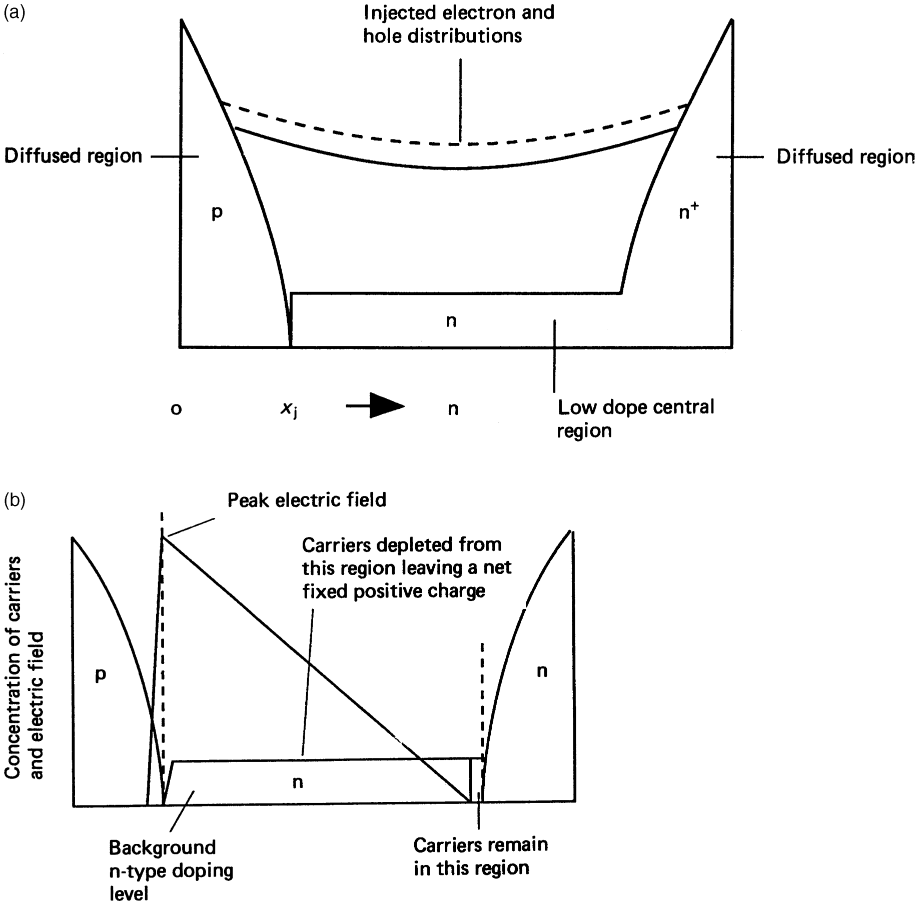

(a)
Injected electron and hole distributions
Diffused region
Diffused region
p
n+
n
o
xj
n
Low dope central region
(b)
Peak electric field
Carriers depleted from this region leaving a net fixed positive charge
Concentration of carriers and electric field
p
n
n
Background n-type doping level
Carriers remain in this region

Figure 17.3 (a) Impurity concentration in a p–n junction diffused diode and the injected carrier concentration under forward bias; (b) The electric field distraction across a p–n junction diode under reverse bias

throughout the n region. However, as can be seen from *Figure 17.3*, the concentration of carriers varies across the n region width. The reason for this is that, once injected into the n region, the excess carriers only have a finite lifetime before they recombine with another excess carrier of the opposite charge. Once injected across the p–n junction the injected carrier can on average only travel a certain distance before recombining, which results in the injected charge distribution shown in *Figure 17.3*. The rate of recombination is governed by the concentration of another type of impurity that provides a recombination centre through which the annihilation occurs. The time taken for the injected carriers to be annihilated is measured by a parameter that is commonly called the minority carrier lifetime. The aim in the construction of converter grade p–n junction diodes is normally to reduce the concentration of these recombination centres to the absolute minimum possible, which typically results in an average carrier lifetime of greater than 100 μs. The conductivity of the n region is therefore increased by the presence of the injected carriers and the on-state voltage of the diode reduced. Thus the greater the minority carrier lifetime, and/or the greater the level of injected charge the lower the on-state voltage.

When the forward current is reduced to zero as the diode recovers its blocking state and reverse bias is applied to the p–n junction diode, it behaves initially like a short circuit until the charge in the central n region is removed. This results in the typical reverse-recovery characteristic shown in *Figure 17.4*.

The length of time it takes for the diode to recover $t_{rr} = t_A + t_B$ is affected by the carrier lifetime; a long lifetime results in a long recovery time. Also the recovered charge Q_{rr}, which is the charge extracted during the recovery time, increases with increased lifetime. As the reverse voltage is increasing across the diode during this recovery phase, there is an inherent energy loss, or recovery loss, determined by the product of volts and amps, during this phase. Thus there is a trade off between the recovery loss and the on-state voltage (loss) of the diode. This highlights one of the fundamental relationships of all bipolar devices: increasing on-state losses can be traded off against decreasing turn off losses by controlling the minority carrier lifetime, and or the carrier injection levels in the p–n junction.

A p–n junction diode is normally designed so that when the maximum reverse-voltage rating is applied the whole of the depletion region is contained within the central n region of the device. Under these conditions an electric field is established across the depletion region as shown in *Figure 17.3*. The maximum peak electric field that can be sustained in silicon is between 5×10^4 and 8×10^5 V/cm. If the reverse voltage is increased above the maximum reverse voltage rating of the device the reverse-leakage current begins to increase rapidly as the peak electric field exceeds the point

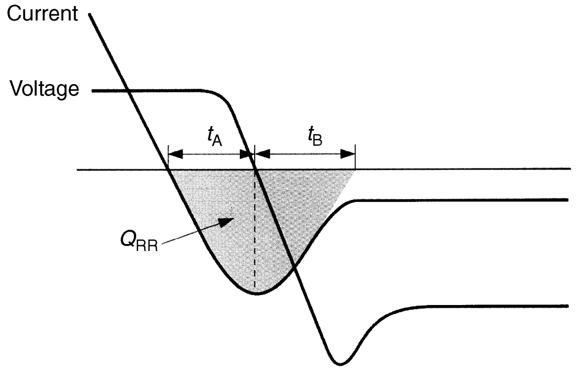


Figure 17.4 Reverse recovery of a p–n junction diode

where avalanche multiplication of carriers occurs in the depleted region. The voltage at which carrier avalanche multiplication is initiated is called the ‘avalanche voltage’ of the diode. A typical characteristic for a silicon p–n junction diode is shown in *Figure 17.5*. As the temperature of the p–n junction diode is increased over its permitted operating range its characteristics also change, as illustrated in *Figure 17.5*. Note that as the junction temperature increases both reverse leakage current and the avalanche voltage increase, whereas the on-state voltage decreases at low current and increases at higher current.

The optimum design p–n junction diode with lowest on-state voltage and smallest reverse-recovery charge possible is obtained by manufacturing the device from a silicon crystal with the lowest resistivity and narrowest n-base consistent with the desired voltage rating of the device.

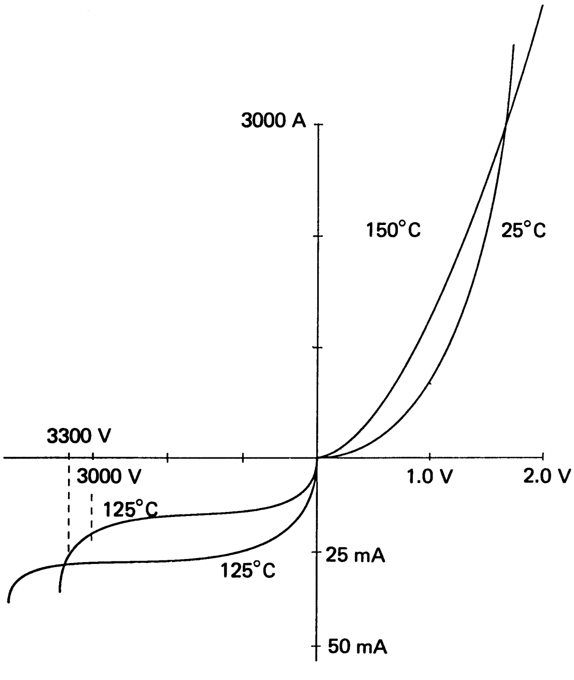


Figure 17.5 Characteristics of a 50 mm silicon p–n junction rectifier. Note the different scales in the different quadrants

As the design target maximum reverse voltage rating of the p–n junction diode is increased the resistivity of the n-type silicon has to be increased to ensure the design voltage is attained before the peak electric field reaches the point where avalanche breakdown occurs across the whole operating temperature range of the device. At the same time the thickness of the n-type region also has to be increased to accommodate the increased depleted region width. Because of the increased n-type-region thickness higher target voltage design p–n junction diodes have a higher on-state voltage at the same forward current and exhibit a larger reverse-recovery charge than lower target voltage design devices of the same surface area. This highlights a second basic relationship for bipolar devices, the device operating losses increase as the voltage rating of the device is increased. A more complete description of the p–n junction design principles is given in reference 1.

17.1.3 Fast-recovery p–n junction power diode

In order to reduce the reverse-recovery charge shown in *Figure 17.4* extracted from a fast-recovery p–n junction diode and, thereby, to increase its maximum permissible operating frequency, the diode designer can apply two techniques. These are to reduce the level of charge injected into the diode, or to reduce the minority carrier lifetime in the central n region of the diode.

The lifetime is reduced by increasing the level of recombination centre impurities, as the concentration level of recombination centres is increased, the annihilation rate of the injected carriers increases with the result that the charge extracted during reverse-recovery decreases. However, the reduced reverse-recovery charge has to be traded off against increases in reverse leakage current and on-state voltage. The recombination centres or impurities that are generally used are gold, platinum and radiation damage. Each has its own individual characteristic that determines the shape of the reverse-recovery current after the peak of the reverse-recovery. Therefore, to optimise fully the performance of a fast-recovery p–n junction diode, not only do those parameters described above concerning central n-region resistivity and thickness have to be carefully controlled, but also the concentration and distribution of these recombination centres in the central n region of the device. Lower target voltage design fast-recovery p–n junction diodes constructed from thinner silicon not only have a lower on-state voltage but can also have very much lower reverse-recovery charge than higher target voltage design devices. This is because it is possible to introduce much greater concentrations of recombination centre impurities into lower voltage design fast-recovery diodes before impairing the reverse leakage and on-state voltage characteristics of the diode. However, in practice it is generally necessary to increase the concentration of recombination centre impurities to the point where their introduction has the effect of increasing reverse leakage current significantly. This reduces the maximum permitted operating temperature of fast-recovery p–n junction diodes compared to converter grade diodes, e.g. from 200°C or 175°C down to 150°C or 125°C. For example, a 600 V target voltage design fast-recovery p–n junction diode with an average current rating of 335 A has a maximum reverse-recovery charge of 4 μC. This can be compared to 25 μC for a 1400 V target voltage design fast-recovery p–n junction diode of the same size with an average current rating of 295 A.

One very important feature of fast-recovery diodes used with fast switching components such as IGBTs is called the snappiness of the diode. During initial the recovery phase the charge is extracted by the formation of the depletion region in the n-base, when this charge is extracted very rapidly the device current can fall towards zero too fast, resulting in what is called snap off, and a consequential spike, or a series of oscillations in the recovery voltage waveform. Snappiness is quantified by the ratio between the two time periods of turn-off, t_A and t_B. A low snappiness factor, combined with a low recovered charge is the ideal situation: this is what is meant by a 'soft' fast-recovery diode. This can be a particular problem with p–i–n type diodes described later.

17.1.4 Epitaxial fast-recovery p–n junction power diode

The central n region of a fast-recovery p–n junction diode designed for optimum performance with a target reverse voltage capability of 800 V or less is usually extremely thin and of the order of 70 μm thickness or less. Unfortunately very thin silicon wafers are difficult to handle in manufacturing. One widely used manufacturing technique to manufacture this type of diode is with an epitaxial silicon deposition process. In this process a single-crystal high resistivity n-type silicon layer is grown from a vapour on top of a more highly doped n^{+} silicon single-crystal substrate. This is followed by a p^{+} type diffusion process to form the p–n junction of the diode. The substrate is of a low resistance and therefore provides a mechanical support for the active epitaxial diode structure without degrading the diode performance. This technique can be used to manufacture fast-recovery p–n junction diodes that have superior characteristics in certain respects compared to any other fast-recovery p–n junction diode designs. In most other respects the epitaxial fast-recovery p–n junction diode is very similar to the fast-recovery p–n diode. There are, however, a number of constraints imposed by the epitaxial process itself, including cost, which inherently limit the highest resistivity of the n-type layer which can be grown, the thickness of the layer and the degree of perfection to which the crystal can be grown over a large area. In practice, these parameters limit the upper voltage capability of this design to 1600 V with a maximum average current rating of approximately 100 A in a single silicon die.

17.1.5 The p–i–n junction power diode

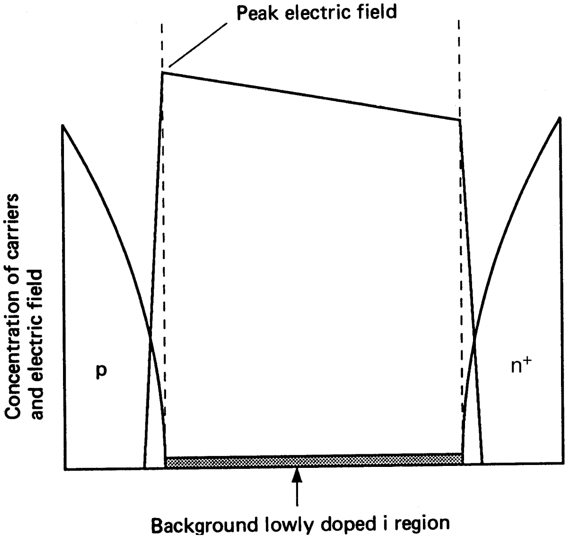

Peak electric field
Concentration of carriers and electric field
p
n+
Background lowly doped i region

Figure 17.6 Impurity concentration and electric field distribution under reverse bias of a p–i–n junction diffused diode

The p–i–n junction diode is another version of the fast-recovery p–n diode. As explained some of the transient forward and reverse-recovery characteristics of this device can generally be improved by reducing the thickness of the central region of the diode to the absolute minimum. However, the normal fast-recovery p–n junction diode is still designed to contain the whole of the maximum reverse-bias depletion region in its n region (see *Figure 17.3*). In the p–i–n junction diode the resistivity of the n region is chosen to be almost intrinsic; that is, the concentration of n-type impurities is very low and its resistivity very high. The thickness of that region is also chosen so that the depletion region of the device penetrates to the n^{+} region when the maximum design reverse voltage is applied as shown in *Figure 17.6*. It is possible in this way to design fast-recovery p–i–n junction diodes that have an n or i region that is as much as half the thickness of the n region of a conventional fast-recovery p–n junction. The resulting device, therefore, can have a much-reduced reverse-recovery charge because being thinner the base region has a lower effective resistance than that of the p–n diode when flooded with carriers. So it is possible to reduce significantly the base carrier concentration, by limited carrier injection or by lifetime reduction, without such an increase in the on-state voltage drop. For example, a 1400 V target voltage design p–i–n diode with an average current rating of 300 A has a maximum reverse-recovery charge of 8 μC. This compares to 25 μC for a 1400 V diode of the same size and an average current rating of 295 A.

One disadvantage of the p–i–n junction diode is that its reverse-recovery characteristic exhibits *snappy recovery* compared to the comparatively *soft recovery* of the fast-recovery p–n junction diode. This results from the fact that the depletion layer spread during recovery is very rapid in the low doped intrinsic base region. By careful design of the fast diode it is possible to minimise the snappy behaviour of the diode, this is of particular importance for IGBT applications as will be discussed in Section 17.6.

17.2 Bipolar power transistors and Darlingtons

17.2.1 Bipolar n–p–n power transistors

The bipolar power transistor is a three terminal power device that relies on charge modulation to reduce device on-state voltage. Carriers are injected from the emitter region and transit across the base region to be collected at the collector and modulate the resistivity of the lightly doped collector. Almost all high power bipolar transistors produced today are based on the n–p–n type of construction, although the alternative p–n–p type of construction is also produced, mainly for use in complementary switching circuits. However, the same principles of operation apply for both types of device. High power bipolar n–p–n power transistors are available today in the type of package shown in *Figure 17.7*, made from a single silicon wafer with continuous current ratings of up to 1000 A and peak currents of up to 1200 A and collector emitter sustaining voltages of up to 1000 V. Construction of the device begins with the diffusion of an n-type impurity to form the n^{+} collector region and this is followed by further p-type and n-type diffusions

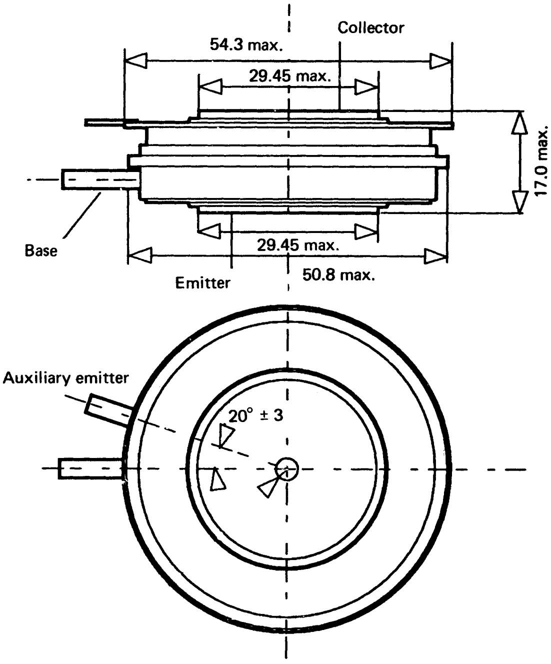

Collector
54.3 max.
29.45 max.
17.0 max.
Base
29.45 max.
Emitter
50.8 max.
Auxiliary emitter
20° ± 3

Figure 17.7 Outline of a high power transistor package

This results in the typical bipolar n–p–n power transistor diffusion profile shown in *Figure 17.8*. An alternative manufacturing procedure replaces the initial n-type diffusion into lightly doped n-type silicon crystal with a high resistivity epitaxial silicon layer grown on a highly doped n-type silicon substrate. The surface structure of the high power bipolar transistor ensures an even distribution of base current over the active area of the device.

Generally, bipolar n–p–n transistors are operated either in a fully off state or in an on-state and driven between these two states as rapidly as possible to reduce instantaneous power losses. In the off state the base of the device is held at zero or negative bias to minimise collector leakage current. The electric field distribution that typically will exist inside the collector structure when the maximum permitted voltage is applied is shown in *Figure 17.9*. As can be seen from *Figure 17.6*, the electric field distribution in the reverse-biased collector region of the bipolar n–p–n power transistor is similar to the reverse bias p–i–n junction power diode. The objective of the two designs is the same, insofar as both aim to achieve as high a reverse blocking voltage capability as possible with the thinnest piece of n-type silicon that can be used.

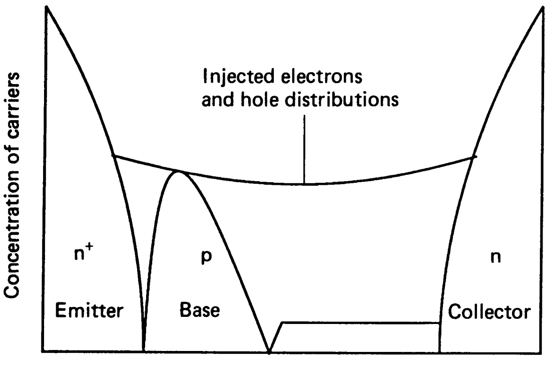

Concentration of carriers
Injected electrons and hole distributions
n+
p
n
Emitter
Base
Collector

Figure 17.8 Diffusion profile of a typical bipolar n–p–n power transistor and on-state injected carrier density

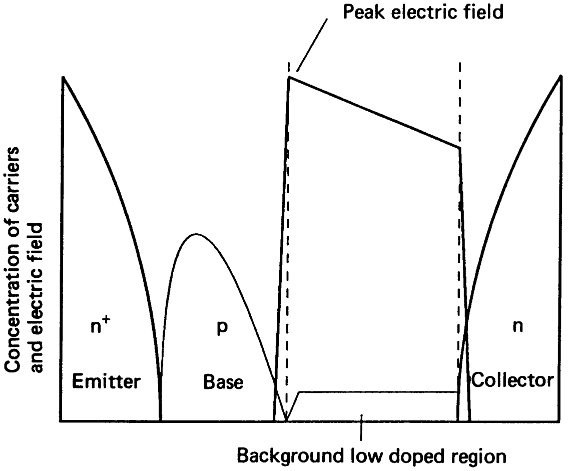

Peak electric field
Concentration of carriers and electric field
n+
p
n
Emitter
Base
Collector
Background low doped region

Figure 17.9 Diffusion profile of a typical bipolar n–p–n transistor and off-state electric field distribution

Bipolar n–p–n power transistors are usually operated in one of two alternative on-states: saturation or quasi-saturation. However, it is sometimes necessary to operate these devices in the active region shown in *Figure 17.10* for certain applications, despite the fact that these devices are not designed and, therefore, not optimised for use in this region. Under these circumstances only a very reduced power-handling capability is possible. In the saturated on-state the collector emitter voltage V_{CE}. is less than that for the forward bias voltage which appears across a simple silicon

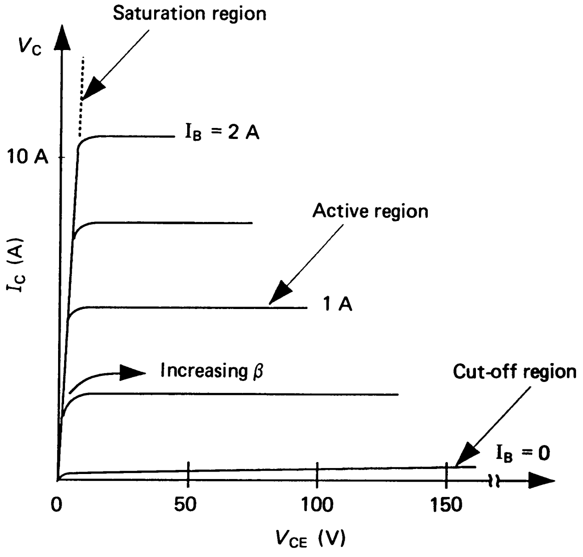


Figure 17.10 Output characteristics of the bipolar junction transistor in the common emitter configuration

p–n junction such as a p–n junction diode. This is because the collector base junction becomes forward biased at high collector current and this subtracts from the voltage across the base emitter junction resulting in typical V_{CE} voltages of 0.2–0.3 V. However, in this condition the whole of the base and collector regions of the bipolar n–p–n power transistor are saturated with injected carriers. When a device is switched off under these conditions a considerable amount of charge has to be extracted and this is associated with a long storage time, t_S. If the collector voltage is held above 1.5 V the device is said to be in quasi-saturation and the collector base junction is prevented from becoming forward biased.

To ensure safe switching of a bipolar n–p–n transistor between the off- and on-state a number of basic rules have to be observed. In addition, a large bipolar n–p–n power transistor has a number of important features that can best be understood by reference to *Figures 17.11* to *17.16*.

Figure 17.11 shows the important waveforms involved in resistive load switching of a typical bipolar power transistor. Line *A–C* in *Figure 17.11*(*a*) shows a voltage-current excursion during both turn-on and turn-off. Turn-on and turn-off switching energy losses can be reduced by minimising the transition time from point *A* to point *C*. The delay time (t_d) and rise time (t_{ri}) as defined in *Figure 17.11*(*b*) can be reduced by increasing the turn-on base current (I_{B1}) defined in *Figure 17.13* and, thereby, the turn-on losses can be reduced. The fall time (t_{fi}) defined in *Figure 17.11*(*b*) of a saturated transistor can be reduced by optimising $\mathrm{d}I_{B2}/\mathrm{d}t$, the rate of application of I_{B2}. Optimisation ensures simultaneous cut-off of the emitter-base and collector-base junctions of the transistor. Once this has occurred the transistor enters the quasi-saturated region. Thereafter, a large value of I_{B2} may be safely applied and this ensures a minimum fall-time and, thereby, reduces the turn-off losses. A transistor already operating in the quasi-saturation region may have I_{B2} applied as rapidly as possible, without damage.

Figure 17.12 shows the waveforms involved in switching an inductive load. The switch-on waveform shown in *Figure 17.12*(*a*) is indicated as that part of the switching locus $A \rightarrow B \rightarrow C$ and the switch-off waveform $C \rightarrow D \rightarrow A$. The freewheel diode ($D_f$) conducts load current (I_L) when the transistor turns off. The diode is assumed to exhibit reverse-recovery characteristics and have an initial load current I_M. Consider first the turn-on sequence of the transistor. Assume D_f is initially in conduction when V_{CC} is applied across the transistor and stray lead inductance. D_f starts to recover but appears initially as a short circuit. When the diode recovers the voltage across the transistor (V_{CE}) falls from V_{CC} to the on-state value and the transistor current increases to full load current value, as shown in *Figures 17.12*(*b*), *17.12*(*c*) and *17.12*(*d*). The rate of rise of this current is determined by the magnitude and rate of rise of I_{B1} and by transistor gain. High-gain transistors driven by large base currents exhibit the lowest turn-on losses. During turn-off the diode will remain non-conducting until V_{CE} reaches the supply voltage V_{CC}. Throughout the rise in V_{CE} all the load current flows through the transistor but, once V_{CE} reaches V_{CC}, I_C drops to zero and D_f carries on conducting the load current. At this point V_{CE} can exceed $\mathrm{V_{CC}}$ by around 10–20 V due to the finite turn-on time of D_f. The total V_{CE}–I_C excursion during inductive switching is as shown in *Figure 17.12*(*a*). By reducing the collector current switching time, energy losses can be reduced and the safe operating area extended. The safe operating area is that portion of V_{CE}–I_C switching boundary graph which a load line can traverse without damaging the transistor.

Figure 17.13 shows the relevant base current waveforms. I_{BI} should initially be high during charging of the emitter-base capacitance C_{BE}, that is usually 2–3 times the value required during the full turn-on phase. During free-wheel diode recovery, the transistor will carry the recovery current causing it temporarily to enter the linear region of the V_{CE}–I_C characteristic. During the conduction phase, sufficient base current must be applied to keep the transistor fully saturated or quasi-saturated. This current is about 1.5 times that for saturation of a minimum-gain transistor. Application of a negative base current I_{B2} will remove charge from the base-emitter junction. $\mathrm{d}I_{B2}/\mathrm{d}t$ must be adjusted so that the emitter-base and collector-base junctions are cut off together. For inductive load switching, as with resistive loads, the quasi-saturated state does not suffer from this limitation.

High-frequency applications tend to favour the quasi-saturated mode of operation. Large I_{B2} values reduce storage and fall times but, once the ratio $I_{B2/B1}$ exceeds 3, little further improvement is observed.

In order to improve the voltage hold-off capability and control any $\mathrm{d}v/\mathrm{d}t$ effects, a reverse bias of at least 3 V needs to be maintained across the emitter-base junction. It is permissible to operate this junction in the avalanche region, provided that power levels are kept low.

Use of a ‘Bakers’ clamp as shown in *Figure 17.14* prevents transistor saturation, keeping the transistor in the quasi-saturated regions. Diode D_{as} conducts base current into the collector and prevents the collector voltage from going below a given value. Anti-saturation techniques such as this give shorter storage times and fast turn-off, since there are few excess carriers in the collector that need to recombine with the collector current.

The use of a low-value resistor R_{BE} of approximately 10–100 Ω between the emitter and base terminals increases the transistor blocking capability. Only a small current is diverted through this resistor and power loss is small. The effect of V_{CE} $\mathrm{d}v/\mathrm{d}t$ can be controlled by applying a negative base-emitter voltage during the off state. Power losses in the R_{BE} resistor then become more significant during the off state than in the transistor on state. As R_{BE} is decreased storage time is also reduced which is due to a lower emitter-base junction capacitance time constant at turn-on.

Switching of transistors, especially with inductive loads, necessarily incurs the simultaneous application of supply voltage V_{CC} and full load current I_C as shown in *Figures 17.15* and *17.16*. Violation of the permitted safe operating area may occur under these conditions which can result in the transistor being driven into secondary breakdown and thermal destruction of the device. Restrictions must therefore be placed on the operating conditions of the transistor.

To avoid simultaneous application of high voltage and high current, the V_{CE}–I_C excursion needs to be modified such that its path across the safe operating region does not produce this situation. This also minimises losses and the full rating of the transistor can be obtained. The modification is effected by components that ensure that peak current and peak voltage do not occur together during turn-on and turn-off.

A turn-on aiding circuit is shown in *Figure 17.15*. The inductance holds off and controls the collector $\mathrm{d}i/\mathrm{d}t$ during turn-on. If the inductor used is non-saturable then $\mathrm{d}i/\mathrm{d}t = \cancel{V}_{CC}/L$. The inductance L also controls the free-wheel diode recovery current.

If the inductance is saturable V_{CE} falls and the inductance initially holds off the supply voltage deficit without much current flow. The inductance then saturates to a low value and supports very little voltage. Losses in the inductance are low since the current required to magnetise it is small. In *Figure 17.15*(*c*) the saturable inductance supports the deficit

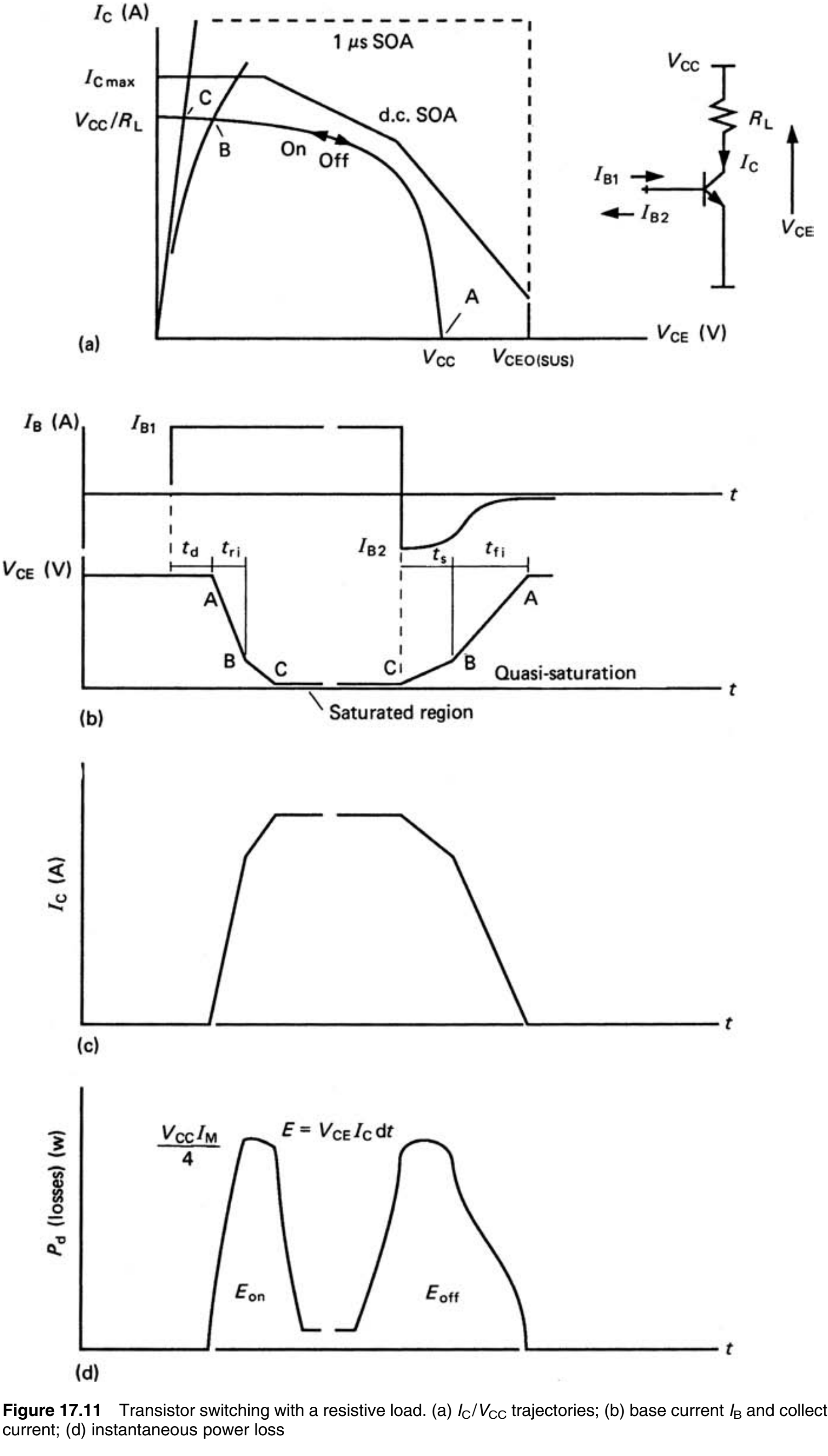

I_C (A)
1 μs SOA
$I_{C\,max}$
C
V_{CC}/R_L
B
On
Off
d.c. SOA
A
(a)
V_{CC}
$V_{CEO(SUS)}$
V_{CE} (V)
V_{CC}
R_L
I_C
I_{B1}
I_{B2}
V_{CE}
I_B (A)
I_{B1}
t
t_d
t_{ri}
I_{B2}
t_s
t_{fi}
V_{CE} (V)
A
B
C
C
B
A
Quasi-saturation
t
Saturated region
(b)
I_C (A)
t
(c)
P_d (losses) (w)
$\frac{V_{CC}I_M}{4}$
$E = V_{CE}I_C\,dt$
E_{on}
E_{off}
t
(d)

Figure 17.11 Transistor switching with a resistive load. (a) I_C/V_{CC} trajectories; (b) base current I_B and collector voltage V_{CE} (in volts); (c) collector current; (d) instantaneous power loss

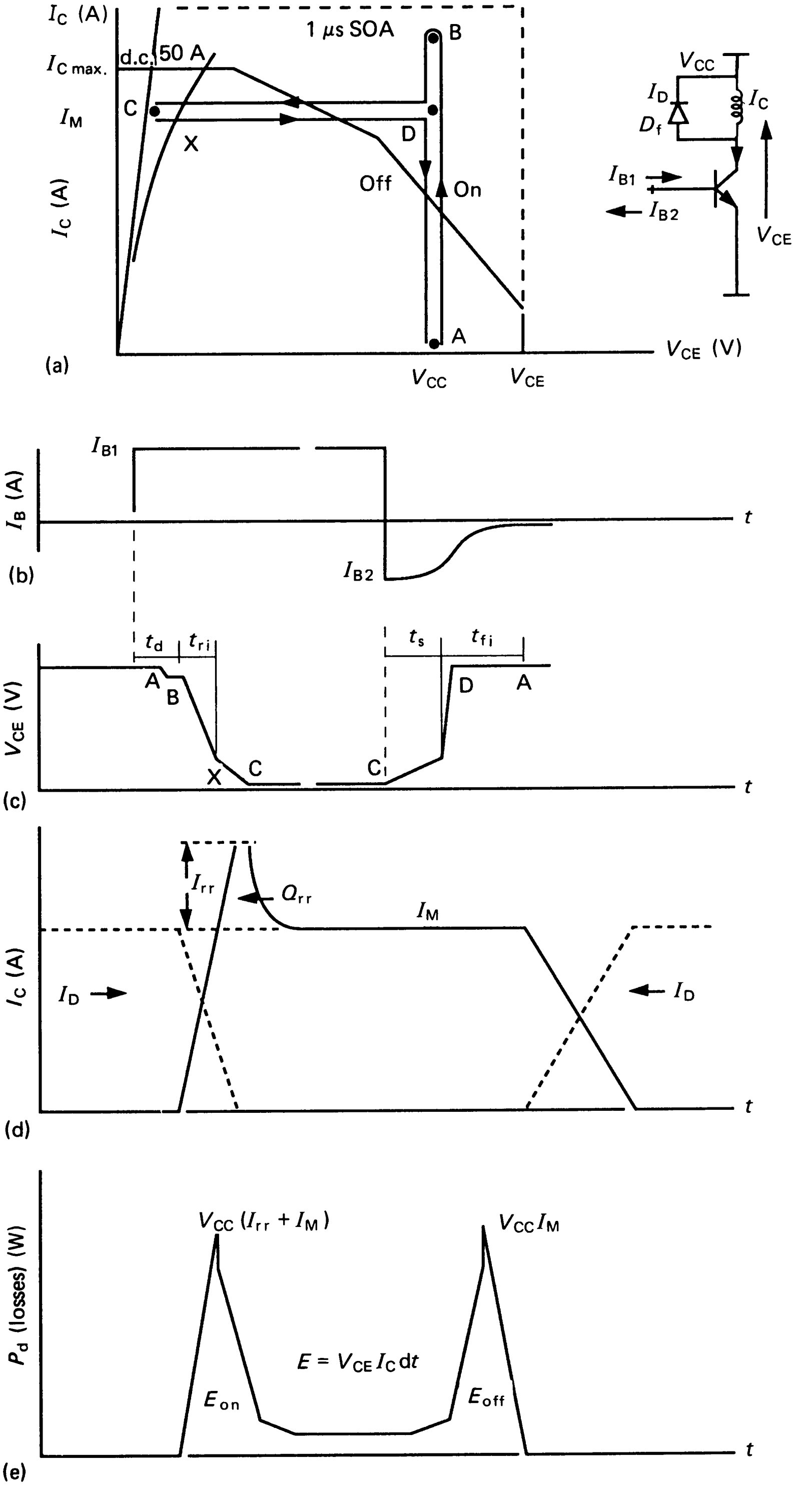

I_C (A)
$I_{C\,max.}$
I_M
I_C (A)
d.c.
50 A
1 μs SOA
B
C
X
D
Off
On
A
(a)
V_{CC}
V_{CE}
V_{CE} (V)
V_{CC}
I_D
D_f
I_C
I_{B1}
I_{B2}
V_{CE}
I_B (A)
I_{B1}
I_{B2}
t
(b)
V_{CE} (V)
t_d
t_{ri}
t_s
t_{fi}
A
B
X
C
C
D
A
t
(c)
I_C (A)
I_{rr}
Q_{rr}
I_M
I_D
I_D
t
(d)
P_d (losses) (W)
V_{CC} (I_{rr} + I_M)
$V_{CC}I_M$
$E = V_{CE}I_C\,dt$
E_{on}
E_{off}
t
(e)

Figure 17.12 Transistor switching with an inductive load. (a) I_C/V_{CC} trajectories; (b) base current; (c) collector voltage; (d) collector current; (e) instantaneous power losses

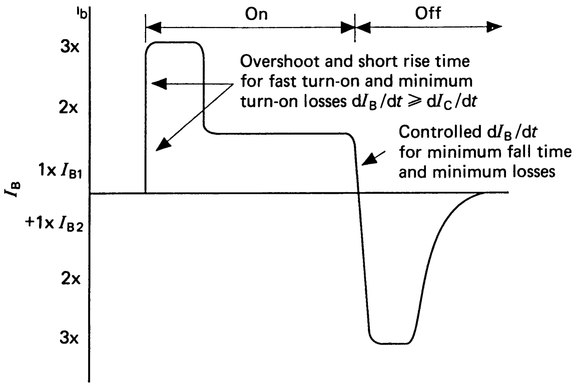


Figure 17.13 Transistor base current switching waveform

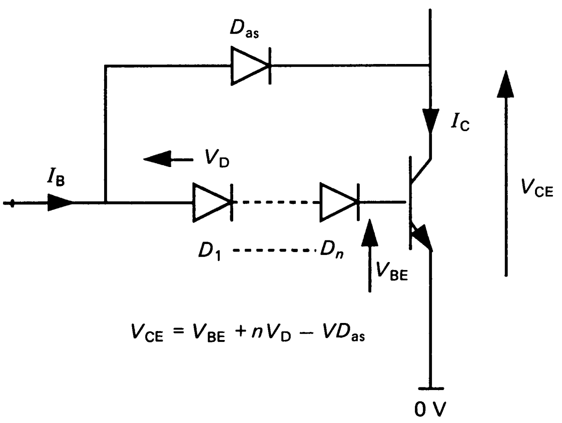


Figure 17.14 The Bakers clamp circuit

supply voltage until V_{CE} has fallen to zero. Once the inductor is saturated, the collector circuit non-load inductance becomes very low. The free-wheel diode reverse recovery d*i*/d*t* is very large, possibly forcing the transistor into the linear region as shown in *Figure 17.15*(*c*) and the addition of a series non-saturable air core inductor may be required.

A resistance R_s in series with the diode speeds up the current decay. Similarly a diode/zener configuration will constrain the inductor voltage and also reduce the current decay time.

A turn-off aiding 'snubber' circuit and the waveform excursions developed are shown in *Figure 17.16*. At turn-off current is diverted from the transistor to the snubber capacitor C_S with the collector voltage clamped to the capacitor voltage. At turn-on D_s blocks and C_S discharges current through R_s and the transistor. The power losses in the resistor are given by $P_c = ½C_S\ V_{CC}^2f$.

If the collector voltage rise time t_{rv} is minimal compared to the I_C fall time (t_{fi}), then maximum energy loss during switching in an unaided transistor is given by $W_T = ½I_C\ V_{CC}\ t_{fi}$.

Losses are less for snubber-aided transistors because switching occurs at low V_{CE} voltages, allowing more efficient transistor operation. As C_S, is increased, more energy is diverted from the transistor to the capacitor. The energy stored in C_S can become very large and must be dissipated as heat in R_S during the subsequent turn-on phase. At high voltages and switching frequencies the power rating of R_S will become large, making it more desirable to recover the energy in C_S.

The 'snubber' circuit maintains the transistor load inside the safe region of the curve RBSOA, increasing device efficiency, since most of the energy dissipated in the snubber would otherwise be dissipated by the transistor.

The size of the snubber capacitor determines the snubber effectiveness and its purpose is to yield a desired rate of rise of V_{CE} at a given peak I_C. The time interval (+*t*) of interest for this calculation is that measured from the instant V_{CE} starts to rise at turn-off until I_C approaches zero. This time interval also is dependent on reverse-base current applied to the transistor.

R_S is chosen to ensure that C_S will discharge completely during the conduction time of the transistor.

The power dissipated in a snubber resistor is given by

$$P_d = 1/2V_{max}^2C_S \times f(\text{watts})\Leftarrow$$

where *f* is the frequency of operation. The snubber capacitor increases the fall time of the transistor by slowing the rate of rise of V_{CE}. The crossover point of voltage and current is lowered by the use of a snubber, indicating a reduction in peak power dissipation, as indicated in *Figure 17.16*. A more complete description of the bipolar n–p–n transistor is given in reference 1 and of the application of the device in reference 2.

17.2.2 Bipolar n–p–n power Darlington transistor

In spite of some disadvantages the monolithic, or one-chip, Darlington transistor continues to be extremely popular for power transistor switching applications. Its main advantage is its high gain with a minimum number of components. The gain of the Darlington is given by the following equation for an unstabilised Darlington

$$\beta\varsigma = \frac{I_C}{I_{B1}} = \beta_1 + \beta_2 + \beta_1\beta_2$$

where β_1, is the gain of transistor T_1 and β_2 is the gain of transistor T_2 in the simple Darlington configuration shown in *Figure 17.17*(*a*).

The addition of stabilising resistors R_1 and R_2 as shown in *Figure 17.17*(*b*), which are normally included in the silicon chip, improve storage time but reduce gain. The gain of the Darlington is then given by

$$\beta\varsigma = \beta_1(\beta_2 + 1) + \beta_2\left(1 - \Leftarrow\frac{V_{BE2}}{R_2I_{B1}}\right)\Bigg]$$

when the effect of R_1 is neglected.

To prevent the Darlington coming out of saturation under overload current conditions, it is necessary to overdrive the base as illustrated in *Figures 17.18* and *17.19* and, although the device gain fails, the saturation voltage V_{CED} can be maintained at an acceptable level:

$$V_{CED} = V_{CE1} + V_{BE2}$$

At turnoff however the input transistor T_1 displays a storage time t_{S1} after which I_{C1} drops resulting in a decrease in I_{B2}. The output from T_2 is quasi-saturated with a storage time t_{s2}, and from *Figure 17.20* it can be seen that

$$t_{SD} = t_{S1} + t_{S2}$$

As I_{B1} increases with I_C held constant, forced gain decreases and t_{SD} increases (*Figure 17.21*). Since the output transistor

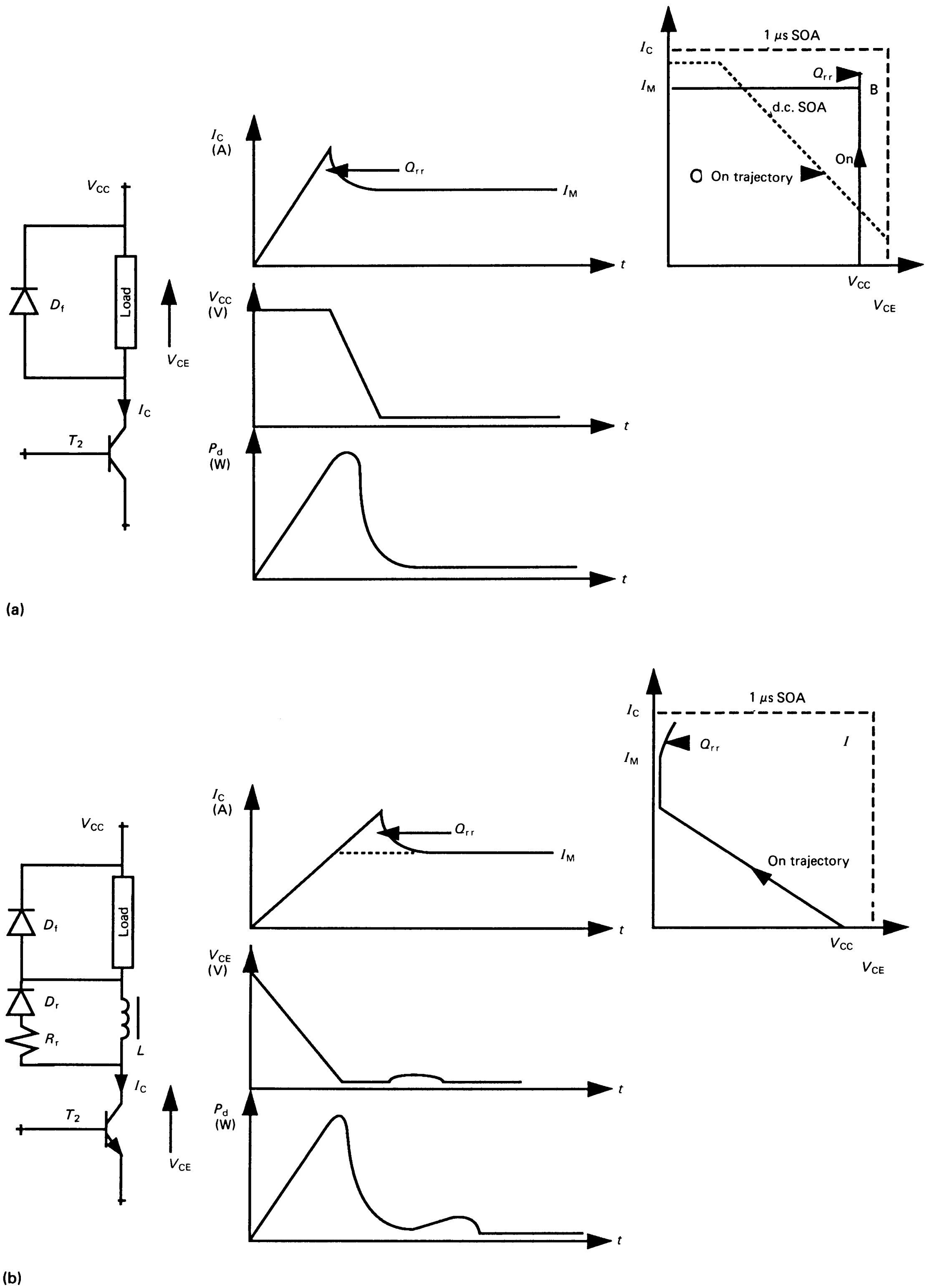
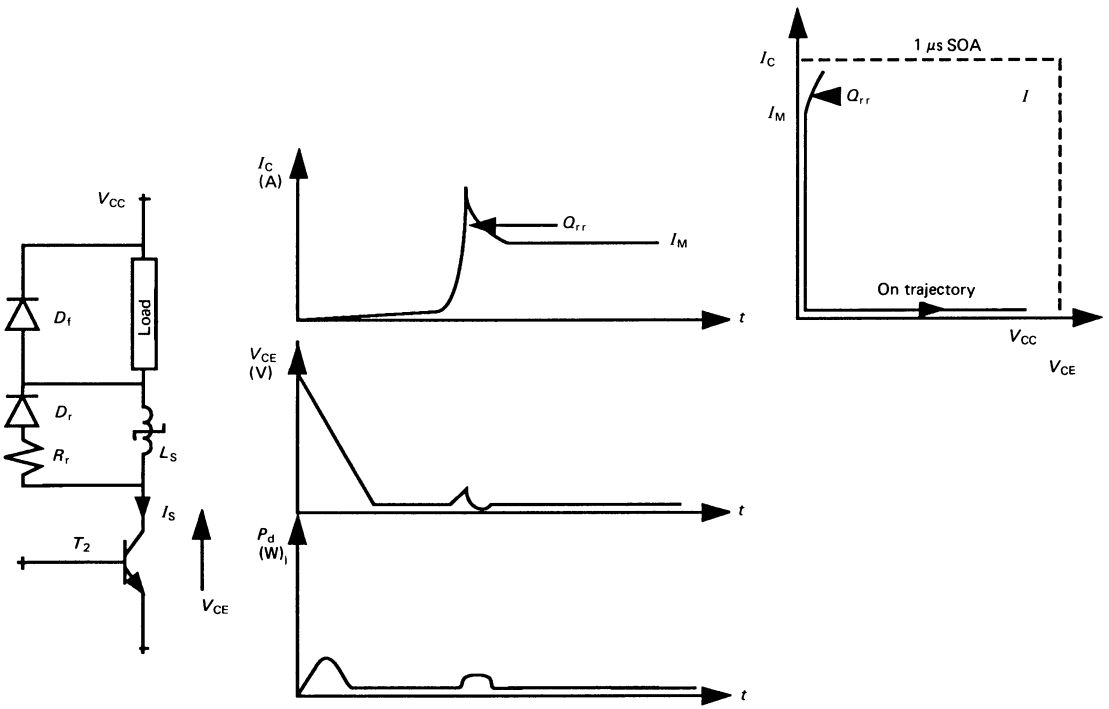

V_{CC}
D_f
Load
V_{CE}
I_C
T_2
I_C (A)
Q_{rr}
I_M
t
V_{CC} (V)
t
P_d (W)
t
I_C
1 μs SOA
Q_{rr}
B
I_M
d.c. SOA
On
O On trajectory
V_{CC}
V_{CE}
(a)
V_{CC}
D_f
Load
D_r
R_r
L
I_C
T_2
V_{CE}
I_C (A)
Q_{rr}
I_M
t
V_{CE} (V)
t
P_d (W)
t
I_C
1 μs SOA
Q_{rr}
I
I_M
On trajectory
V_{CC}
V_{CE}
(b)

Figure 17.15 Transistor switch-on: (a) without a switch-aiding circuit; (b) with a switch-aiding circuit; (c) with a saturable reactor switch-aiding circuit



(c)

Figure 17.15 (*continued*)

is in quasi-saturation, its storage time t_{S2} is independent of forced gain. Therefore, overdriving the Darlington (*Figure 17.17*) only influences the storage time of the driver transistor t_{S1}. This can be reduced if a negative base current is withdrawn from T_1. However, if the ratio of negative to the positive base current is increased, the effect on t_{SD} diminishes, as shown in *Figure 17.22*. Any further reduction in t_{SD} can only be brought about by reducing t_{S2} that can be achieved by fitting a speed-up diode D_1, as shown in *Figure 17.23*. When D_1 is fitted the emitter-base junction of T_1 is by-passed. A negative base current will then flow into T_2 if a negative bias $-V_{BB}$ is applied to R_1, which reduces T_{S2}, but only after the emitter-base diode of T_1 begins to become reverse biased. Current flowing in D_1 then reverse biases the emitter-base junction of T_1. However, the reverse-bias voltage may be insufficient to allow efficient turn-off of T_1. Therefore, several diodes may be connected in series to increase the reverse bias on T_1.

The Darlington turn-off sequence is complete after the fall time t_{fD}. This is dependent only on T_2 (T_1 having been previously turned off). The requirement for optimum turn-off of T_2 is dependent on simultaneous cut-off of its emitter-base and collector-base junctions that can be achieved by carefully controlling the value of $\mathrm{d}I_{B2}/\mathrm{d}t$.

17.3 Thyristors

17.3.1 The basic thyristor

The basic thyristor structure is shown as a two transistor analogue model in *Figure 17.24*(*a*). As can be seen from this figure, the thyristor is a four-layer junction silicon device that performs as a controllable rectifier. Functionally, it can be described in terms of two internally connected complementary transistors, with current gains, α_{npn} and α_{pnp} operative within a positive feedback loop. When the sum of their gains is raised to unity, the forward-biased thyristor switches from a high- to a low-impedance state. In normal operation, current injection through the third terminal or gate initiates this regenerative action. Two-terminal breakover also occurs, as shown in *Figure 17.24*(*b*) by carrier multiplication when the forward bias is increased to a critical value defined as the breakover voltage V_{BO}.

The vertical doping concentrations and the four semiconductor layers of the thyristor are shown in *Figure 17.25*. Forward- and reverse-blocking voltages are supported across the charge-depletion layers that develop around junctions J_2 and J_3 of the thyristor shown in *Figure 17.25*. The vertical layers of the p–n–p–n structure must be designed, therefore, in terms of widths and impurity concentrations to accommodate the depletion regions: this is achieved by the correct specification of the high-resistivity n-type starting material and of the detailed sequential diffusion and fabrication processes used to form the active structure. Under reverse bias, the anode is negative with respect to the cathode and both junctions J_1 and J_3 are reverse biased. Because the p and n regions at J_1 are heavily doped this junction avalanches at a low voltage, and so the device behaves in essentially the same manner as a p–n junction diode, with a low leakage current being maintained up to a critical breakdown voltage.

Today, production thyristors are available with voltage ratings up to about 6000 V. Even higher voltages are feasible, but only at the expense of severe de-rating on other important parameters, including current rating and switching speed.

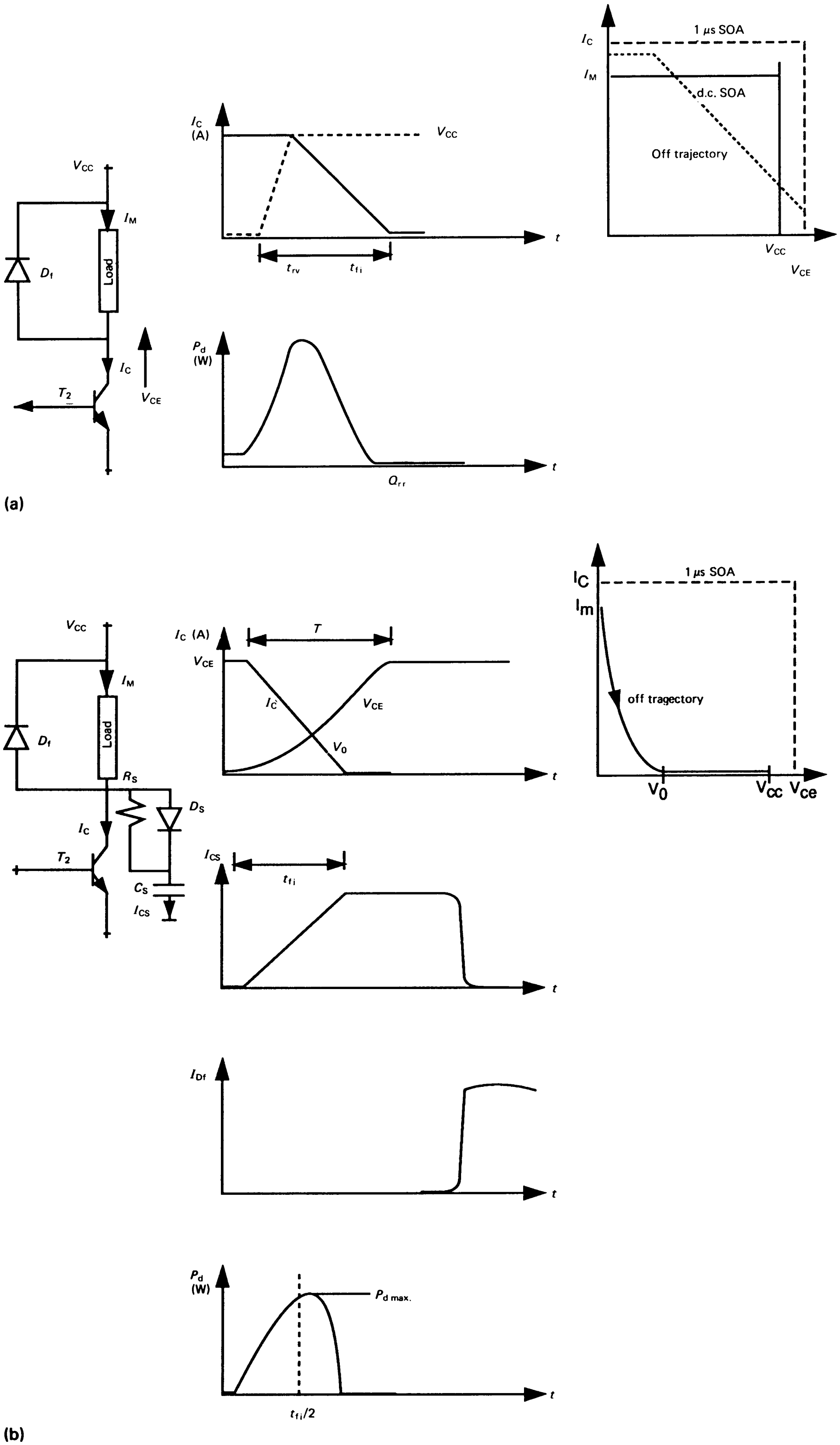

(a)
V_{CC}
I_M
Load
D_f
I_C
V_{CE}
T_2
I_C (A)
V_{CC}
t
t_{rv}
t_{fi}
P_d (W)
t
Q_{rr}
I_C
I_M
1 μs SOA
d.c. SOA
Off trajectory
V_{CC}
V_{CE}
(b)
V_{CC}
I_M
Load
D_f
R_S
D_S
I_C
T_2
C_S
I_{CS}
I_C (A)
T
V_{CE}
I_C
V_{CE}
V_0
t
I_{CS}
t_{fi}
t
I_{Df}
t
P_d (W)
$P_{d\,max.}$
t
$t_{fi}/2$
I_C
I_m
1 μs SOA
off trajectory
V_0
V_{cc}
V_{ce}

Figure 17.16 Transistor switch-off: (a) without a switch-aiding circuit; (b) with a snubbing active-capacitor switch-aiding circuit

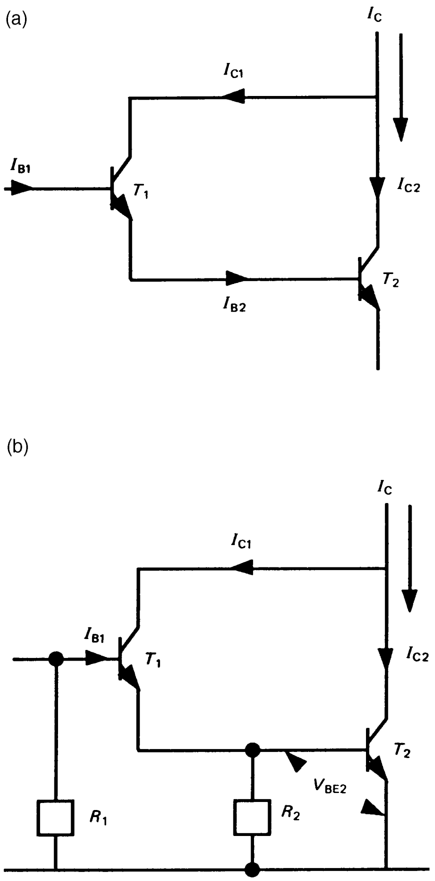


Figure 17.17 (a) Simple Darlington configuration; (b) stabilised

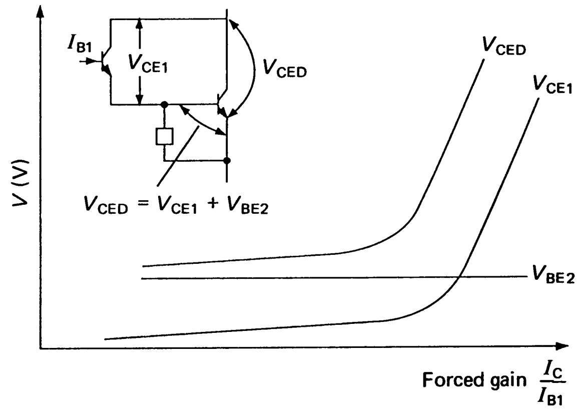


Figure 17.18 Effect of forced gain on Darlington collector emitter saturation voltage

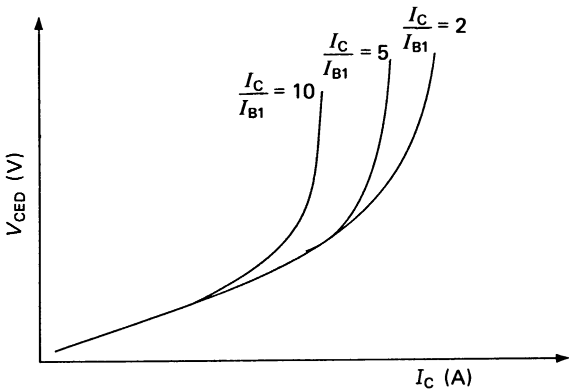


Figure 17.19 Effect of forced gain on $V_{CE(sat.)}$

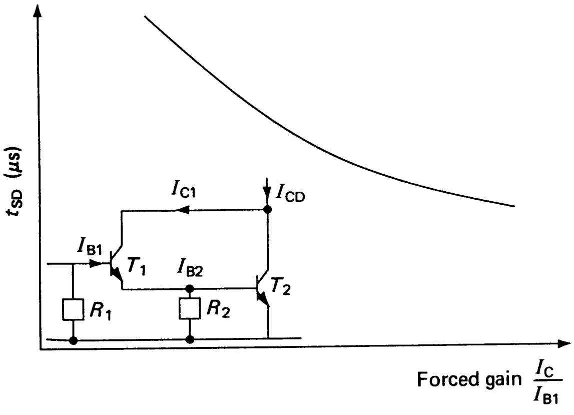


Figure 17.20 Darlington storage time as a function of forced gain

The limit to reverse blocking is set by avalanche multiplication of carriers traversing the high electric field in the space charge region of J_3 that occurs at a critical value of maximum field. A wide depletion layer is thus necessary to support a high blocking voltage, and this in turn implies the use of high resistivity, or lowly doped n-type starting material. Although complicated to some extent by the loop-gain, similar considerations apply in setting the level of forward-blocking voltage.

This simple picture is true within the bulk of the thyristor structure, but neglects a factor of crucial importance: this is the termination of the p–n junctions at the silicon surface. Unless corrective steps are taken, electric-field enhancement can occur locally at the surface with a lowering of the breakdown voltage and destructive channelling of the avalanche current. The solution of this problem is by ensuring adequate field relief at the surface. *Figure 17.26* shows one approach that can be used, that remains the basis for many of today's high-voltage 'alloyed' thyristors. The edge of the silicon pellet has a compound bevel, with a surface at angle θ_1 to the junction plane to control reverse blocking, and a surface at θ_2 to control forward blocking. These sections are frequently termed 'positive' and 'negative' bevels, respectively. The exact measure of the field relief afforded depends on both the diffused concentration gradients and the surface conditions, including the density of surface charge and the dielectric properties of the surrounding ambient. The positive bevel acts by spreading the equipotentials in the high-resistivity n-base region, and a fourfold reduction in peak field is obtained with a bevel angle of about 20°. Likewise, the negative bevel acts by equipotential spreading in the diffused p-region. Its effect is less pronounced and more dependent on the dopant concentration

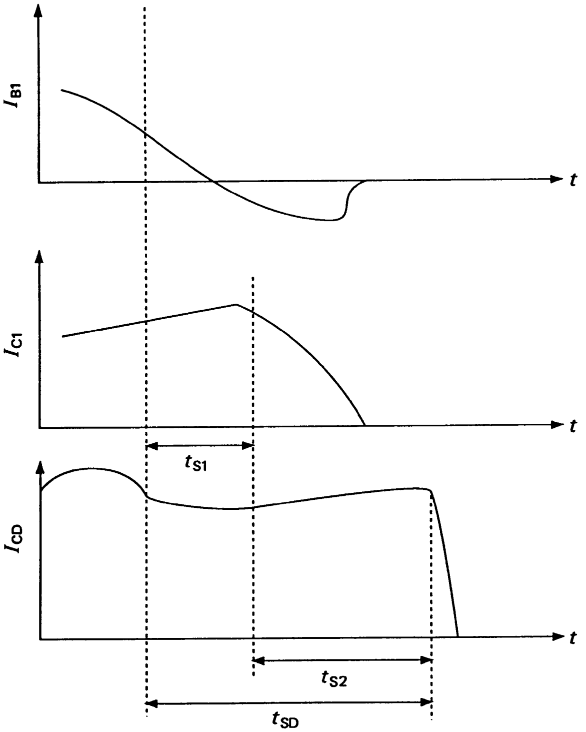


Figure 17.21 The relationship between base current (I_{B1}) the driver transistor collector current (I_{C1}) and storage time (t_{S1}), Darlington current (I_{CD}) and the Darlington storage time (t_{SD})

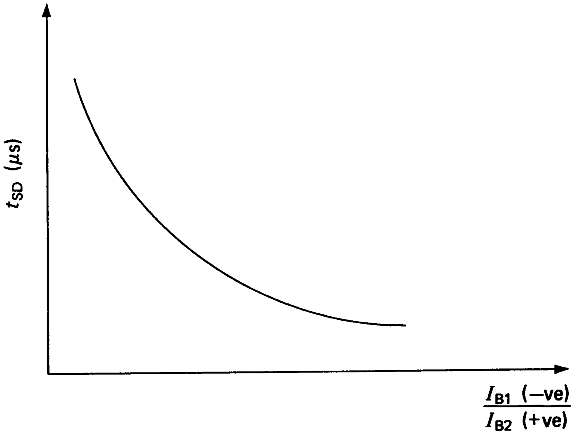


Figure 17.22 The effect on t_{SD} of a negative bias applied to the driver transistor

gradient close to the junction. Low-angle bevels, typically 2–4°, are necessary.

Other types of edge termination are also used on higher power higher voltage thyristors: these include double negative bevels and double positive bevels. These techniques are generally applied to the so-called 'fully floating' thyristors. In the 'alloyed' device the anode contact is provided by a permanently 'alloyed' molybdenum disc, with the cathode contact using being a direct pressure contacted disc. In the free floating type both the anode and the cathode surfaces are contacted using a pressure plate. From the user perspective the alloyed type can offer a better transient thermal response, but it presents difficult manufacturing problems, particularly at high current ratings owing to the large diameter of the alloyed joint. In broad terms devices below 75 mm silicon diameter are more usually alloyed, those of larger size more generally fully floating.

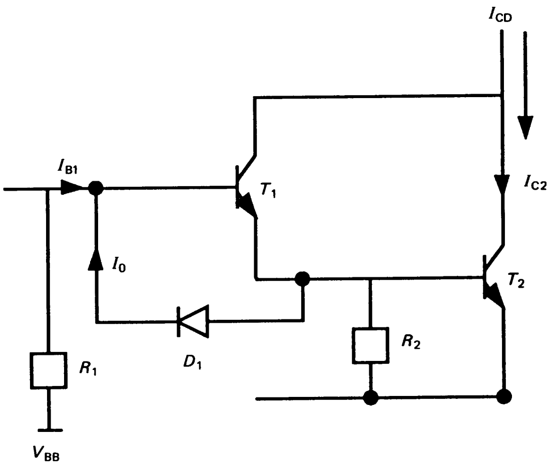


Figure 17.23 The reduction of t_{S2} by means of an anti-parallel diode connected across the emitter-base junction of a driver transistor

The thyristor conducting area is clearly one key factor in determining the on-state voltage, mean and surge current ratings of a thyristor. Also, when the thyristor is in the on-state, all three junctions are forward biased and the anode-to-cathode voltage drop approaches 1 V at low current densities. As the current is increased, additional non-linear voltage drops arise within the base regions, so that the on-state voltage drop becomes particularly dependent on their widths, and the injected carrier lifetime within them. (For a brief description of p–n junctions and carrier injection see Section 17.1.) Both high-voltage or wide-base, and fast-switching or low-lifetime design requirements imply an increase in on-state voltage drop, and hence reduction in current density rating.

The electrical performance of the thyristor is also dependent on the mechanical design and the thermal performance of the housing in which the silicon pellet is encapsulated. Typically high power thyristors are assembled in press pack housings such as that shown in *Figure 17.2(a)*. But smaller devices use standard stud based or plastic packages. Moving outwards from the silicon pellet, low-resistance non-rectifying contacts must be made to the three electrode regions. As well as carrying the load current without a significant voltage drop, these contacts provide the path for heat flow away from its source within the silicon to an external heat sink or cooling system. This cooling is vital, since all the thyristor characteristics are to some degree temperature sensitive, and the maximum operating temperature of the silicon must be limited to about 125°C. Power ratings depend, therefore, on the thermal resistance between the silicon and the external heat sink, which is cooled by either natural convection or forced-fluid methods. Thermal resistance from junction to mounting base achieved by these techniques ranges from about 3°C/W for a low-power stud-mounted thyristor to below 0.02°C/W for a high-power double-side-cooled encapsulation.

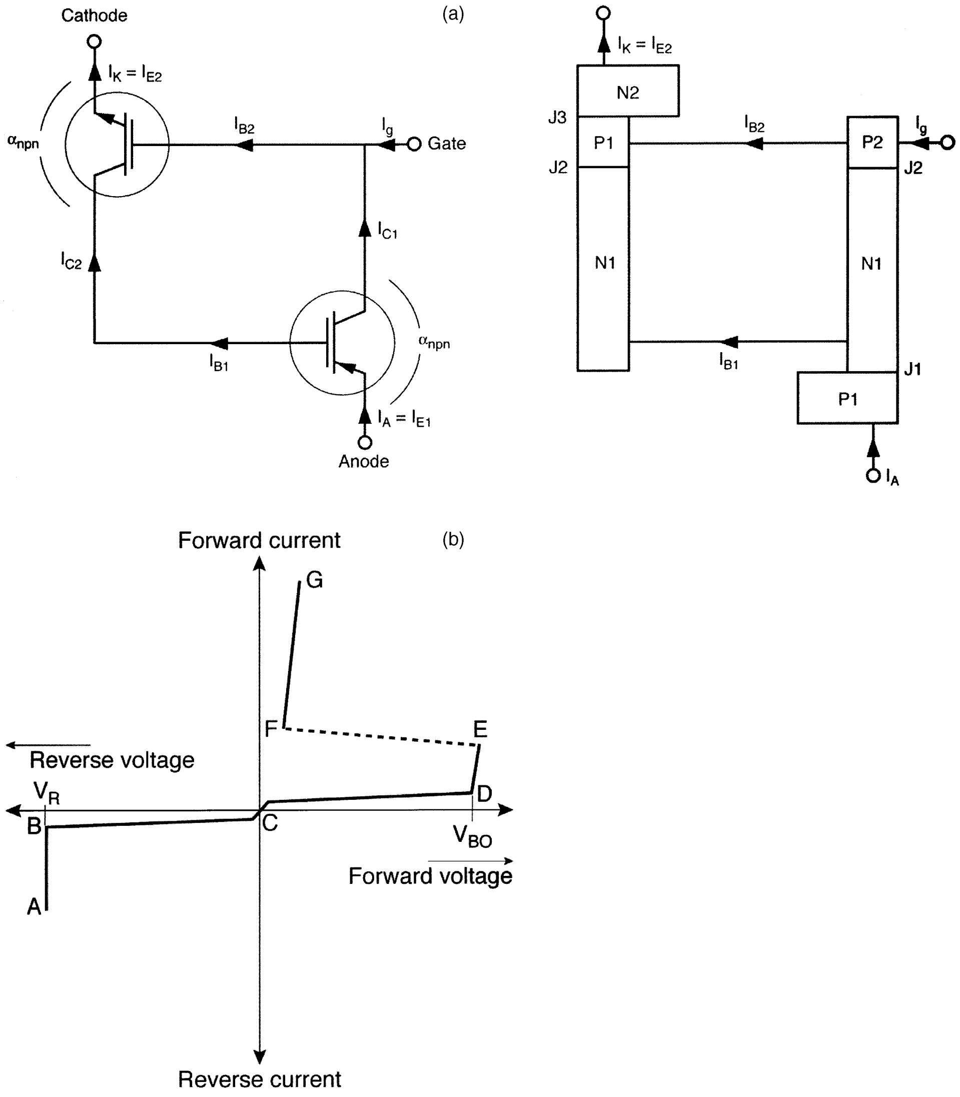


Figure 17.24 (a) Two transistor model of the thyristor; (b) Thyristor current-voltage characteristics

Long-term operational reliability is also a vital feature of all thyristor designs. The packages are hermetically sealed by techniques including electric resistance, projection welding, argon arc welding and cold pressure welding, after thorough vacuum out gassing of all components and filling with a dry nitrogen ambient. The materials used in the packaging are arranged so that the device can withstand many temperature excursions and as a result thyristors can be extremely reliable and survive under extreme conditions for several decades.

It is also possible to deposit a special passivation on the silicon surface, which allows highly reliable thyristor construction in moulded plastic packages similar in external appearance to those used for high power IGBTs (*Figure 17.2*(*b*)). An important application of these protected silicon pellets is in assemblies that provide most cost-effective arrangement for many medium and low-power control applications.

The switching behaviour of the thyristor is explained with reference to *Figure 17.27* where it is defined in terms of the turn-on and turn-off times, together with the associated dI/dt and dV/dt ratings. The dI/dt rating is the maximum permissible rate of increase of load current, and the dV/dt rating is the maximum rate of rise of off-state voltage that will not initiate conduction.

The three phases of turn-on are shown in *Figure 17.27*. The delay time is associated with establishment of transistor, and hence regenerative, action in response to the gate current. Its duration is set by the level of gate drive and by the minority carrier transit times across the transistor bases. A high gate drive reduces the delay time, and wide base regions, or high blocking-voltage rating, increase it. Following the delay time the voltage across the thyristor collapses. Regeneration is well established during the rise time, with charge modulation of the base regions aiding carrier transport so that the rate of current build-up is more

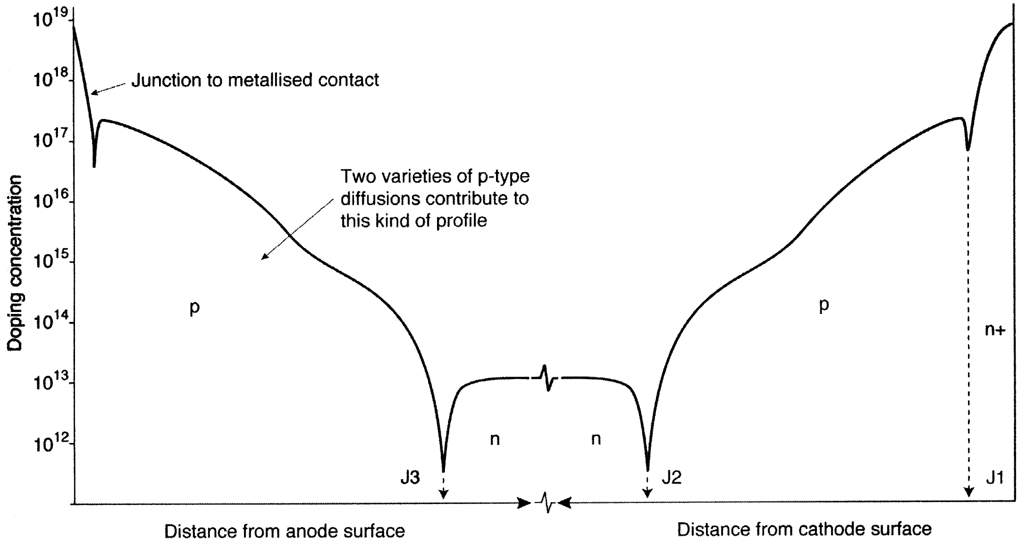


Figure 17.25 Doping concentration across section of a thyristor. Junction positions J_1, J_2, J_3 are shown

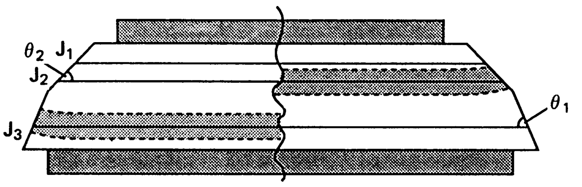


Figure 17.26 Section of a thyristor compound edge bevel. The depletion layer is shown around J_3 under reverse bias on the left-hand side, and around J_2 under forward bias on the right-hand side

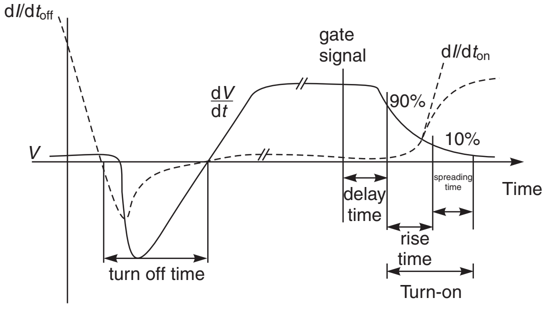


Figure 17.27 The anode-to-cathode and voltage waveforms during thyristor turn-on and circuit-commutated turn-off

rapid. Again, longer rise times are to be expected in high-voltage thyristors with wide base regions, particularly when it is remembered that the collapse of the centre-junction depletion layer during turn-on widens the effective base widths. Finally, there is the spreading time. A fundamental feature of thyristor turn-on is its three-dimensional nature. Because of the finite transverse or sheet resistance of the p-base region, the initiating gate current influences only those regions of the cathode nearest to the gate contact. The gate contact is most usually in the centre of the thyristor. Regenerative switching action is restricted to the turned on regions close to the gate thus formed a narrow conducting plasma. The establishment of the equilibrium current flow over the cathode area follows by outward spreading from this conducting plasma by field-aided diffusion. Plasma spreading is a relatively slow process, with a typical ‘velocity’ of 0.1 mm/μs. When the area of conduction is small the anode/cathode voltage is considerable, leading to high turn on power densities and excessive local heating may result if the current is allowed to increase too rapidly (di/dt).

This turn on spreading phenomenon is affected not only by geometrical factors but also by base widths and such other factors as the recombination rate of carriers within them. The spreading velocity is decreased in both high-voltage and fast turn-off structures.

Where higher di/dt performance is required the amplifying gate structure shown in *Figure 17.28*(*a*) is employed. It may be viewed as the integration of two radially disposed thyristors with a common anode contact. The initiating gate pulse turns on the inner small area, or auxiliary thyristor, this draws current from the external load and provision is made to feed this current back into the silicon via an overlapping metal contact. This amplified current provides the gating current for the main power thyristor. Once the latter is switched, the auxiliary structure is extinguished due to lateral self-biasing voltages. In this way, the repetitive dI/dt rating can easily be increased to about 500 A/μs or greater and switching losses are reduced.

An extension to the amplifying-gate principle is to increase further the area of initial conduction by extending the edge length of the auxiliary thyristor coupling contact and ultimately incorporating an interdigitated gate structure on the silicon surface, as shown in *Figure 17.28*(*b*). The technique is applicable to the highest power large area thyristors. During the on-state all three junctions of the thyristor are forward biased, and both base regions contain excess minority and majority charge. In switching off the thyristor and reverting to the blocking state, charge must either be swept out by an electric field, or decay through regenerative processes within the silicon.

Practical circuits use anode commutation, or bias reversal, to turn off a thyristor, and this is represented by the waveforms as shown in *Figure 17.27*. When the reverse voltage

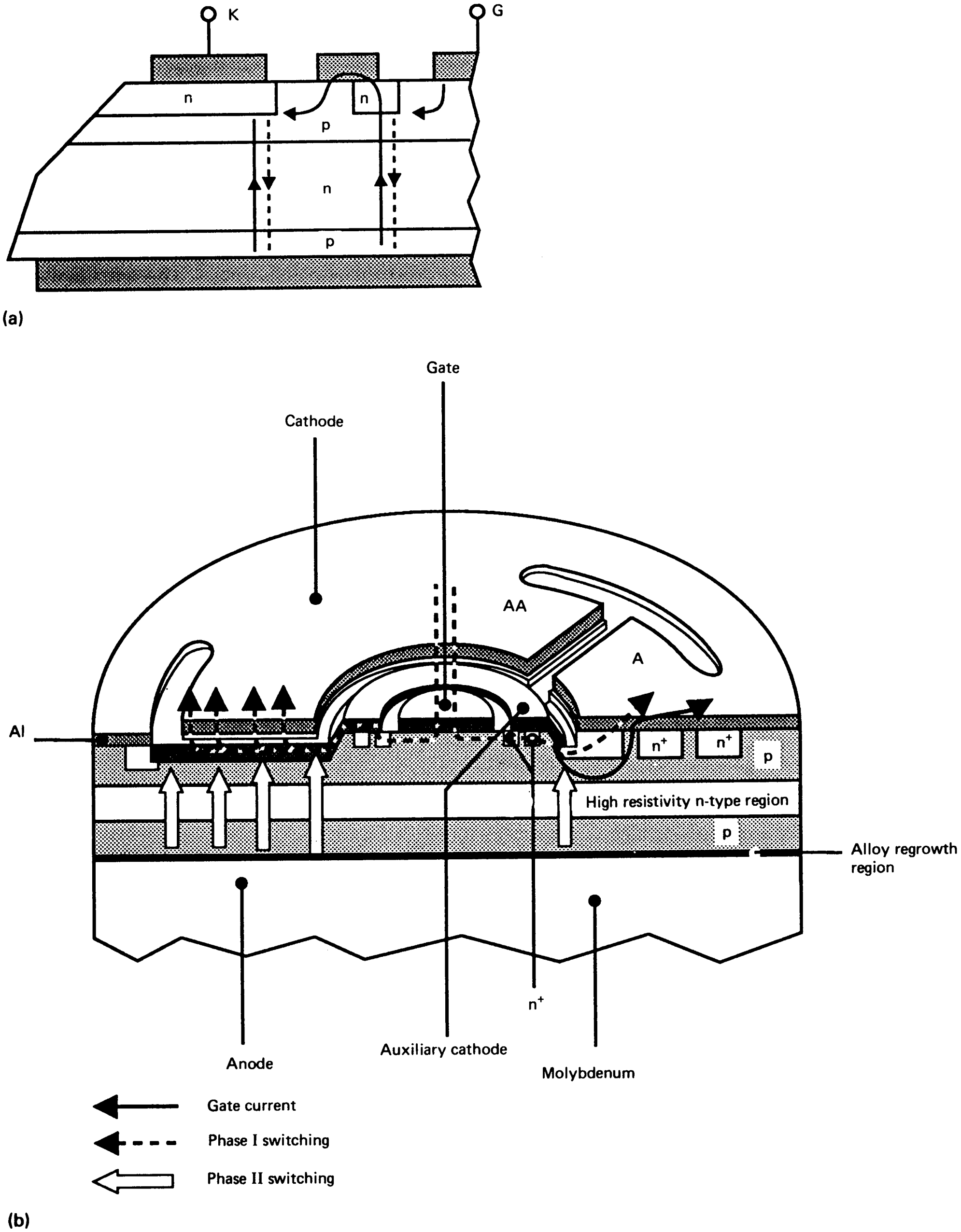

K
G
n
n
p
n
p
(a)
Gate
Cathode
AA
A
Al
n+
n+
p
High resistivity n-type region
p
Alloy regrowth region
n+
Anode
Auxiliary cathode
Molybdenum
Gate current
Phase I switching
Phase II switching
(b)

Figure 17.28 (a) Amplifying gate thyristor; (b) interdigitated amplifying gate thyristor

is applied it forces the current to fall to zero at a rate defined as dI/dt. Once it reaches zero, current flow reverses since the minority carrier concentration at the junctions can support this current as they are extracted before depletion-layer build-up. The peak value of reverse current is reached when the excess hole concentration at the anode junction has fallen to zero. At this time the voltage across the thyristor reverses, with development of the anode depletion layer, and the current decays in a near-exponential manner as a result of charge recombination within the n-base region. It must fall to, at most, the holding current if the thyristor is to block when forward voltage is reapplied. How long this takes is critically dependent on the mean lifetime of injected carriers in the n-base region, and on any extraction of stored charge (in the n-base region) into the load circuit or out from the gate connections.

Special steps are taken to reduce the n-base minority-carrier lifetime in thyristors designed for use in high-frequency applications. Gold, diffused through the structure, or radiation damage is used to provide a controlled concentration of impurity centres or traps through which the injected carriers recombine. Any method used to reduce minority-carrier

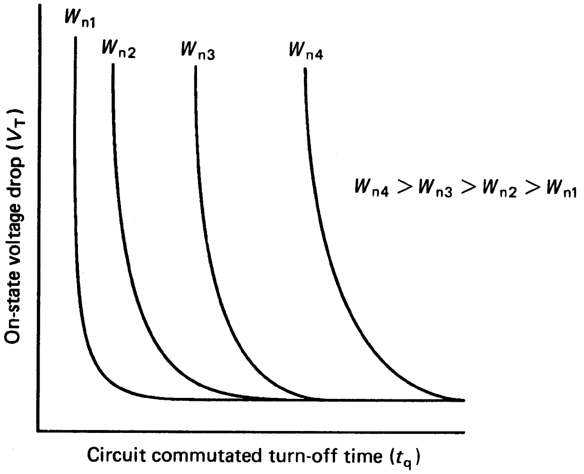


Figure 17.29 The relationship between on-state voltage (V_T) and circuit commutated turn-off time (t_q) for fast switching and gold doped thyristors of different n-base width (W_n) and hence different voltage rating

lifetime leads inevitably to some increase in the on-state voltage drop, principally through a reduction in conductivity modulation of the n-base region. As the recombination centre concentration is raised to reduce the turn-off time, this effect assumes increasing significance, particularly in the case of wide-base, high-voltage devices. The form of the relationship between on-state voltage and turn-off time is illustrated in *Figure 17.29*.

Under normal operation the forward-blocking voltage will be re-applied to the thyristor at a specified rate, or d*V*/d*t* as shown in *Figure 17.27*, and this gives rise to a uniform charge-displacement current as the depletion layer develops around the centre junction. Its magnitude is simply the product of the voltage-dependent junction capacitance, and the value of d*V*/d*t*. If no steps are taken to counteract this, the cathode junction will become forward biased by the flow of this displacement current, and this will lead to thyristor switching action.

The so-called 'shorted-emitter' technique shown in *Figure 17.30* is effective in preventing this two-terminal switching up to high values of d*V*/d*t*. It is a method for controlling carrier injection by structural design via a regular array of small-area electrical short circuits in the n emitter to the underlying p region. These 'shorts' are usually in the form of an array of distributed dots over the cathode emitter. All high power thyristors use this technique. The displacement current flows laterally to these points and bypasses the emitter regions, this minimises the forward bias developed at the cathode junction. The limit on d*V*/d*t* is set by the ohmic voltage drop caused by transverse displacement-current flow in the p-base region. When any part of the cathode emitter is biased above about 0.6 V, it starts to inject and the thyristor may switch on. By correct selection of shorting dot diameters, spacing and array symmetry, d*V*/d*t* ratings exceeding 1000 V/μs can be realised without sacrificing much gate-triggering sensitivity. However, some loss in gate sensitivity is inevitable since a fraction of the gate signal will always flow to shorts near to the cathode edge and play no part in the turn-on process. With correct design this can be minimised and the threshold gate current to fire the device kept below a reasonable value of 100–150 mA. There will also be some degradation to the on-state voltage as the emitter shorts give an effective reduction in the cathode emitter efficiency.

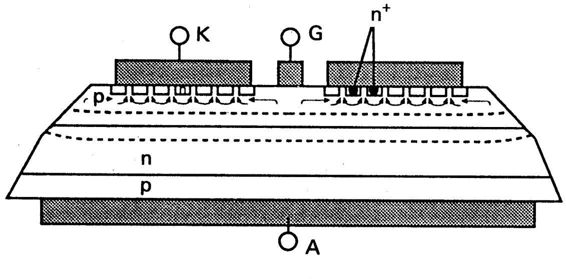


Figure 17.30 Distributed shorting of the thyristor–cathode junction to permit high rates of application of forward-blocking voltage or d*V*/d*t* rating

From overall design considerations an array of closely spaced minimum area shorts is preferred for a given d*V*/d*t* rating. They then occupy the smallest fraction of the cathode area and have the least effect during on-state conduction, and provide the minimum impedance to conduction spreading during turn-on.

The turn-off performance of thyristors can also be improved through the use of increased n-emitter shorting. When the shorting density is in excess of that needed to meet the required d*V*/d*t* withstand, forward anode voltage can be re-applied earlier in the charge-decay period, and the residual charge extracted safely from the n base. This is because during turn off the emitter shorts give direct access to the base regions of the thyristor which assists charge extraction and reduces any emitter injection during the turn off period. In this way it is possible to realise high-power fast-switching thyristors for use at frequencies in excess of 10 kHz. A more complete description of thyristors is given in the references 1, 3 and 4 at the end of this chapter.

17.3.2 The converter thyristor

The converter thyristor is designed and optimised primarily for operation at 50 or 60 Hz (mains frequency) and is available up to 5000 V and 3500 A average current rating. The basic design of the converter thyristor is the same as the basic thyristor described above. The introduction of gold or electron irradiation recombination centres is not normally necessary in this type of device, where the main objective of the design is to achieve the lowest on-state voltage possible for the voltage rating required. However many high power applications interconnect several thyristors in series where both static and dynamic voltage sharing is essential between the thyristor levels. In this case it is necessary to achieve an accurate control of the recovered charge during turn off: this control is usually achieved by gold diffusion or radiation damage in a similar manner to that used for fast recovery diodes. As with the fast diode this results in a compromise between low on state voltage, recovered charge, and leakage current.

17.3.3 The fast thyristor

Fast thyristors are designed and optimised for operation frequencies above 400 Hz and up to above 20 kHz. The basic design principles of fast thyristors are the same as those for general thyristors as described above. Their maximum operating frequency is dependent on the switching losses during turn on and turn off, and the turn-off time. The latter is influenced by the intensity of the cathode emitter shorting, the concentration of the recombination centres introduced in the same n-base region, i.e. the minority carrier lifetime, and also by the base width as shown in *Figure 17.29*. The turn on performance is improved by the use of interdigitated amplifying gates (*Figure 17.28*), and the use of narrow high gain base regions. As these devices are used under high frequency conditions their safe operation is determined by the impact of switching losses, and the designer requires detailed rating data as shown in *Figure 17.31*. The frequency-rating curves ((*a*) and (*b*)) give the

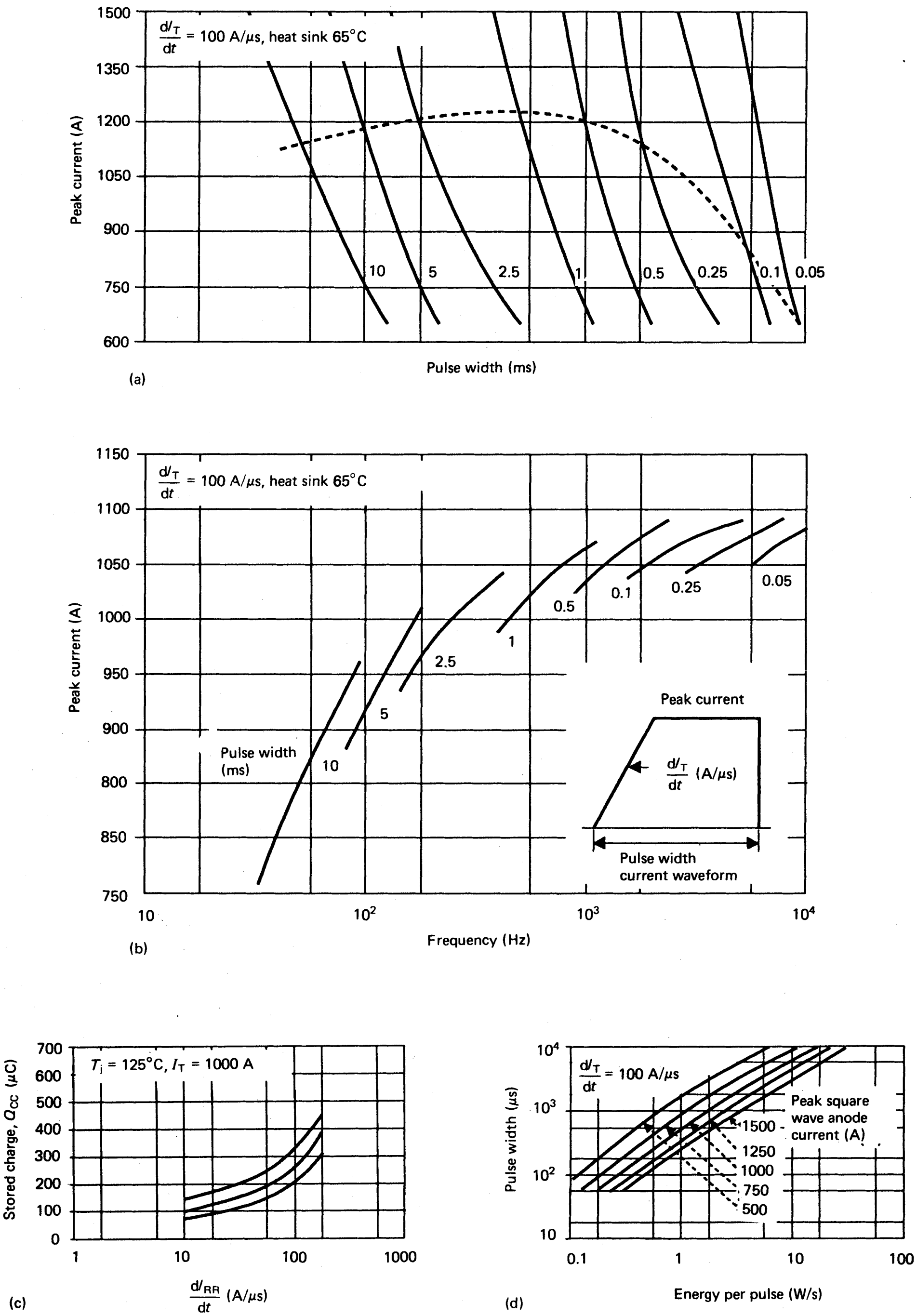


Figure 17.31 Examples of thyristor square-wave rating data. For ratings tests the gate-source impedance is 10 Ω, with a gate-pulse rise time to 1 A of 0.5 μs. The thyristor is switched from 600 V and a resistive capacitive (RC) snubber network of 10 Ω, 0.25 μF is connected

limiting case for operation under a particular forward-conduction duty, to which must be added the reverse recovery losses derived from stored-charge curves (*c*). From this total average power loss, together with the steady-state thermal impedance, the maximum thyristor case temperature can be determined to retain the junction temperature within its defined rating. Data on the energy dissipation per pulse as a function of pulse base width are also needed to specify the allowable pulse width for a particular duty and an example of such curves is given in *Figure 17.31*(*d*).

Fast thyristors are available with voltage ratings of 600–3000 V, average current ratings upto 2000 A, turn-off times down to 5 μs and d*I*/d*t* ratings over 1000 A/μs.

17.3.4 The asymmetric thyristor

The basic principle of the design of the p–i–n structure which is incorporated in the construction of asymmetric thyristors is similar to the p–i–n diode and is explained in Section 17.1.5. The other features of the basic thyristor structure are retained as shown in the section through an asymmetric thyristor shown in *Figure 17.32*. In this device, the n-base width is minimised for a particular forward-blocking voltage rating by making use of the limited base-width approach of p–i–n rectifiers and an additional heavily doped n^+-layer is included within the n-base region of the device. As junction J_3 is now a p^+n^+ junction it cannot support a high reverse voltage. By sacrificing this reverse blocking capability which is not required in many applications a significant advantage can be obtained in an improved combination of forward blocking characteristic, on-state voltage and turn-off time. Asymmetric thyristors are available with voltage ratings up to 3000 V, or average current ratings over 500 A and turn-off times down to 2 μs.

Asymmetric thyristors may be found in two types, fast turn on and fast turn off. The fast turn on asymmetric thyristor is used in power conversion applications, such as a resonant converter, where the device doesn't require forced commutation, or turn off. In this case the device can operate at high frequencies, in the 20 kHz and above range, at high power, and most efficiently. A further useful application of such a device is as a pulse power component. The fast turn off device is an extension of the fast turn off thyristor described in the previous section, this type has the advantage of even faster turn off, at the expense of no reverse voltage, and increased manufacturing complexity and cost.

17.3.5 Triacs

The triac, or bidirectional thyristor, was evolved specifically for low to medium current a.c. load control as, for example, in lamp-dimming circuits. Its construction (shown in *Figure 17.33*) comprises two thyristors in inverse parallel within a five-layer n–p–n–p–n structure. A compound gate region is also included with balanced trigger sensitivities in both the so called first and third electrical quadrants with either a positive or a negative polarity gate bias. The main design problem with a triac is two fold; firstly the two thyristors interact and this will severely limit the turn off and d*v*/d*t* capability of the component, and secondly the gates are placed on the same surface, resulting in difficulties in triggering one of the two thyristors. Yet although the triac poses its own set of design compromises, it can be made by conventional all-diffused thyristor processes. In general, triac ratings were limited to about 50 A r.m.s., 800–1000 V and of only real use in main frequency applications due to their d*v*/d*t* and switching loss limitations but they provide the lowest-cost solution in some a.c. control circuits. For higher power applications, thermal and other considerations make an inverse pair of individual thyristors a more satisfactory proposition.

17.3.6 Light fired thyristors

Light fired or triggering of power thyristors offers the advantage of eliminating the normal gate-firing circuit and overcomes the problems of high-voltage isolation normally associated with gate-firing circuits. Light firing also offers advantages in applications in environments where electrical

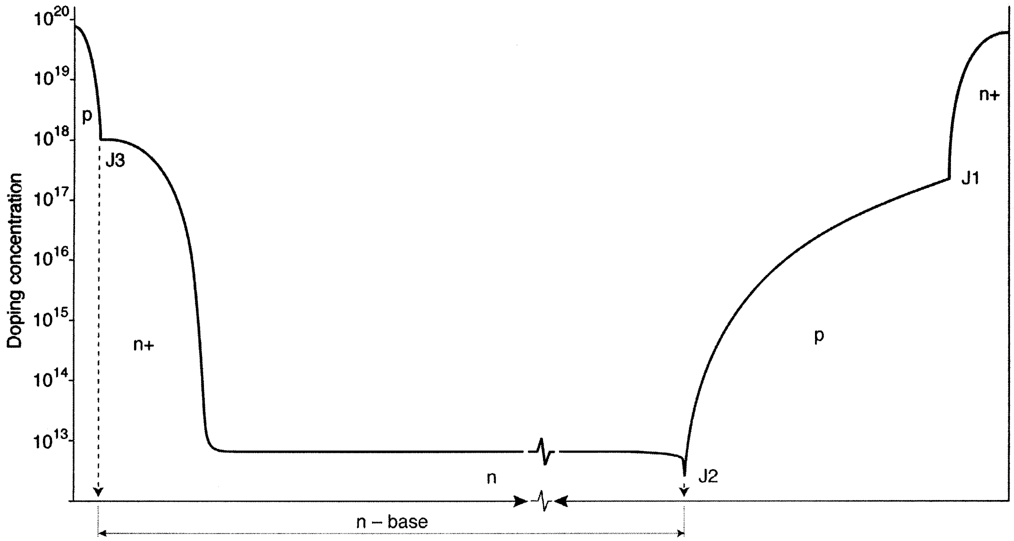


Figure 17.32 Doping concentration across a section of an asymmetric thyristor

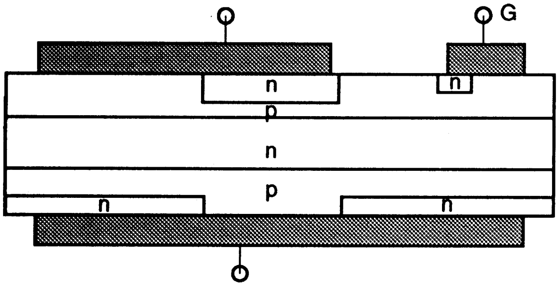


Figure 17.33 The bidirectional thyristor or triac for a.c. load control

noise is likely to cause problems by interfering with gate signals.

In the optical-triggering process, the light pulse generates electron and hole pairs. These pairs induce a current pulse that is amplified in turn via an amplifying gate region of the thyristor. This provides sufficient gate current to turn-on the thyristor.

Various light sources can be used such as the light emitting diode (LED) or laser. The light from these sources is pulsed at the appropriate moment into an optical fibre that delivers the light directly to the silicon surface to trigger the thyristor. One of the main problems is the reliability of the light source that is often not as good as the high-power thyristor. Special high-reliability LEDs have been developed for this application. However, although in principle this type of thyristor offers many advantages, it has not been widely used. This is partly because of the limited number of manufacturers who produce this type of device but also because the inclusion of the light-sensitive region on the thyristor adds to the cost and complexity of the device. A more complete explanation of light-fired thyristors is given in reference 3. The main application area for light triggered thyristors was for ultra high voltage electric utility equipment such as HVDC converter stations.

17.3.7 HVDC thyristors

The HVDC thyristor is a special type which although based on the standard converter thyristor structure has several demanding specification requirements. These include high fault current capability, high di/dt, very low on state voltage, a low turn off time, and accurately controlled stored charge. Because generally each HVDC system built is very large, each system design tends to have a specially tailored HVDC thyristor produced for it, using 75–150 mm diameter silicon. The special features that might be incorporated into this type of device are, for example, self-protection in voltage breakover, excess dV/dt, low voltage triggering and light firing. However these features and the highly demanding specification add significantly to the thyristor cost. Therefore such thyristors tend to be too expensive for general industrial applications but are used in other similar applications such as static valve active reactance (VAR) compensation. They are available with up to the highest voltages, e.g. 8000 V blocking voltage and 3000 A average current rating with carefully tailored characteristics.

17.3.8 The gate turn-off thyristor

The gate turn-off (GTO) thyristor has a unique advantage over the conventional thyristor in its ability to be turned off by the application of a negative gate current. Therefore, in applying the GTO thyristor to thyristor inverter or chopper circuits, the forced commutation components are not needed. This offers the advantages of a simpler power-circuit configuration with an associated reduction in cost, volume and weight, less electrical and audible noise, and improved power-conversion efficiency. The GTO revolutionised high power AC motor drives when they became available and were the device of choice for this application prior to the advent of high power IGBT modules.

The GTO is a four-layer p–n–p–n semiconductor device similar in construction to a thyristor but with several design features which allow it to be turned on like a conventional thyristor and off by reversing the polarity of the gate signal. The turn-on action shown in *Figure 17.34* is similar to a conventional thyristor. Injection of hole current from the gate forward biases the cathode p-base junction causing electron injection from the cathode, these electrons flow to the anode and induce hole injection by the anode emitter. Injection of holes and electrons into the base regions continues until charge multiplication effects bring the GTO into conduction. As with a conventional thyristor only the area of cathode adjacent to the gate electrode is turned on initially, and the remaining area is brought into conduction by plasma spreading. However, unlike the thyristor, the GTO consists of many narrow cathode elements, such as the device shown in *Figure 17.35*. This shows the heavily interdigitated

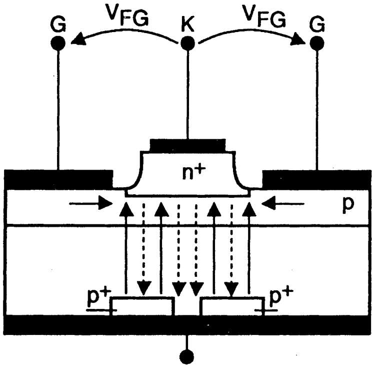


Figure 17.34 A GTO thyristor during turn-on. Turn-on is initially along the edge of the n^+ emitter, the remainder of the emitter is brought into conduction by plasma spreading

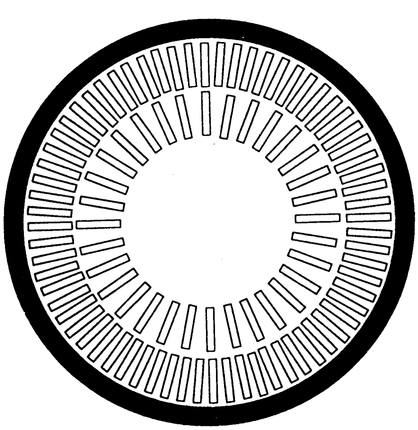


Figure 17.35 The emitter pattern of a GTO

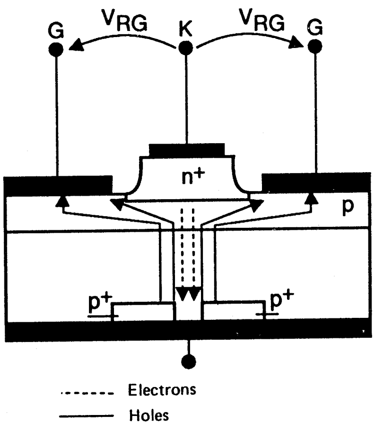


Figure 17.36 A GTO thyristor during turn-off. The gate is negatively biased and holes are extracted by the gate from the p base. The conducting region is squeezed to an area remote from the gate until sufficient charge is extracted to turn the device off

cathode and gate region. Therefore the initial turned on area is very large and the time that is required for plasma spreading is small. The GTO, therefore, is brought into conduction very rapidly and can withstand a high turn-on di/dt. This excellent turn on capability also makes the GTO attractive as a fast turn on thyristor for low frequency pulse duty applications.

In order to turn-off a GTO, the gate is reverse biased and holes are extracted from the p base, as shown in *Figure 17.36*. Hole extraction continues until the excess carrier concentration is low enough for carrier multiplication to cease and the device reverts to the forward-blocking condition. During turn-off the conducting area is squeezed down to an area remote from the gate electrode and an anode emitter short located in this area to assist turn-off. A lifetime control technique, such as gold or radiation damage is also used to reduce the device p–n–p current gain, thereby improving the turn-off power loss. The switching waveforms that can be observed in the GTO circuit shown in *Figure 17.37* are shown in *Figure 17.38*. Note, the active snubber network consisting of D_S C_S and R_S in *Figure 17.37* which restricts the dV_D/dt that appears across the GTO at switch-off. The meaning of the common terms in the switching waveform are also as shown in *Figure 17.38* (t_d, t_r, etc.). The turn-on delay time (t_d), the rise time (t_r), and the turn-on energy (E_{ON}) are all strong functions of the magnitude of the gate pulse (I_{FG}) as shown in *Figure 17.39*. To minimise the turn-on times and energy losses the value of I_{FG} should be as large as possible, consistent with considerations of the drive power loss and cost.

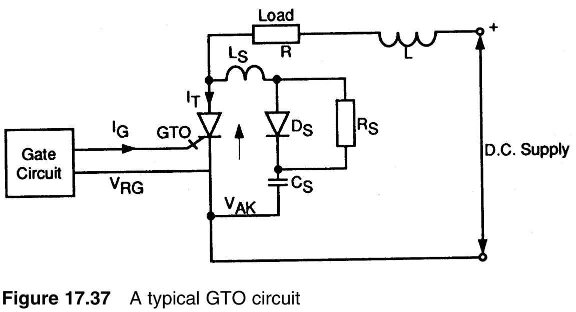


Figure 17.37 A typical GTO circuit

The turn-off of a GTO proceeds in three stages: storage time, fall time and tail time. During the storage time (t_{gs}) the conducting area is reduced due to a squeezing action of the gate drive, and the negative gate current increases to a peak value I_{GQM}. The duration of the storage time depends on how rapidly the excess charge (Q_{GQ}) is removed by the gate and is therefore a strong function of the rate of rise of reverse gate current dI_{GQ}/dt. At the end of storage time the anode current falls rapidly, and the reverse gate current begins to decay as the gate junction is forced into avalanche, $V_{(RG)BR}$. The avalanche condition is not damaging to the GTO thyristor and assists in the removal of stored charge from the device. During this period the blocking junction of the GTO recovers and the anode voltage rises. Unless a snubber circuit (as shown in *Figure 17.37*) is connected across the GTO the anode voltage could rise at a rate in excess of the device dV_D/dt capability and destroy the GTO.

After the fall time t_{gf}, the anode current has fallen to a low value I_{tail}, which decays away more slowly. This tail period is an important factor for the switching losses since the anode voltage is high at this time. In order to reduce these losses GTO thyristors use anode shorting and minority carrier lifetime control to give low values of I_{tail}.

The storage time, fall time, gate turn-off charge (Q_{GQ}) and switch-off energy are all functions of the applied negative gate current as shown in *Figure 17.40*.

In the on state the GTO operates in a similar manner to the thyristor. If the anode current remains above the holding current level then positive gate drive may be reduced to zero and the GTO will remain in conduction. However, as a result of the turn-off ability of the GTO it does possess a higher holding current level than the normal thyristor and, in addition, the cathode of the GTO thyristor is subdivided into small finger elements to assist turn-off. Thus, if the GTO thyristor anode current transiently dips below the holding current level, localised regions of the device may turn off. Forcing a high anode current back into the GTO at a high rate of rise of anode current, after this partial turn-off, could be potentially destructive. Therefore it is usually recommended that the positive gate drive is not removed during conduction but is held at a value $I_{G(on)}$ that is greater than the maximum critical trigger current (I_{GT}) over the expected operating temperature range of the GTO thyristor.

Unlike the standard thyristor, the GTO thyristor does not include cathode emitter shorts to prevent non-gated turn-on effects due to dV/dt forward-biased leakage current. In the off-state of the GTO thyristor, therefore, steps should be taken to prevent such potentially dangerous triggering. This can be accomplished by either connecting the recommended value of resistance between gate and cathode (R_{GK}) or by maintaining a small reverse bias on the gate contact. This will prevent the cathode emitter becoming forward biased and, therefore, sustain the GTO thyristor in the off-state.

Gate turn-off thyristors, or GTOs, are available with repetitive controllable current ratings up to 4000 A and 4500 V forward-blocking voltages. A more complete description of the GTO is given in references 1 and 3.

17.4 Schottky barrier diodes

A Schottky barrier diode, shown for example in *Figure 17.41*, is based on the formation of a potential barrier by the so called Schottky effect at the interface between a metal such as molybdenum, platinum, chromium or tungsten, and a semiconductor surface such as silicon that is in

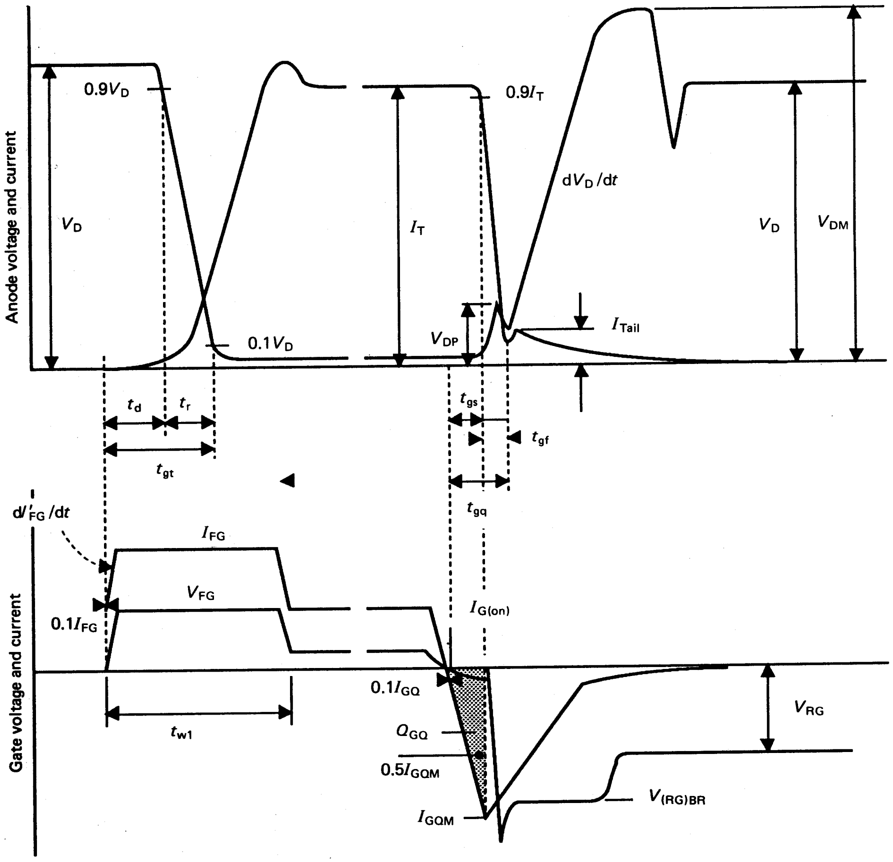

Anode voltage and current
$0.9V_D$
V_D
$0.1V_D$
I_T
$0.9I_T$
V_{DP}
dV_D/dt
I_{Tail}
V_D
V_{DM}
t_d
t_r
t_{gt}
t_{gs}
t_{gf}
t_{gq}
dI_{FG}/dt
I_{FG}
V_{FG}
$0.1I_{FG}$
t_{w1}
$I_{G(on)}$
$0.1I_{GQ}$
Q_{GQ}
$0.5I_{GQM}$
I_{GQM}
V_{RG}
$V_{(RG)BR}$
Gate voltage and current

Figure 17.38 GTO switching waveforms

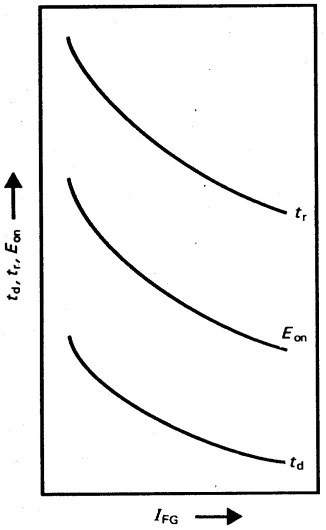


Figure 17.39 The effect of turn-on gate drive I_{FG} on t_r, t_d and E_{on}

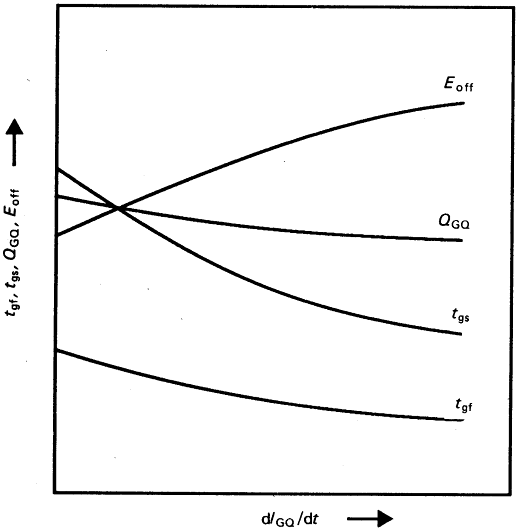


Figure 17.40 The effect of turn-off gate drive dI_{GQ}/dt on E_{off}, Q_{GQ}, t_{gs} and t_{gf}

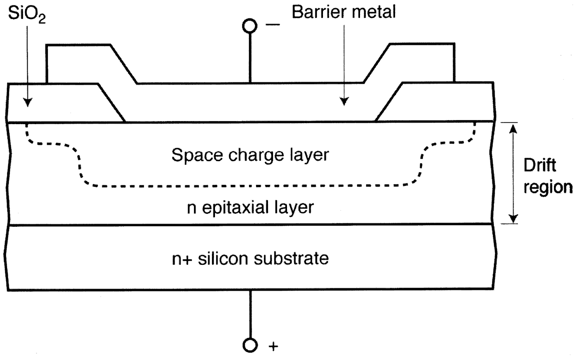


Figure 17.41 Reverse bias depleted region applied to a Schottky barrier diode

intimate contact with it. This phenomenon results in the non-linear current transport across the metal semiconductor interface. Forward conduction results from the electrons passing over the potential barrier from the n-type silicon into the metal, reverse conduction is impeded by the formation of a space charge layer. Therefore in conduction very little minority carrier charge is accumulated in the silicon and such conduction is predominantly by the majority carrier electron. The Schottky barrier diode is therefore a unipolar device that operates without the effect of injected carrier modulation. During reverse recovery it behaves almost like a perfect diode switching very rapidly from forward conduction to reverse blocking. This makes the Schottky diode an ideal device for use as a rectifying diode in very high frequency, and fast switching, applications.

In forward conduction the Schottky on-state voltage depends on the contact potential (barrier height) formed between the metal and the semiconductor and, in series with this, the resistance of the semiconducting layers (drift and substrate regions). Because there is no minority carrier injection the on-state voltage is much greater for thicker, and therefore higher reverse blocking, devices. Nevertheless for thin drift and substrate regions the on-state voltage is very low, (for example 0.3 to 0.5 V in silicon) and is therefore significantly better than the p–n diode.

The reverse off-state voltage of the Schottky diode is identical to that of an abrupt p–n junction diode, and is determined by the thickness and resistivity of the drift region. As the reverse voltage rating increases so does the silicon thickness needed to support the space charge region. Unfortunately, the reverse leakage current is also strongly dependent on the Schottky barrier height, and increases as the barrier decreases. This results in the type of characteristics shown in *Figure 17.42*.

Diffused guard ring structures are usually employed to improve reverse-blocking characteristics by reducing the electric field at the edge of the metal contact area. When combined with an optimised n-type epitaxial layer (*Figure 17.41* but guard ring not shown) the best Schottky diode performance is achieved.

Schottky barrier diodes are commercially available with reverse voltages up to 150 V and forward currents of up to 300 A. Higher reverse voltages are prevented in silicon by the high leakage current and consequential operating temperature limitations.

Silicon is not the only semiconductor material that can be used for Schottky barrier diodes, although it is the only material in commercial use at the time of writing. However there is interest in the use of silicon carbide, owing to the high breakdown field and high Schottky barrier height of this type. With further development this will lead to the availability of much higher voltage and higher current Schottky barrier diodes in the future with the potential of higher temperature operation than silicon.

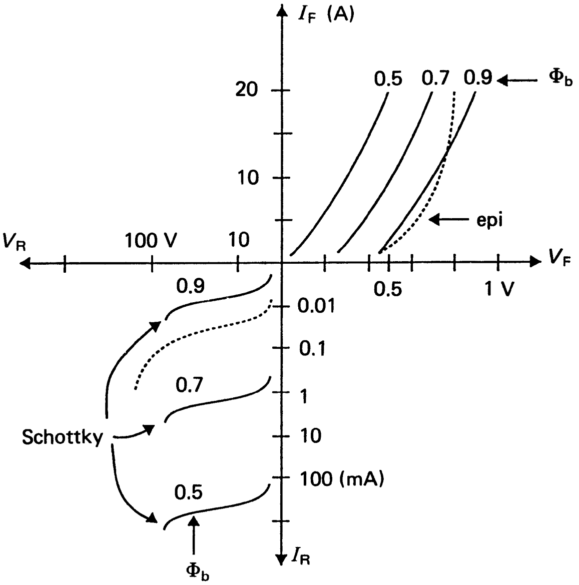


Figure 17.42 Schottky barrier diode forward- and reverse-bias characteristics

17.5 MOSFET

In the power Metal Oxide Silicon Field Effect Transistor (MOSFET) majority carriers move across an induced channel to carry forward an on-state current. This is better understood by reference to the basic MOS structure shown in *Figure 17.43*. In this structure the active MOS gate area of the device is on the silicon surface and the MOS gate is made up of three layers: the conductive layer (labelled *M* which could be a metal), the isolating layer (labelled *O* for oxide), and the semiconductor layer (labelled *S* for semiconductor). Layer *M* is the control electrode or gate. Layer *O* prevents any d.c. current flow from the gate electrode to the other two electrodes, but does allow the electric field imposed by the gate electrode to influence the surface of the semiconductor. This is how the term 'MOS' in MOSFET was originally derived. The MOS device switches on or off depending on the electric field imposed on the p-type semiconductor surface between the two n regions which act as the transistor source and drain regions. If the gate electrode is made positively biased it can induce an inversion layer on the p-zone surface and open a conductive n channel between the two n-type zones. In the on-state the residual resistance in the induced channel determines the level of on-state conduction in the device just below the silicon surface. In the off-state a depleting region is formed at one of the two p–n junctions when a bias is applied between them and the resistivity and width of the p region determines the device voltage rating. This type of structure is limited practically to a few amperes and 50–100 V although high voltage types are available for special applications. In more practicable high-power MOSFET

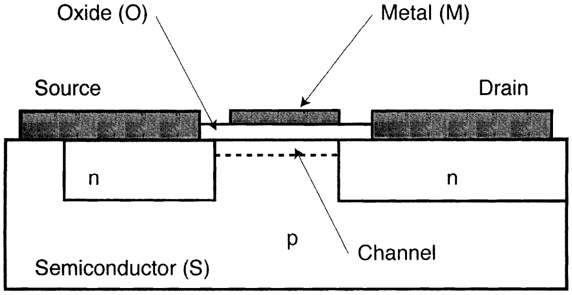

Oxide (O)
Metal (M)
Source
Drain
n
n
p
Channel
Semiconductor (S)

Figure 17.43 The enhanced n-channel in a basic MOS structure

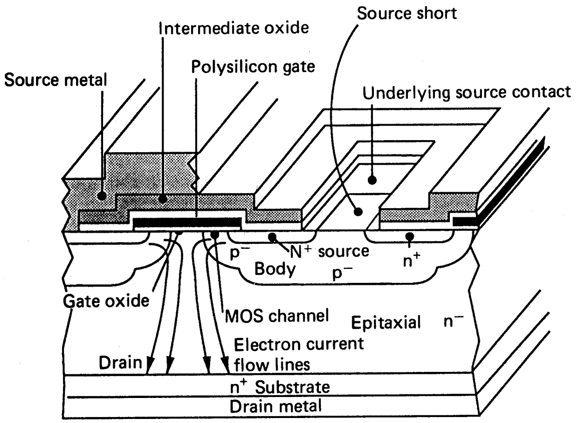

Source short
Intermediate oxide
Polysilicon gate
Source metal
Underlying source contact
p⁻
N⁺ source
Body
p⁻
n⁺
Gate oxide
MOS channel
Epitaxial n⁻
Electron current
flow lines
Drain
n⁺ Substrate
Drain metal

Figure 17.44 Detail of a DMOS structure

structures the semiconductor is exploited both vertically as well as horizontally. One of the two n-doped regions (the drain region) shown *in Figure 17.43* is extended to the bottom face of the semiconductor which in effect connects all the elementary structures together forming the drain of the device as shown in *Figure 17.44*. The p–n junction formed between the channel region and the drain also provides reverse-blocking performance with the depletion region penetrating predominantly into the lowly doped n region. At the same time, because this junction is positioned vertically, it also avoids the waste of horizontal space compared to the structure shown in *Figure 17.43* and enables the conductive channel to become very short. To obtain the performance necessary in practicable power MOSFETs this channel is normally designed to be of the order of 1–2 μm. The resulting power MOSFET device therefore consists of multiple-MOS basic cells, with all the n$^+$-type source zones connected in parallel on the top side of the semiconductor chip, together with the gate cells as shown in *Figure 17.48*. As mentioned above, the substrate of the chip then forms the common drain for the entire multiple source cells on the other side of the chip.

This vertical double diffused DMOSFET silicon gate device described above has evolved to become the design that offers the best combination of characteristics and represents a culmination of the development of the power MOSFET. Because the power MOSFET has evolved and is continuing to develop through various stages of design, it is possible to find several different types in use, illustrated in *Figure 17.46* and *17.47*.

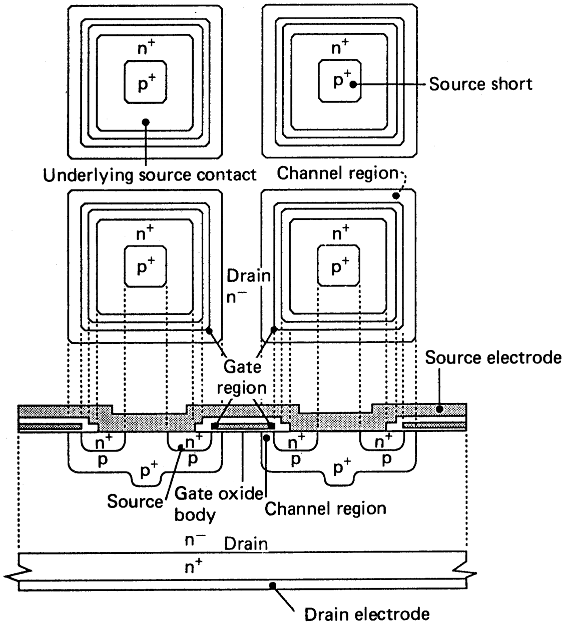

n⁺
p⁺
n⁺
p⁺
Source short
Underlying source contact
Channel region
n⁺
p⁺
n⁺
p⁺
Drain
n⁻
Gate
region
Source electrode
n⁺
p
p⁺
p
n⁺
p
p⁺
p
Source
Gate oxide
body
Channel region
n⁻ Drain
n⁺
Drain electrode

Figure 17.45 Horizontal layout and structure of a DMOSFET

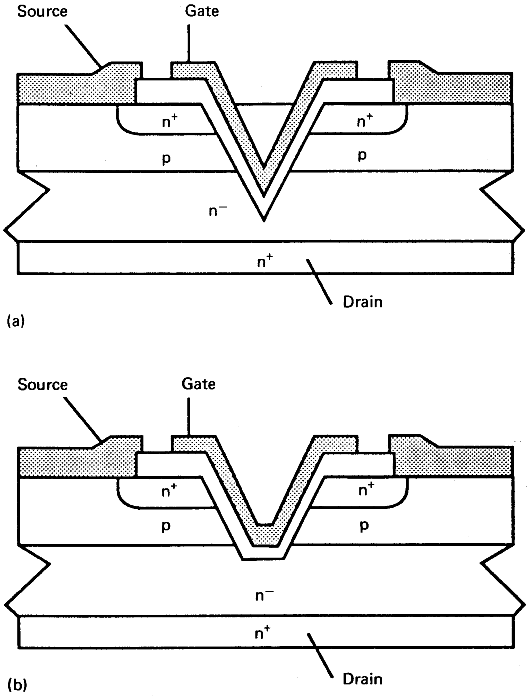

Source
Gate
n⁺
n⁺
p
p
n⁻
n⁺
Drain
(a)
Source
Gate
n⁺
n⁺
p
p
n⁻
n⁺
Drain
(b)

Figure 17.46 (a) V-groove MOSFET; (b) Trench U-groove MOSFET

All these structures have a common feature in that the current flows through a vertical path like conventional high power bipolar devices and, as a consequence, all the devices have two electrodes on the top which are the gate and source, and one on the bottom which is connected to

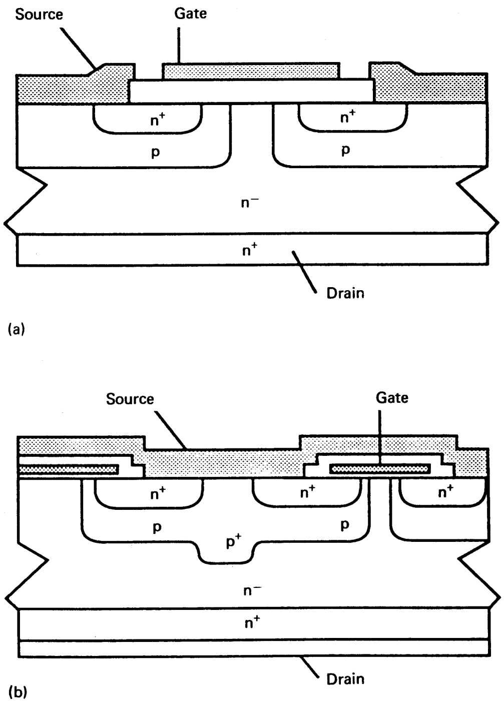


Figure 17.47 (a) Double-diffused MOS metal gate FET; (b) Double-diffused MOS silicon gate FET

the drain. Another common feature for all these n-channel devices is that the n-region largely supports the applied drain potential because its doping level is much lower than that of the p body region. However, it can be seen that the position of the p channel in the VMOSFET and the UMOSFET differs from its position in the DMOSFET. Although the DMOSFET silicon gate structure needs a more sophisticated technology very similar to that needed to produce CMOS integrated circuits, it is generally possible to produce a far superior device using this process and, therefore, the other three power MOSFET structures are less widely utilised. This is because it can be manufactured using a self-aligned process that makes it far easier to produce. In this process the MOS channel regions are obtained by a difference in lateral diffusion of the two impurity distributions and the use of this double-diffusion technique achieves very short channel lengths of less than 1.5 μm. This increases the potential source/gate packing density that directly reduces the cost and improves the performance of the device. At the same time the use of a polycrystalline gate also reduces the possibility of sodium ion contamination in the gate oxide, and results in a higher stable threshold voltage $V_{GS(th)}$.

However the UMOSFET structure has recently been given increased attention as fabrication techniques have improved to a degree that it could be practically realised. This type of MOSFET device is also referred to as a Trench MOSFET, and correspondingly the DMOSFET is called a planar MOSFET. The advantage of the Trench MOSFET is the capability to achieve far greater cell packing densities, but at the expense of increased manufacturing difficulty.

As mentioned previously, the n-MOSFET structure is switched on by applying a voltage between the drain and the source and positively biasing the gate with respect to the source. This bias creates an electric field in the channel region that reverses the polarity of the charge carriers in the body region to create a majority carrier path from the source to the drain. Electron current flows from the source metal to the source contact, laterally through the channel and then vertically through the n^- and n^+ drain regions to the drain contact as shown in *Figure 17.44*. This structure also contains an internal parasitic diode, formed by the source short-p body contact, and a bipolar n–p–n transistor as shown in *Figure 17.48*.

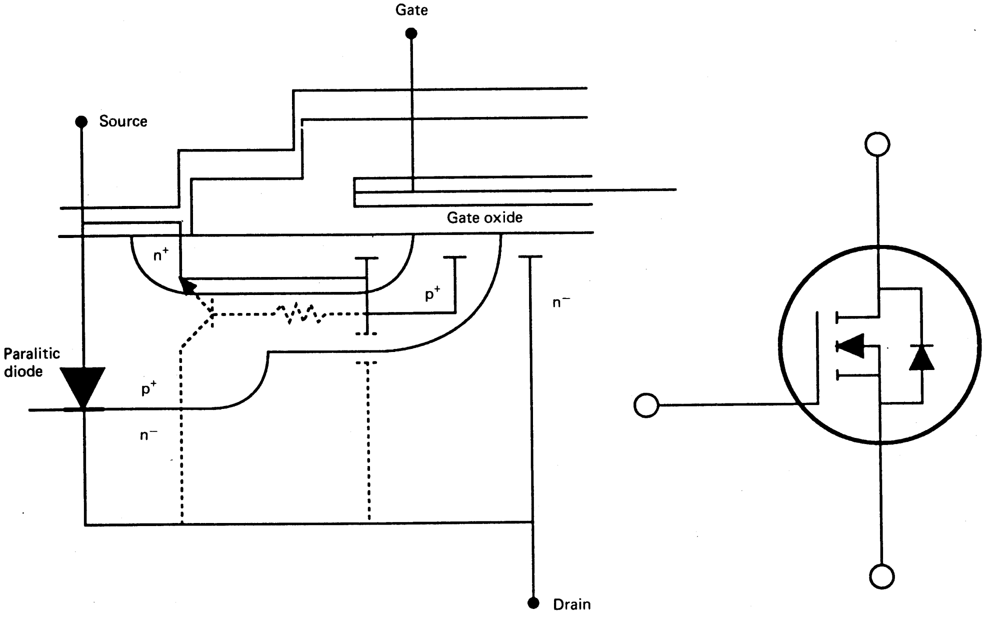


Figure 17.48 Schematic representation of a power MOS structure

The diode conducts when the drain is negatively biased with respect to the source and usually can conduct forward current at the same level as the MOSFET part of the device. Under these conditions charge is injected into the n-region and the parasitic diode behaves almost like the epitaxial fast-recovery p–n junction power. It does, however, suffer from the disadvantage that it cannot be made a very fast-recovery type, as normally required for high frequency MOSFET applications, without affecting the other characteristics of the MOSFET. Therefore, although it can be used in certain lower frequency applications as an internal inverse parallel diode, this is often avoided by inserting an external reverse blocking diode in series with the MOSFET and providing another inverse parallel to carry reverse current. Where the internal diode is used an allowance must be made for the additional power losses dissipated in the MOSFET structure and the parasitic diode's reverse recovery characteristics resulting in a reduced power rating and/or a lower permissible switching frequency. The advantage in using the internal diode is of course the elimination of the two external components. Where it is essential to use the external diodes they should be of the very fast-recovery type such as a Schottky barrier diode, or a fast switching p–n diode, to keep switching losses in the MOSFET to a minimum.

If the current flow in the diffused base resistance in the $p^{+\Leftarrow}$ region of the parasitic transistor shown in *Figure 17.48* (which may be very high under MOSFET current conditions) is sufficiently large, the p^{+}–$n^{+\Leftarrow}$emitter junction of the parasitic transistor will become forward biased. This injects current that will be conducted by the parasitic n–p–n transistor. Usually only a few of the MOSFET cells reach this condition first wherein a very high current is focused that destroys the device. This is one of the basic failure mechanisms of the MOSFET that is avoided by preventing the load current of the MOSFET exceeding a critical value. The critical value of current also decreases with increasing temperature but is usually well beyond the manufacturer's rating specification. However it is a mechanism which should be borne in mind together with the fact that a MOSFET does not have the capability of passing an unlimited amount of current for a very short time.

To minimise on-state power losses in a MOSFET the device on-resistance ($R_{\mathrm{DS(on)}}$) needs to be the minimum compatible with desired breakdown voltages. To achieve the maximum possible ($R_{\mathrm{DS(on)}}$) it is necessary to optimise the power MOS channel perimeters to achieve the highest packing density per unit area possible. As a result, planar DMOS low-voltage devices have a packing density of 1000–10 000 cells or more per square millimetre This has resulted in the development of advanced 'Trench' MOSFET structures which can reach even higher cell densities up to 20 000 cells per square millimeter. For high-voltage devices the epitaxial layer resistance has a greater effect on the overall resistance and to optimise ($R_{\mathrm{DS(on)}}$) it is necessary to minimise the bulk resistance by choosing a lower packing density which increases the area of the epitaxial drift region. High-voltage devices therefore often have a packing density of about 500 cells per square millimetre. Many novel techniques have been considered for reducing the epitaxial drift region resistance, with such proprietary names as '*CoolMOS*'.

The maximum breakdown voltage rating of the DMOSFET (BV_{DSS})would be limited by junction edge breakdown effects below the bulk avalanche breakdown due to the curvature of the diffused junctions and surface electrical field crowding, unless an edge structure such as that shown in *Figure 17.49* is used. The field plate allows higher breakdown voltages by spreading the field laterally along the surface of the device.

The gate threshold voltage ($V_{\mathrm{GS(th)}}$) is another important parameter. This is the voltage at which strong inversion begins to occur in the MOS gate and sets the minimum applied gate voltage for conduction in the channel. The threshold voltage is related to the thickness of the gate oxide and to N_{A}, the maximum peak impurity concentration in the region between the source and the drain. Channel punch-through can occur as the result of insufficient impurity charge in the channel, under strong reverse bias and to avoid punch-through a trade off between $V_{\mathrm{GS(th)}}$ and length must be made. Due to the negative temperature coefficient of $V_{\mathrm{GS(th)}}$ its value cannot be allowed to become too low otherwise the gate becomes too sensitive at high temperature. Also, if it becomes too high, the MOSFET cannot be directly driven by low-voltage logic circuits or requires too much gate power and, therefore, the value of $V_{\mathrm{GS(th)}}$ is normally designed to be in the range 2–4 V for power devices, although lower values of 1–2 V are found for logic-compatible devices

Power MOSFETs require a certain amount of gate charge driven into their gates or extracted during the charging and the discharge phases of the input capacitance and behave quite differently from bipolar devices. In particular the switching performance of the device is influenced by the time it takes for the voltage to change across the device capacitance. *Figure 17.50* show a power MOSFET driven by a gate source with an internal resistance R_{I} and an open-circuit voltage V_1 where R_{L} is the load in the main switching circuit.

The gate input capacitance $C_{\mathrm{Gi}} = \text{€}_{\mathrm{GS}} + C_{\mathrm{GD}}$ is non-linear during the switching cycle. The capacitance between the gate electrode and the drain C_{GD} is a function of the depleted drain layer thickness which in turn is dependent on V_{DS} (drain source voltage). As V_{DS} increases the depleted layer thickness increases and the static input capacitance C_{GD} decreases.

There is also a displacement charge which has to be supplied when $\mathrm{V_{DS}}$ decreases from V_{DD} to $V_{\mathrm{DS(on)}}$ and C_{GD} behaves as a higher equivalent capacitance than during the drain 'on' transitions. C_{GD} is called the Miller Capacitance and it causes the total input capacitance to exceed the sum of the static capacitance during the dynamic turn on phase.

During the turn on period the voltage across the gate increases until it reaches the threshold voltage ($V_{\mathrm{GS(th)}}$), at this point drain current begins to flow and C_{GS} becomes charged. When the gate source capacitance is fully charged, the drain voltage begins to fall while the drain current is almost constant charging the Miller Capacitance C_{GD}. During this phase the $\mathrm{V_{GS}}$ is also constant, as it cannot increase until the input capacitance is fully charged. This is the so-called Miller effect. Following this V_{GS} increases again until it reaches the gate supply voltage V_{i}.

The minimum charge that must be delivered by the gate drive to ensure turn-on is therefore given by the charge required to fully charge the input capacitance, however in practice gate drive design will require the use of greater capacity so that the gate voltage reaches the gate supply voltage, i.e. time t_4 in *Figure 17.51*.

At turn-off, the behaviour of the input capacitance is exactly opposite, and the phenomena described above occur in reverse order.

Power MOSFETs are available with current ratings up to 200 A and voltage ratings to 1000 V and are capable of switching at frequencies up to 10 MHz. The high switching speed, wide SOA, high peak current capability and ease of

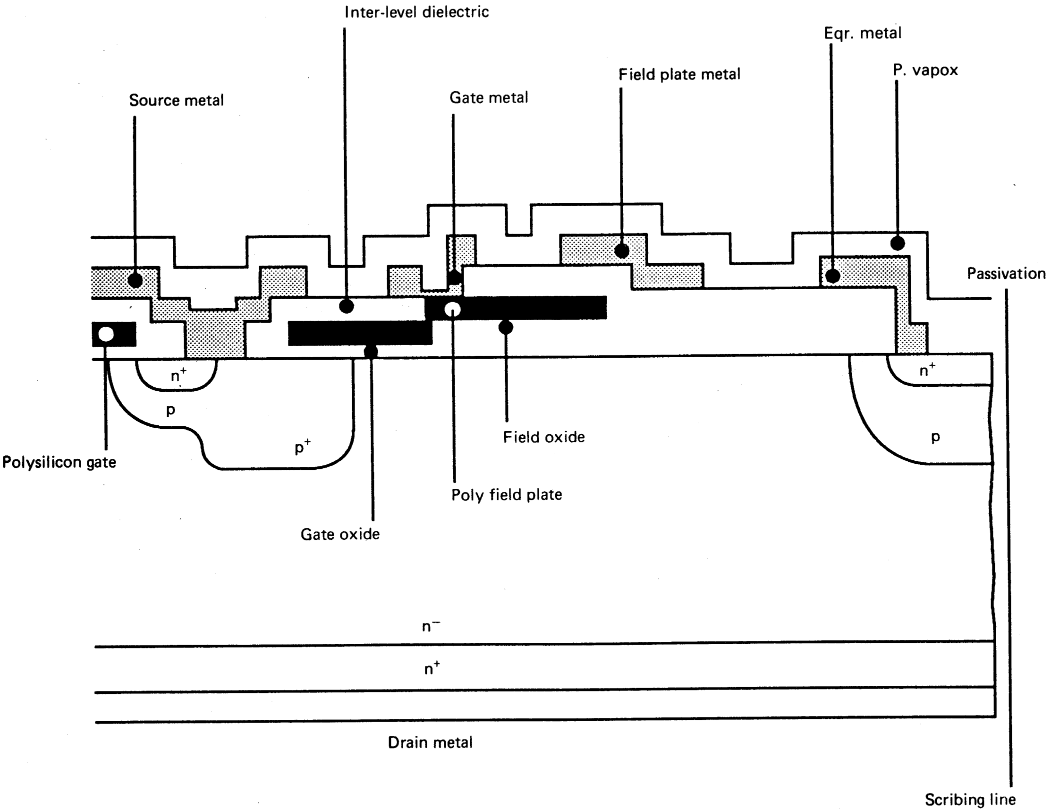

Inter-level dielectric
Eqr. metal
P. vapox
Field plate metal
Source metal
Gate metal
Passivation
n+
p
p+
n+
p
Polysilicon gate
Field oxide
Poly field plate
Gate oxide
n−
n+
Drain metal
Scribing line

Figure 17.49 Power MOS edge structure

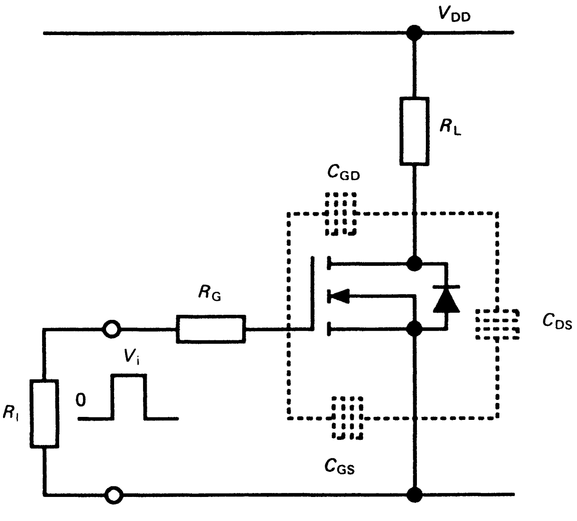

V_DD
R_L
C_GD
R_G
C_DS
V_i
0
R_i
C_GS

Figure 17.50 Power MOS equivalent circuit

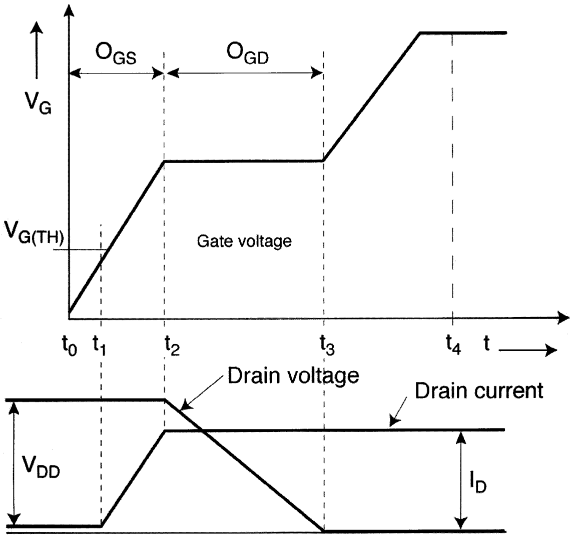


Figure 17.51 MOSFET turn-on characteristics

control has made the power MOSFET the favoured device for the majority of power electronic applications. Unfortunately it is the unipolar operation of the device, that delivers the above benefits that also leads to highly temperature and blocking voltage dependent conduction characteristics, and therefore restrict the useful application area to the lower power regime. A more complete description of the MOSFET is given in reference 2.

17.6 The insulated gate bipolar transistor (IGBT)

In searching for the ideal power device, designers have considered many potential design concepts that combine the voltage control of the MOSFET gate with the superior conduction characteristics of the bipolar device. Of the many candidates there is only one that has become widely commercially available, the insulated gate bipolar transistor (IGBT). There is much research and development work in hand to implement improved versions of the IGBT, for example incorporating Trench Gates and making use of accumulation layer emitter effects, however this falls outside the scope of this chapter where we will focus on the commercially available DMOS IGBT. The IGBT employs injected-charge modulation in the base region and, due to the need for this charge to be extracted or extinguished at turn-off, has higher switching losses compared to the MOSFET. But as it can be realised at much higher power ratings it has become the power device of choice for a wide range of medium to high power electronic applications. The main advantages of the IGBT are the simplicity with which it can be driven (which is comparable to a power MOSFET), its lower on-state conduction losses and the capability of switching high voltages. These characteristics, together with the ability of IGBT to survive a wide reverse bias safe operating area (RBSOA) make it superior to the power MOSFET in high-voltage applications. Typically, IGBTs are used for switching circuits requiring high voltage (up to 3300 V) and high current (up to 3000 A), with a switching frequency of the order of 1–40 kHz.

17.6.1 Device physics

Figure 17.52(*a*) shows a cross section through a single cell of the multi-cell structure of an IGBT and *Figure 17.52*(*b*) shows the power MOSFET for comparison. The drain region of the IGBT which contains a p$^+$-doped layer represents the major difference in structure to that of power MOSFETS. The presence of the p$^+$–n junction drastically reduces the on resistance of the IGBT during conduction because holes are injected from the p region, which results in the n$^-$-region becoming conductivity modulated. In a power MOSFET significant proportions of the conduction losses occur in the n$^-$-region, for devices above 500 V this can amount to more than 70% of the total.

As is also shown in *Figure 17.52* the Source of the MOSFET becomes the Emitter of the IGBT, and the MOSFET drain has become the Collector. This change in terminology correctly demonstrates that the IGBT is a bipolar device and that the effective 'Drain' of the MOSFET section is not directed connected to the external circuit.

A circuit model of the IGBT, shown in *Figure 17.53*(*a*), illustrates the internal arrangement. The MOS section provides a gate drive to the base of a p–n–p transistor (formed by the p$^+$–n$^+$–n$^-$–p$^+$-layers in *Figure 17.52*(*a*)) where in this p–n–p transistor its emitter is directly connected to the IGBT collector and the p–n–p collector feeds into the base of a second n–p–n transistor (formed by the n$^-$–p$^+$–n$^+$ layers). Thus it can be seen that the IGBT resembles a four layer n–p–n–p thyristor structure, see the two transistor analogue of the thyristor in *Figure 17.24*. In this thyristor analogy the gate drive to the thyristor derives from the MOSFET gate. As with the thyristor, should the sum of the n–p–n and p–n–p gains ($\alpha_{\mathrm{npn}} + \alpha_{\mathrm{pnp}}$) exceed unity then the IGBT will latch on, as a thyristor, with consequential loss of gate control and potentially destruction of the device. In order to prevent latch up IGBTs are designed to achieve control over the gains α_{npn} and α_{pnp}. This gain control is usually achieved by three approaches. Firstly the resistance R_{B} is minimised (this is analogous to efficient cathode emitter shorting in the thyristor, see Section 17.3.1 and *Figure 17.30*). Secondly the injection efficiency of the p$^+$-layer (at the collector) is reduced by various techniques, for example the structure in *Figure 17.52*(*a*) uses an n$^+$-layer. Finally, the p–n–p gain may be reduced by minority lifetime control.

With the above mentioned techniques to de-sensitize the 'thyristor' action of the IGBT, the NPN transistor may be ignored and the equivalent model is simplified to that shown in *Figure 17.53*(*b*). In both models R_{mod} represents the effective resistance of the n$^-$-region in the 'drain' region of the device (i.e. the MOSFET region drain). Referring to *Figure 17.53*(*a*) or (*b*) the drain region of the IGBT is the region in the centre below the gate, in this region the n-base region is surrounded by p-base regions. In the forward biased condition, and when the IGBT is conducting the

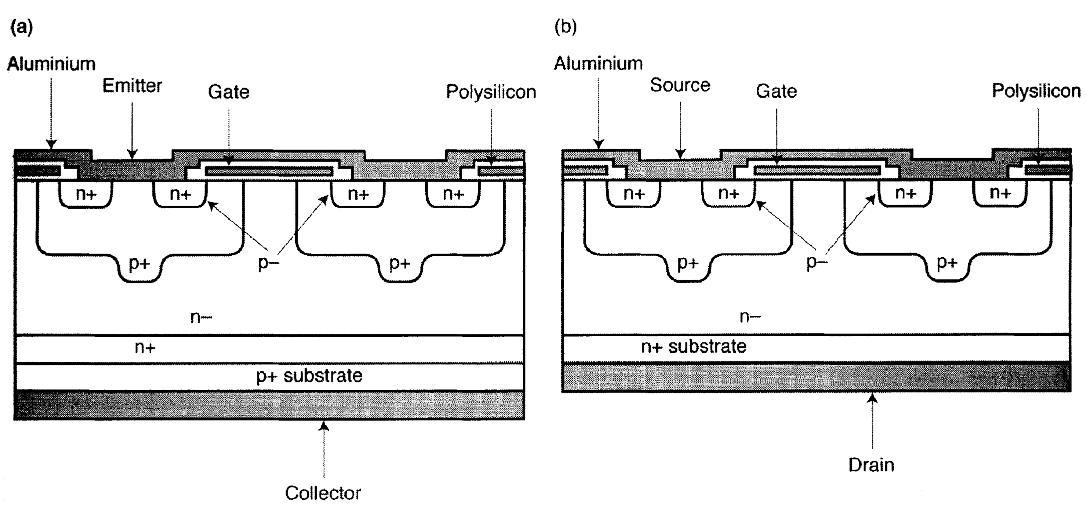
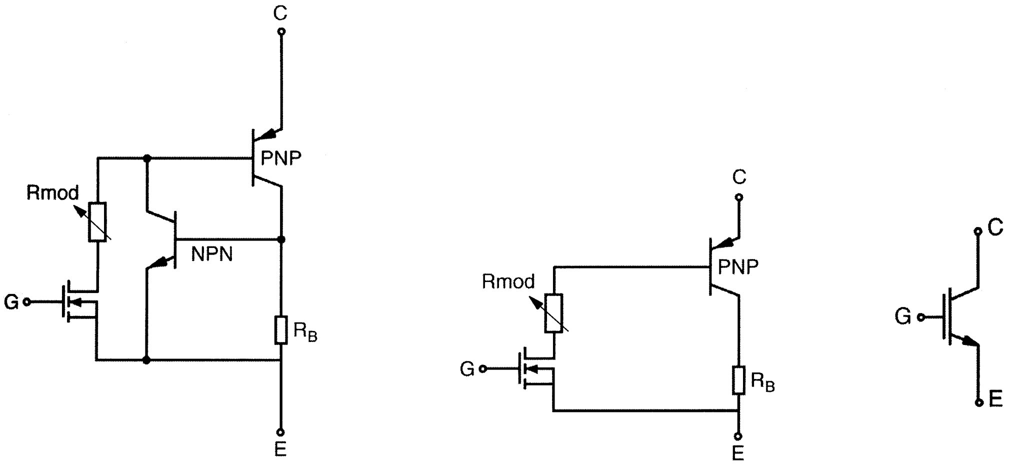


Figure 17.52 (a) IGBT structure; (b) Vertical DMOSFET structure



(a) Equivalent circuit (b) Simplified circuit (c)

Figure 17.53 Equivalent circuits for IGBTs

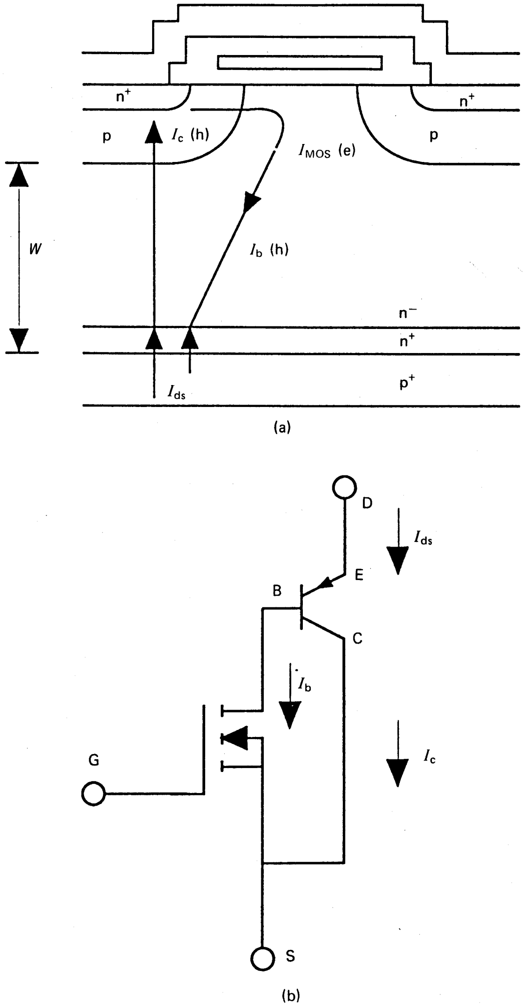


Figure 17.54 Current flow in an IGBT during conduction

p–n junction between the p- and n-base is reverse biased. Thus this reverse biasing effect may increase the effective resistance R_{mod}. As a result many IGBT designs take particular care to maximise the conductance in this region of the device by increasing the n-type doping in this region for example. In some equivalent circuit models this is shown as a JFET (a Junction Field Effect Transistor) because the p^{+}-base regions surrounding this n^{-}-drain act as the junction of a JFET structure. For an explanation of the JFET see reference 5.

There are two major types of IGBT manufactured, either Punch-Through (PT) or Non-Punch-Through (NPT). The IGBT structure shown in *Figure 17.52(a)* is a PT IGBT and this has the similar structure to the Power MOSFET with the addition of the p^{+}-layer (Collector) to the drain side of the MOSFET. In this structure the off-state voltage is supported by the formation of a space charge region across the wide n-region, but this space charge region punches through to the n^{+}-buffer region before the junction breaks down. In this respect it is similar to the p–i–n junction (Section 17.1.5) in that it achieves a thinner device structure, and therefore minimum on-state voltage and improved dynamic characteristics, for the same blocking voltage.

In an NPT IGBT structure the n^{+}-buffer layer is omitted, and so the thickness and the doping levels in the n^{-}-region are chosen such that the junction breaks down before the space charge region reaches through to the p^{+}-collector region.

The choice between NPT and PT types is based on a number of factors surrounding the manufacturing process cost and yield and the IGBT characteristics. In the earlier IGBT types PT has been restricted to IGBTs at voltages less than 1200 V owing to the economics of growing the very thick epitaxial layers needed for higher voltages, but many manufacturers have developed both PT and NPT types across the full voltage range.

As with the Power MOSFET the IGBT is available in both DMOS and UMOS (or more usually 'Trench Gate') construction. The Trench gate is generally used in the high power IGBT to improve the effective conduction area of the IGBT and reduce the $V_{\text{CE(on)}}$, although it can also be used to increase the cell packing density to improve the switching speed for lower voltage designs. At the time of writing the trench IGBT was just becoming commercially available at

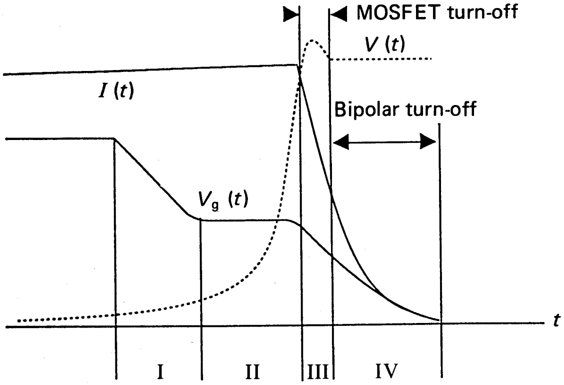


Figure 17.55 Turn-off of an IGBT showing three turn-off regions

1200 V levels with higher voltage designs in development, whereas the DMOS device is available up to 6500 V.

17.6.2 Packaging and thermal considerations

Although IGBT are available in discrete packages, i.e. packages that contain one IGBT die, the majority of IGBTs are supplied in module configurations. A module is a package that contains more than one IGBT die, and may more usually contain a combination of IGBT and fast recovery diode die. An example of a high power IGBT module, with one half of the lid cut away to illustrate the internal construction, is shown in *Figure 17.56*.

Typically the module has a conductive base plate of copper or alternatively a metal matrix material (such as AlSiC), the IGBT and diode die are soldered onto a DCB and one or more DCBs are soldered to the base plate. The DCB (Direct Copper Bonded substrate) consists of an insulating material, such as Aluminium Nitride or Alumina, with a copper layer bonded to both sides, and with the side to which the die are attached patterned to produce the necessary interconnections between the die. The DCBs may be further interconnected within the module and the main terminal electrodes (cathode, anode and gate) are connected to the DCBs. Clearly the detailed design and construction of the module is extremely complex and requires consideration of the electrical, thermal (for high transfer) and mechanical requirements. Of prime importance is the reliability of the package, and manufacturers go to great lengths to optimise the life expectancy of the module.

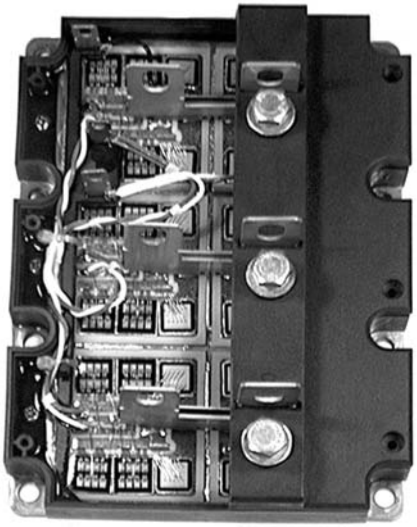


Figure 17.56 IGBT module construction

The IGBT module will contain fast recovery diodes connected across the IGBTs, in anti-parallel. The inclusion of the diodes results in ease of assembly and low parasitic inductance to optimise the switching and operating area of the module. The diodes used in the modules are of the fast-recovery type but are specially designed to have soft-recovery. The use of soft-recovery diodes is essential in order to prevent the snappy recovery of the diode applying excess voltage stress on the IGBT at turn on.

In the assembly of equipment using IGBT modules the user usually connects to the power terminals by parallel busbar rails and the layout of the power terminals facilitates the use of extremely low inductance rails. The base of the module is connected directly to a heaksink surface and because the base plate is electrically isolated from the IGBT die several modules may be mounted side by side on the same heatsink. Thus the IGBT module lends itself to ease of assembly and extremely compact equipment designs.

17.6.3 On state characteristics and conduction losses

Current flow in the IGBT during conduction is shown in *Figure 17.54*. The total current I_{DS} is the sum of the two currents I_c(h) and I_b(h), where I_c(h) is the current due to holes injected from the substrate p^+ layer, which are collected on the surface of the source region, and I_b(h) is the current due to holes that combine with electron current flowing through the MOS channel in the base region of the p–n–p bipolar junction transistor. The IGBT, therefore, behaves like a bipolar junction transistor whose base current is provided by a MOSFET, and the drain current of the IGBT is thus the sum of the bipolar base and collector currents.

$$I_{DS} = I_{MOS}(1 + \beta_{pnp})$$

From this it can be seen why the IGBT has a higher current carrying capability and lower $R_{DS(on)}$ than a power MOSFET of equal chip size. However, the presence of the p^+–n junction in the drain produces an initial offset in the output characteristics of 0.6–0.8 V, and also limits the reverse breakdown voltage of the IGBT to 30–70 V. However for most applications (e.g. voltage source inverters) the lack of reverse blocking capability is not a problem.

17.6.4 Turn on and turn off

When the gate-source voltage (V_{GS}) is greater than the threshold voltage (V_{TH}), the MOSFET section of the structure turns on. The MOS drain current is then the base current of the parasitic p–n–p transistor and it turns on the p–n–p transistor in less than 100 ns. The IGBT turn-on time is therefore, a function of the impedance of the gate driver circuit and the voltage applied to the gate. During this turn-on time the falling voltage and rising current can result in high instantaneous power levels, leading to a total energy loss during this phase, called the turn-on switching loss.

Turn-off in the IGBT has typical features of both a power MOSFET and a bipolar junction transistor turn-off due to the use of conductivity modulation for increased current density. *Figure 17.55* shows how the turn-off can be divided into four regions, during the first phase (I) the gate voltage decreases to a point where the Miller effect begins and V_{DS} begins to rise. In the second phase (II) the gate voltage is constant due to the Miller effect. During this period the gate capacitance decreases as V_{DS} increases. The gate voltage polarity is reversed with respect to the drain voltage, as V_{DS} rises above the gate potential and V_{DS} rises to its maximum value at a rate controlled by the driver circuit. These two phases (I and II) are dependent on the MOSFET behaviour as the base-collector junction of the p–n–p transistor becomes reverse biased. Regions III and IV define t_{fall}. In region III the turn-off process is MOS controlled which is very fast. In region IV the MOS channel is closed and the p–n–p transistor has an open base. During this period the current falls at a rate determined by recombination of excess carriers in the n^--region. The fall time can therefore be controlled by the gate drive-circuit, in region III, but in region IV is only dependent on the p–n–p transistor lifetime and gain. *Figure 17.57(a)* shows how t_{fall} varies with V_{DS}. It can be seen that, as V_{DS} increases, so the gain of the p–n–p transistor increases due to the reduction in the base thickness. This is caused by the depletion region increasing with increasing V_{DS}. The dependence of t_{fall} on temperature is shown in *Figure 17.57(b)* and confirms the correlation between t_{fall} and the p–n–p transistor gain. The lifetime of the minority carriers in the n^--region versus t_{fall} is shown in *Figure 17.57(c)*.

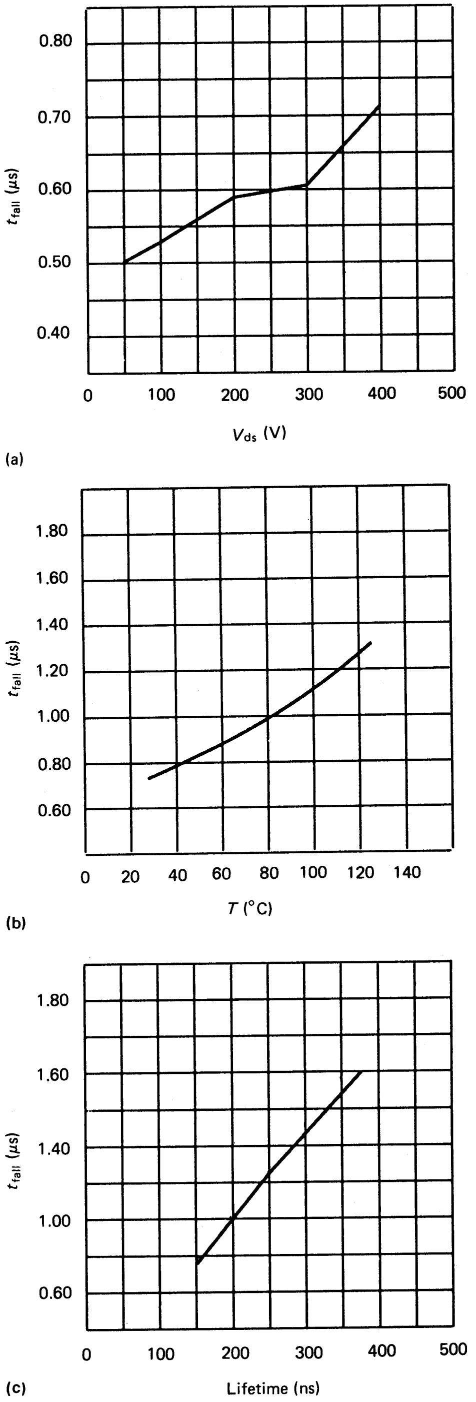


Figure 17.57 (a) Variation in t_{fall} with V_{DS}, (b) variation in t_{fall} with temperature; (c) variation in t_{fall} with the lifetime of the minority carriers in the n^--region

17.6.5 Safe operating area

Protecting the IGBT against over-current or over-voltage effects that can be potentially destructive is a most important part of the design for the majority of applications. These can be specified by the ability of the device to withstand a short circuit load, the short circuit safe operating area (SCSOA) and to survive during switch off, the turn-off switching safe operating area (SOA).

The SOA is illustrated by considering the switching of an IGBT in a half bridge circuit shown in *Figure 17.58*. The switching waveforms are shown in *Figure 17.59*. The SOA curve is derived from the locus of maximum permissible current and voltage during the turn off phase,

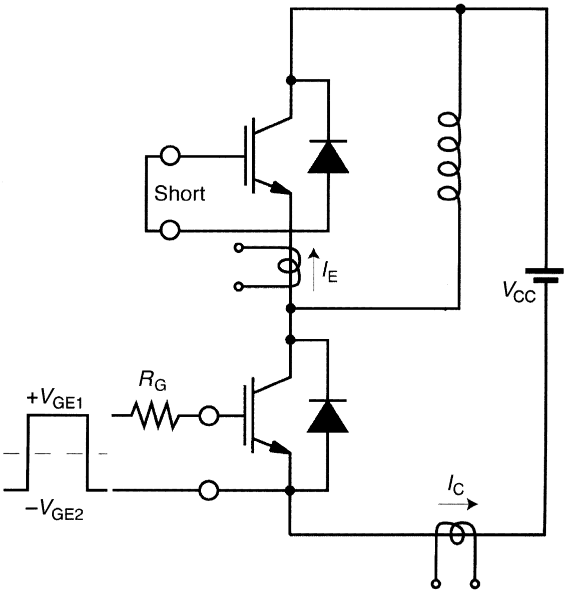
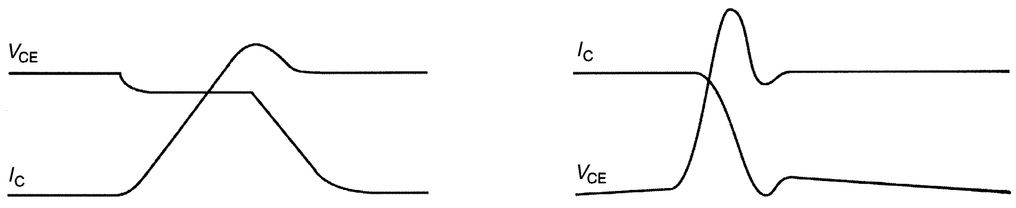


Figure 17.58 IGBT half bridge test circuit



(a) Turn-on

(b) Turn-off

Figure 17.59 IGBT switching waveforms (a) turn-on; (b) turn-off

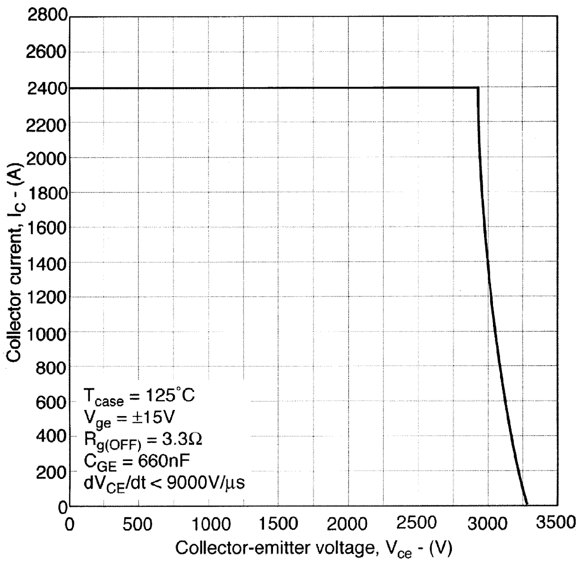
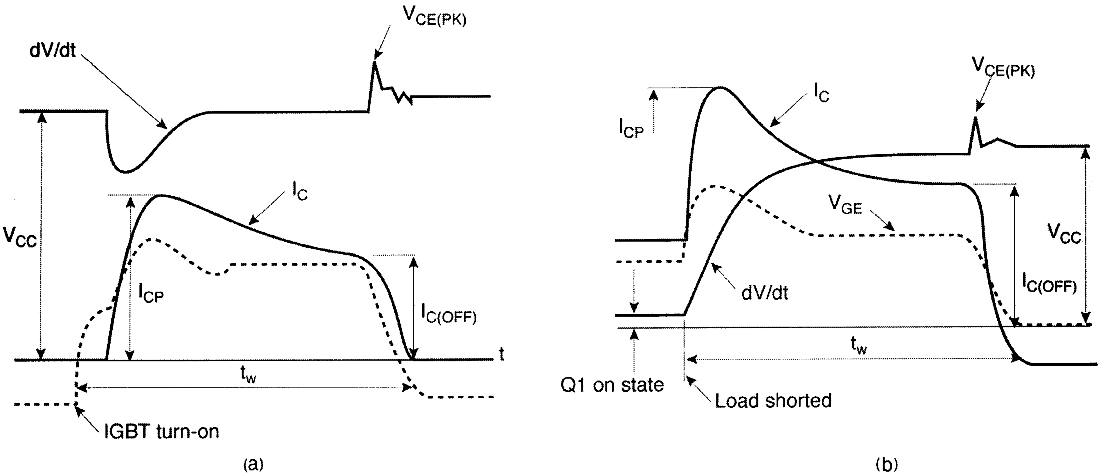


Figure 17.60 Turnoff switching safe operating area for a 3300 V, 1200 A IGBT

for example *Figure 17.60*. The failure mechanism is typically due to the rapid extraction of charge from the bipolar section of the IGBT, this charge may flow laterally in the base regions and can lead to latch up of the equivalent thyristor section of the device.

In many applications it is important that the IGBT can withstand a short circuit event in the system, for example in the motor. The are two situations that may occur. Firstly where the IGBT is turned on into a short circuit *Figure 17.61(a)*. In this case the IGBT is initially off and supporting the supply voltage V_{cc}, the IGBT is turned on into a short circuit where the only load is the stray circuit inductance L_s and the IGBT voltage then drops due to the discharge of the stray inductance and then rises rapidly back to the supply level. This rapid d*v*/d*t* is picked up by the gate due to the internal capacitance of the IGBT, and this rise in gate voltage results in a further increase in the collector current. In this condition the IGBT must be turned off within a defined time of the order of 10 μs, and the peak over-current should be limited or the device will fail.

The second condition for SCSOA is where the device is already in conduction when the load is made short circuit. *Figure 17.61(b)*. The short circuit forces the IGBT into de-saturation, the collector voltage rises rapidly towards the supply voltage V_{cc}, as in the previous example this d*v*/d*t* forward biases the gate strongly leading to higher values of collector current. Again the peak



(a)

(b)

Figure 17.61 IGBT switching waveforms under short circuit conditions: (a) IGBT turned on into a short circuit; (b) load short circuited while IGBT is conducting

collector current and the duration of the short circuit must be limited to avoid device failure. Most manufacturers will define the conditions to limit SCSOA in their data sheets.

IGBTs are available covering a broader spectrum of power than almost any other power device, with the exception of the thyristor. IGBTs are available from a few amps and 300 V up to 3000 A and 6500 V. These devices are used over a broad frequency range from a few hundred Hz up to and exceeding 100 kHz. At the time of writing no suitable alternative has been released that might challenge the dominance of the IGBT, but there is no doubt that as this device technology becomes mature the device designers will turn their attention beyond the IGBT. For further reading a more complete description of IGBTs is given in reference 2.

Acknowledgements

Much of this text is derived from the first edition prepared by Dr Roger Bassett when he was Marketing Manager for what was to become the power business of Dynex Semiconductor Ltd in Lincoln England, but was at the time owned by GEC Plessey Semiconductors. The recent revisions and additional material are included with the kind permission of Dynex Semiconductor Ltd.

References

1 GHANDI, S. K., *Semiconductor Power Devices,* Wiley, New York (1977)
2 WILLIAMS, B. W., *Power Electronics, Devices, Drivers and Applications*, MacMillans Education, New York (1987)
3 TAYLOR, P. D., *Thyristor Design and Realisation*, Wiley, New York (1987)
4 BENDA, V., GOWAR, J. and GRANT, D., *Power Semiconductor Devices: Theory and Application*, Wiley, New York (1999)
5 BALIGA, B. J., *Modern Power Devices*, Wiley, New York (1987)

18

Electronic Power Conversion

T C Green PhD, MIEE, CEng
Imperial College of Science, Technology and Medicine

Contents

18.1 Electronic power conversion principles 18/3
  18.1.1 Switch-mode electronics 18/3
  18.1.2 Power loss in switch-mode circuits 18/4

18.2 Switch-mode power supplies 18/5
  18.2.1 Buck SMPS 18/5
  18.2.2 Boost SMPS 18/10
  18.2.3 Flyback SMPS 18/11
  18.2.4 Capacitor coupled SMPS 18/11
  18.2.5 Isolated flyback SMPS 18/14
  18.2.6 Transformer isolated Buck SMPS 18/17
  18.2.7 SMPS control 18/18

18.3 D.c./a.c. conversion 18/20
  18.3.1 Single phase bridge 18/20
  18.3.2 Three phase bridge 18/22
  18.3.3 Current source inverters 18/25

18.4 A.c./d.c. conversion 18/26
  18.4.1 Line-frequency-switched rectifiers 18/26
  18.4.2 Wave-shape controlled rectifiers 18/29

18.5 A.c./a.c. conversion 18/34
  18.5.1 A.c. voltage regulator 18/34
  18.5.2 Direct frequency converter 18/35
  18.5.3 Indirect frequency converter 18/35

18.6 Resonant techniques 18/37
  18.6.1 Quasi-resonant SMPS 18/37
  18.6.2 Resonant SMPS 18/39

18.7 Modular systems 18/39
  18.7.1 Interleaved SMPS 18/41
  18.7.2 Multi-pulse rectifiers 18/42
  18.7.3 Multi-level inverters 18/42

18.8 Further reading 18/43

Electronic processing of power provides the freedom to optimise the generation, conversion and use of electrical power. Most notably, it frees a system designer from the constraints of a fixed voltage supply (whether a 50/60 Hz a.c. public electricity supply or a d.c. source such as a battery). For example, wind turbines are more effective at capturing energy from the wind if freed from the constraint of rotating at synchronous speed, and pump flow rate is much more efficiently regulated through variable voltage/frequency control of the motor than through valves. But to realise the efficiency savings from varying the voltage or frequency of a power source we must ensure that the energy conversion process is itself efficient. Power electronics provides efficient means of energy conversion and has long since displaced electromechanical means.

Efficiency has been important in promoting the widespread adoption of power electronics, but the speed and degree of control that can be exercised over the power conversion process is at least as important. This control is essential in conditioning the power for an electrical machine in a machine tool or robotic manipulator. The speed of response that we can achieve from an electronic power converter depends on its implementation and, in turn, its power rating. In general, we can expect a response much faster than any electromechanical element of the system.

18.1 Electronic power conversion principles

The terms *power electronics* and *switch-mode electronics* are almost synonymous for the simple reason that semiconductors are energy efficient only when used as switches.

18.1.1 Switch-mode electronics

Voltage drop across a current carrying element is the indication of inefficiency. A transistor used in its active region (also know as the linear or amplifier region) dissipates power because it has voltage across it while carrying current. *Figure 18.1* illustrates the active region of three common devices on graphs of their output characteristics (i.e. graphs of the current through the main terminals and the voltage across those terminals for a range of controlling voltages). The three devices, the Bipolar Junction Transistor (BJT), the Metal-Oxide-Semiconductor Field Effect Transistor (MOSFET) and the Insulated-Gate Bipolar Transistor (IGBT) were discussed in Chapter 17.

The effect this has on system efficiency is well illustrated by two alternative ways of regulating the voltage delivered to a load. *Figure 18.2* shows a transistor used as a linear voltage regulator to reduce an input voltage, V_I down to a controlled output voltage, V_O.

As an example, consider a 5 Ω load and a regulator set to produce a 5 V output from 15 V input. The load will draw a current, I_O of 1 A. Because the current drawn from the source of the transistor is matched by current drawn in through the drain, the input current I_L will be 1 A (assuming that the control circuitry draws negligible current). The power flow will be as follows:

$$P_\mathrm{out} = V_\mathrm{O} I_\mathrm{O} = 5\ \mathrm{W}$$

$$P_\mathrm{in} = V_\mathrm{I} I_\mathrm{I} \approx V_\mathrm{I} I_\mathrm{O} = 15\ \mathrm{W}$$

$$P_\mathrm{lost} = V_\mathrm{T} I_\mathrm{O} = 10\ \mathrm{W}$$

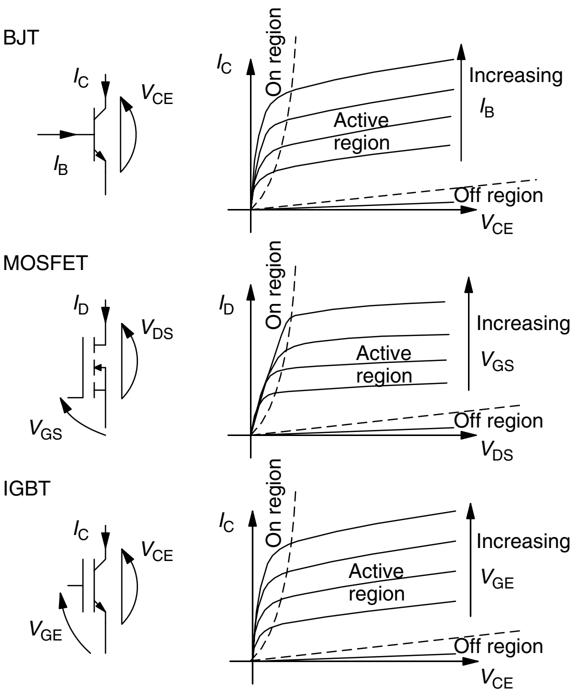


Figure 18.1 BJT, MOSFET and IGBT output characteristics and operating regions

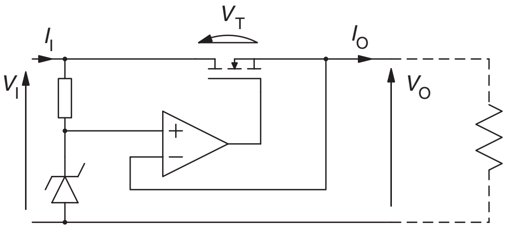


Figure 18.2 A MOSFET based linear voltage regulator

This circuit shows an efficiency of only 33%. Efficiency is important in its own right because of the need to utilise primary sources of energy well, but there is a second important reason to aim for high efficiency. The energy lost in this circuit is lost as heat in the semiconductor material. Thermal management of the semiconductor is difficult because it can be difficult to remove heat from the silicon and so its temperature will rise far above ambient. Further, the semiconductor is a small mass and its temperature will rise quickly during short periods of high dissipation. Typical maximum temperatures of silicon devices are in the range 125°–150 °C.

The alternative way of regulating the voltage across a load using a transistor is to switch the input voltage into a series of pulses. The average voltage of the pulse train is necessarily less than the input voltage. By filtering the pulse train, the ripple component of the voltage can be removed to leave only the average, that is, d.c. component to be applied to the load. *Figure 18.3* shows a transistor used to apply a pulse train to an LC filter.

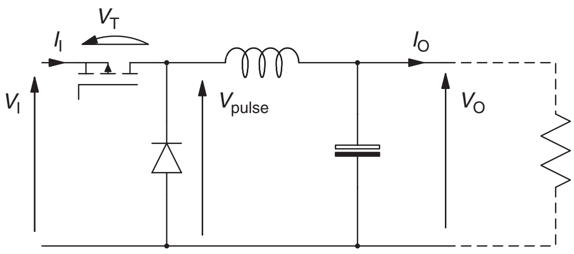


Figure 18.3 A switch-mode voltage regulator

The ratio of the on-time to the off-time can be adjusted to control the voltage delivered to the load. The output voltage will be equal to the average value of the pulse train.

$$V_O = V_{\text{pulse}}^{\text{avg}} = V_I \frac{t_{\text{on}}}{t_{\text{on}} + t_{\text{off}}} = V_I \delta \tag{18.1}$$

The ratio of the on-time to the period, δ, is known as the duty-cycle. If we wish to produce a 5 V output from a 15 V input, as we did with the linear regulator example, we would choose a duty-cycle of $\frac{1}{3}$.

Because the switch is either off (supporting voltage but not carrying current) or on (carrying current but without a voltage appearing across it) there is no coincidence of current flow and voltage drop. Therefore, there is no power dissipated in the transistor. We can also show this by considering the input and output powers. First we note that the inductor current, I_L will have the same magnitude as the output current (with no net current in the capacitor). When the switch is on, the inductor current flows as input current, I_I. When the transistor is off, the inductor current flows in the diode and no current is drawn from the input.

On state:

$$P_{\text{out}}^{\text{on}} = V_O \cdot I_O = 5\,\text{W}$$
$$P_{\text{in}}^{\text{on}} = V_I \cdot I_I = V_I \cdot I_O = 15\,\text{W}$$

Off state:

$$P_{\text{out}}^{\text{off}} = V_O \cdot I_O = 5\,\text{W}$$
$$P_{\text{in}}^{\text{off}} = V_I \cdot I_I = 0$$

On average:

$$P_{\text{out}} = V_O \cdot I_O = 5\,\text{W}$$
$$P_{\text{in}} = \frac{P_{\text{in}}^{\text{on}} \cdot t_{\text{on}} + P_{\text{in}}^{\text{off}} \cdot t_{\text{off}}}{t_{\text{on}} + t_{\text{off}}} = \frac{P_{\text{in}}^{\text{on}} \cdot t_{\text{on}}}{t_{\text{on}} + t_{\text{off}}} = P_{\text{in}}^{\text{on}} \cdot \delta = 5\,\text{W}$$

The average input power matches the output power and therefore the circuit is ideally efficient.

18.1.2 Power loss in switch-mode circuits

In principle, a switch-mode circuit can be perfectly efficient whereas a linear-mode circuit cannot. In practice, of course, there are power losses in switch-mode circuits including power loss within the semiconductors. During the off-state, a semiconductor will leak current, and therefore there will be some power dissipation. The leakage properties of semiconductors are very good and the off-state dissipation is negligible. During the on-state, a semiconductor will drop some voltage and there will be further power dissipation. The on-state power loss is normally significant. There is also power loss as the semiconductor passes through its active region as it is switched between its on- and off-states. The average power loss is found by dividing the energy loss per cycle by the period, T. The energy loss can be expressed as an integration of the product of the instantaneous voltage and current. We can partition the cycle of switch operation into four stages: the on-state, the off-state, the turn-on transition and the turn-off transition.

$$\begin{aligned} P_{\text{loss}} &= \frac{1}{T}\int^{T} v_T(t) \cdot i_T(t) \cdot \mathrm{d}t \\ &= \frac{1}{T}\left\{ \begin{aligned} &\int^{t_{\text{on}}} v_{\text{on}}(t) \cdot i_T(t) \cdot \mathrm{d}t + \int^{t_{\text{off}}} v_T(t) \cdot i_{\text{off}}(t) \cdot \mathrm{d}t \\ &+ \int^{t_{\text{turnon}}} v_T(t) \cdot i_T(t) \cdot \mathrm{d}t + \int^{t_{\text{turnoff}}} v_T(t) \cdot i_T(t) \cdot \mathrm{d}t \end{aligned} \right\} \\ &= \frac{1}{T}\{E_{\text{cond}} + E_{\text{leak}} + E_{\text{turn-on}} + E_{\text{turn-off}}\} \end{aligned} \tag{18.2}$$

The on-state (conduction) energy loss is relatively straightforward to calculate. MOSFET devices are resistive in the on-state whereas most other devices have a near constant on-state voltage. Assuming a constant on-state voltage:

$$E_{\text{cond}} = V_{\text{on}} \cdot I_{\text{on}}^{\text{avg}} \cdot t_{\text{on}} \tag{18.3}$$

Calculation of switching power loss requires detailed knowledge of the trajectory taken through the active region. This is dependent on the gate drive arrangement and the circuit topology.[6] *Figure 18.4* illustrates a simple case where the load is inductive (and treated as a constant current element, I_L) and is switched from a supply voltage of V_S. Diode recovery has been ignored and the transistor has been assumed to operate as a controlled current source while in its active region. The switching trajectory of MOSFETs and IGBTs and the reverse-recovery characteristics of diodes were discussed further in Chapter 17 (Sections 17.5, 15.6 and 17.1.2).

(I) During the off-state, the load current flows in a loop involving the diode, and the transistor must support the full supply voltage (plus diode forward voltage drop).

(II) At turn-on, the transistor begins to take current from the load and this current is assumed to rise linearly. The portion of load current not taken by the transistor continues to flow in the diode. Therefore, the diode is held in forward bias and the full supply voltage still appears across the transistor. Once the transistor current has risen to equal the load current, the diode ceases to conduct. (It is at this point that reverse-recovery will occur in a real diode as discussed in Section 17.2.1). The transistor voltage falls as the parasitic capacitances (of the diode and transistor) charge/discharge and the diode becomes reverse biased. The figure is drawn assuming the change of voltage is also linear.

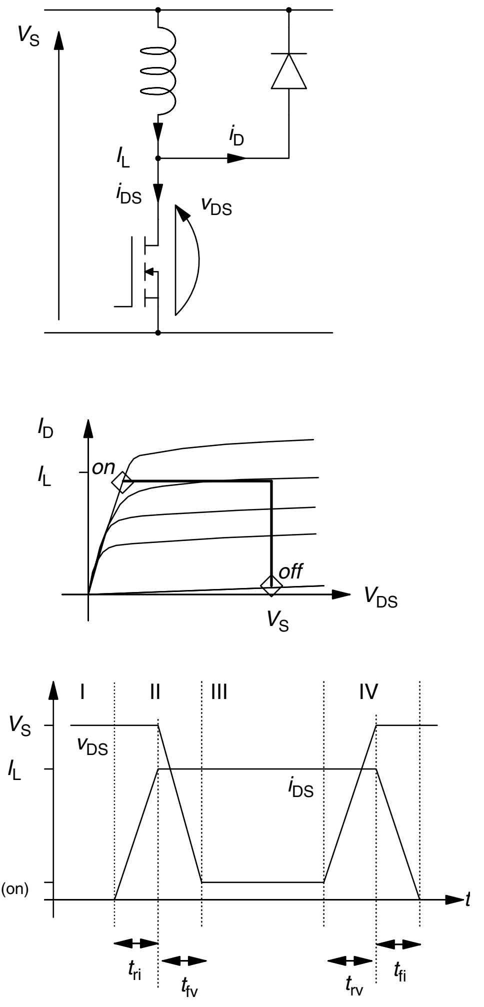


Figure 18.4 A MOSFET switching an inductive load, its trajectory through output characteristic and its drain current and drain-source voltage waveforms

(III) During the on-state, the load current flows in the transistor.

(IV) When turn-off of the transistor is initiated, the decrease of current flow through the transistor diverts current into the parasitic capacitances of the transistor and diode. The voltage across the transistor rises as a result of this. The voltage rises until it exceeds the supply voltage and the diode is forced into conduction. Once the diode is conducting, the transistor current can fall and the load current diverts into the diode.

With the assumptions of linear rise and fall of both voltage and current, the turn-on and turn-off energy loss can be found:

$$\begin{aligned} E_{\text{turn-on}} &= \tfrac{1}{2}V_S I_L(t_{ri} + t_{fv}) \\ E_{\text{turn-off}} &= \tfrac{1}{2}V_S I_L(t_{rv} + t_{fi})\Leftarrow \end{aligned} \qquad (18.4)\Leftarrow$$

Some devices, notably MOSFETs, are specified in terms of the rise and fall times during switching (under standard conditions). Others, notably IGBTs, that have more complex switching transients are specified in terms of the energy loss during switching (again, for standard conditions).

It is normal to express the loss in terms of power rather than energy and so we divide the energy loss per cycle by the period of the cycle.

$$\begin{aligned} P_{\text{loss}} &= \frac{1}{T}(E_{\text{cond}} + E_{\text{turn-on}} + E_{\text{turn-off}})\Leftarrow \\ &= \delta\varsigma\; V_{\text{on}} \cdot I_{\text{on(avg)}} + f \cdot (E_{\text{turn-on}} + E_{\text{turn-off}})\Leftarrow \\ &= \textit{Conduction loss} + \textit{Switching loss} \end{aligned} \qquad (18.5)$$

The power loss due to leakage in the off state has been neglected.

It is important to note that because the switching loss is an energy loss per operation, the power loss is proportional to switching frequency, *f*; whereas the conduction loss is simply an average power loss as indicated by the duty cycle, *δ*. The dependence of power loss on frequency acts against the desire to increase switching frequency that would otherwise bring benefits of faster response and the opportunity to use smaller passive components.

Proper thermal design of a power electronic system is essential to its success. There needs to be management of the power loss through operating mode or circuit design. Choice can be exercised over the package style for its heat transfer properties. The heat-sink to which the semiconductors are attached is often a bulky and expensive part of the system. Its type (natural convection, forced-air, liquid-cooled) and rating (temperature rise per unit power loss) must be carefully chosen.[6]

18.2 Switch-mode power supplies

There is a large variety of circuits available for converting power at one d.c. voltage to another. The linear regulator, introduced in *Figure 18.2*, offers only a step-down of voltage. Switch-mode power supplies, SMPS, can offer step-down, step-up and negation of the voltage through the three basic circuit topologies, *viz*., Buck, Boost and Flyback. Three further topologies, *viz*., Čuk, SEPIC and Zeta, add extra passive components in order to improve secondary aspects such as current and voltage ripple. A family of circuits has been derived from the basic family that incorporate mutually coupled inductors to offer galvanic isolation between input and output. Another family of circuits, to be discussed in, Section 18.6 has been developed that incorporate a resonant element so that the switching trajectory of the transistor is modified to reduce power loss and enable higher frequency operation. The sections that follow cover a selection of the many circuits that are described in the literature. A fuller description and a wider selection of circuits are to be found.[2,3,6] Here, the description of the Buck SMPS is used to introduce several of the basic principles of analysis that apply to all SMPS circuits.

18.2.1 Buck SMPS

The Buck circuit has already been introduced in *Figure 18.3* as the direct competitor to the linear regulator. Its operating principle is the storage and release of energy in the inductor during its two states as illustrated in *Figure 18.5*.

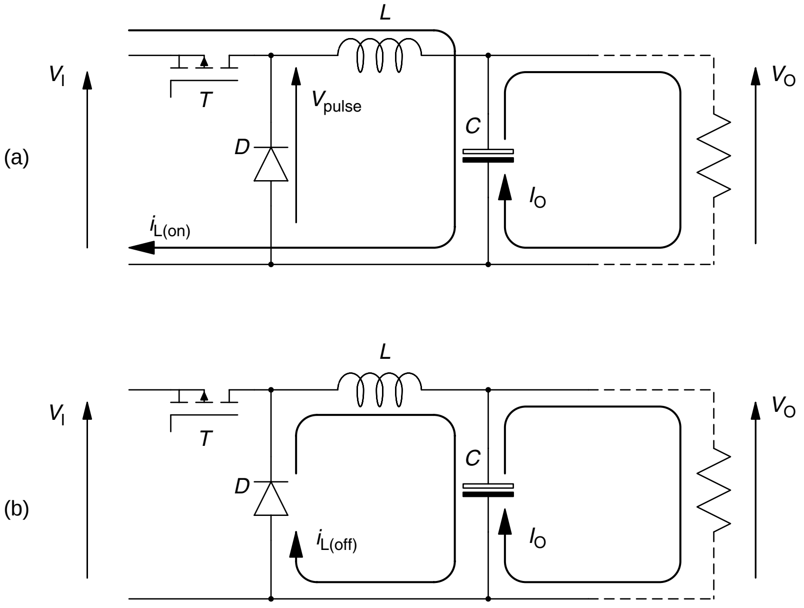


Figure 18.5 Current paths in the on- and off-states of a Buck SMPS

The operation of the circuit can be summarised as:

On-state (*Figure 18.5*(*a*)):

- A voltage is imposed across the inductor of:

$$\begin{aligned} V_L &= V_I - V_{DS(on)} - V_O \\ &\approx V_I - V_O \end{aligned}$$

- The current in the inductor increases according to:

$$\frac{\mathrm{d}i_L}{\mathrm{d}t} = \Leftarrow \frac{V_I - V_O}{L}$$

- The energy stored in the inductor increases.
- Energy is also delivered to the output.

Off-state (*Figure 18.5*(*b*)):

- The current flow through the inductor is maintained by its stored energy.
- The diode is forced into conduction to provide path for the current.
- A voltage is imposed across the inductor of:

$$\begin{aligned} V_L &= -V_{DS(on)} - V_O \\ &\approx -V_O \end{aligned}$$

- The current in the inductor decreases according to:

$$\frac{\mathrm{d}i_L}{\mathrm{d}t} = \Leftarrow \frac{V_I - V_O}{L}$$

- The energy stored in the inductor decreases.
- Energy continues to be delivered to output.

This analysis shows that the inductor current ripples up and down. The circuit will settle into a periodic steady-state condition such that the inductor current at the end of the cycle is the same value as at the beginning. The capacitor is chosen to be sufficiently large that the voltage across it changes very little during the cycle. It too will have ripple, and again the voltage at the end of the cycle will match that at the beginning.

There are two cases to examine. The first is known as continuous conduction and is the case where the decrease of current does not take it to zero. *Figure 18.6*(*a*) illustrates the principal voltage and current waveforms. The second case is known as discontinuous conduction and is where the decrease of current does take it to zero and the inductor has released all of its stored energy. The current will remain at zero until the switch is next turned on *Figure 18.6*(*b*).

Applying Kirchhoff's current law to the node at the output, we can state that for every instance in time:

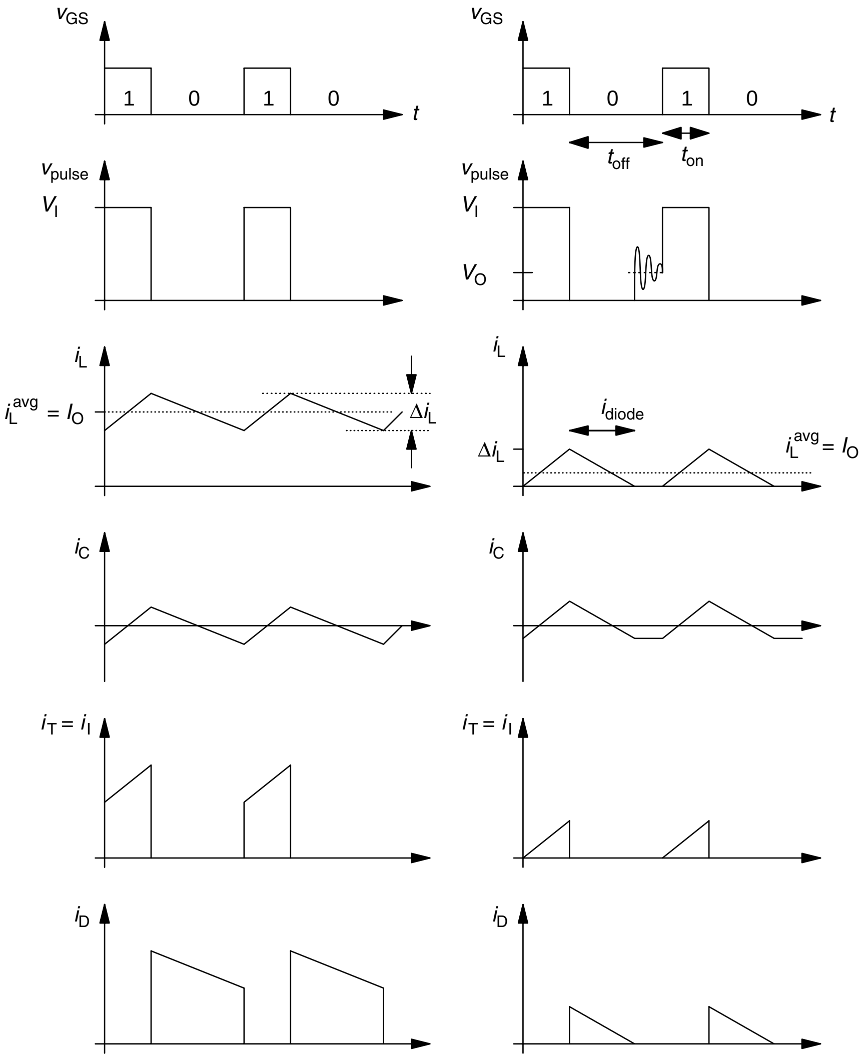
$$i_L = i_C + i_O$$

If we assume that the output current is constant (because the output voltage is very nearly so) we can write $i_O = I_O$. Then taking the average current:

$$i_L^{avg} = i_C^{avg} + I_O$$

In periodic steady-state there can be no average current flow through the capacitor. We can consider the inductor current to be composed of a constant component that flows through the load and a ripple component that diverts through the capacitor.

A relationship between the output voltage and the switching of the transistor follows from the statement that there is no net change in inductor current over a cycle. That is, the increase in current during the on-time of the transistor is matched by the decrease of current when the current flows in the diode. Here we make a distinction between the diode conduction time and the off-time because in the case of discontinuous conduction the diode does not conduct for all of the off-time.



(a) (b)

Figure 18.6 Voltage and current waveforms of the Buck SMPS in *Figure 18.5* for (a) continuous and (b) discontinuous inductor current

$$\Delta i_{L(on)} + \Delta i_{L(diode)} = 0$$

$$\Delta i_{L(on)} = \frac{di_L}{dt} t_{on} = \frac{V_I - V_O}{L} t_{on} \qquad (18.6)$$

$$\Delta i_{L(diode)} = \frac{di_L}{dt} t_{diode} = \frac{-V_O}{L} t_{diode}$$

$$\frac{V_O}{V_I} = \frac{t_{on}}{t_{on} + t_{diode}} \qquad (18.7)$$

For the continuous conduction case, $t_{diode} = T_{off}$ and the output/input ratio is simply the duty-cycle of the switch, δ.

$$\frac{V_O}{V_I} = \delta \qquad (18.8)$$

This is a very convenient relationship. It means that the output voltage is a linear function of the duty-cycle of the switch, as shown in *Figure 18.7*. This is something we can set easily and accurately. However, if the circuit operates in discontinuous mode the relationship is more complex. The conduction becomes discontinuous if the average inductor current (and therefore the output current) becomes small in comparison to the ripple. It is the magnitude of output current that dictates the proportion of the off-time for which the diode conducts. The output current in discontinuous mode is calculated by taking the average of the inductor current illustrated in *Figure 18.6(b)*.

$$I_O = i_L^{avg} = \frac{1}{T}\int^{T} i_L \cdot dt = \frac{\frac{1}{2}(t_{on} + t_{diode}) \cdot \Delta i_L}{T} \qquad (18.9)$$

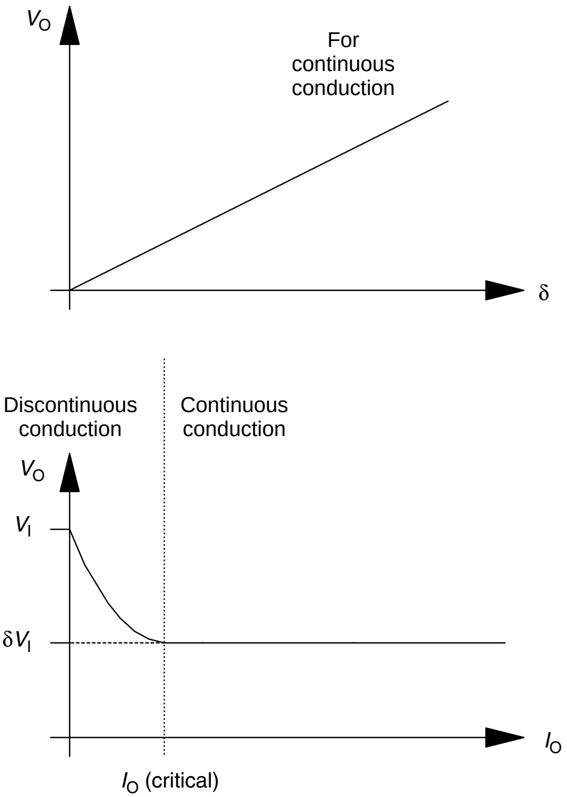


Figure 18.7 Output voltage as a function of duty-cycle and output current

Substituting Equations 18.6 and 18.9 into Equation 18.7 and using the relation $f = 1/T$ yields:

$$\frac{V_O}{V_I} = \frac{1}{1 + \frac{2 I_O L f}{V_I \delta^2}} \tag{18.10}$$

Figure 18.7 also shows a typical graph of output voltage of a buck converter operated at constant duty-cycle as the output current is increased. There are two regions: discontinuous and continuous operation. The critical conduction point is when $I_O = \frac{1}{2}\Delta i_L$. At currents higher than this, the output voltage is constant, as predicted by Equation 18.8. In practice, the voltage will decrease slightly with current because of voltage drops across the semiconductors and the resistance of the inductor. Below the critical current, the voltage rises with decreasing current, according to Equation 18.10 until reaching the input voltage at zero output current.

The switching action of the transistor leads to voltage ripple on the output capacitor. This ripple may interfere with the operation of the circuit being supplied at the output. It is important to analyse this ripple and make design choices that reduce it. The capacitor current waveform was shown in *Figure 18.6*. The ripple voltage developed across the capacitor can be found by integrating this current.

$$v_C = \frac{1}{C} \int i_C \cdot dt \tag{18.11}$$

Since the ripple must be centred on zero to have an average value of zero, we can find the ripple amplitude from the area under the positive excursion of the current (i.e. the charge delivered to the capacitor while the inductor current exceeds the load current). For continuous conduction the ripple is:

$$\Delta v_C = \frac{1}{8 f C \Delta i_L} \tag{18.12}$$

Using typical values of switching frequency and current ripple it becomes clear that capacitors of the order of 100 μF are sufficient to keep the voltage ripple in the order of 10 mV. However, this is a misleading calculation because the parasitic components of the capacitor make a significant, in fact a dominant, contribution to the ripple voltage. *Figure 18.8* shows the form (but not relative magnitude) of the ripple contributions. Another consideration is that much larger capacitors than indicated by Equation 18.12 are often used for providing some energy storage for supply-loss ride-through.

The magnitude of the voltage ripple due to the Effective Series Resistance, ESR is given by:

$$\Delta v_{ESR} = R_{ESR}\, \Delta i_L \tag{18.13}$$

The magnitude of the voltage ripple due to the Effective Series Inductance, ESL, is given by:

$$\Delta v_{ESL} = L_{ESL}\, \Delta i_L \left(\frac{1}{t_{on}} + \frac{1}{t_{off}}\right) = L_{ESL}\, \Delta i_L\, f \frac{1}{\delta(1-\delta)} \tag{18.14}$$

So, for a given current ripple, the voltage ripple due to the capacitance is inversely proportional to frequency and is normally insignificant. The voltage ripple due to the ESR is independent of frequency and is significant. The voltage ripple due to the ESL is proportional to frequency and is

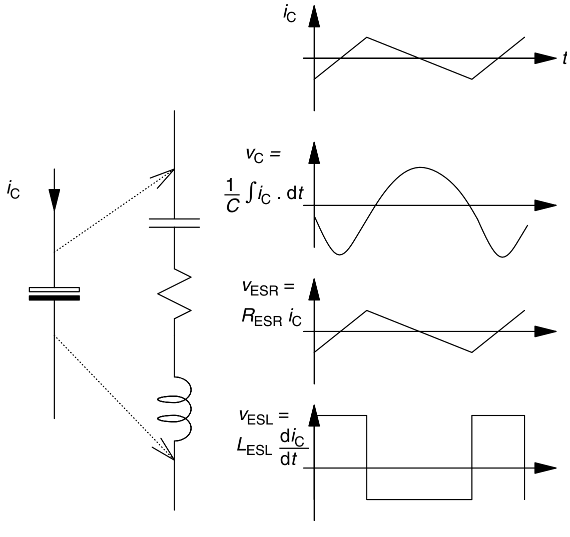

i_C
t
i_C
$v_C = \frac{1}{C}\int i_C \cdot dt$
$v_{ESR} = R_{ESR}\, i_C$
$v_{ESL} = L_{ESL}\frac{di_C}{dt}$

Figure 18.8 Contribution to ripple voltage from the effective series resistance (ESR) and inductance (ESL) of a capacitor

significant. In other words, the output capacitor is chosen for its ESR and ESL properties and its capacitance is normally satisfactory by default.

The output capacitance is not the only component for which parasitic effects are observed. *Figure 18.6*(*b*) illustrates a parasitic oscillation of V_{pulse}, the voltage across the diode, as the diode falls out of conduction when the inductor current reaches zero. In an ideal circuit we would expect V_{pulse} to rise to V_{O} so that no voltage appears across the inductor. In practice, this requires the junction capacitance of the diode to charge via the inductor as the diode enters reverse bias. Thus, a second order system is formed and a lightly damped oscillation of the voltage occurs.

18.2.1.1 Two and four quadrant chopper

The Buck SMPS is used at power ratings of less than 1 W to more than 1 MW. The larger versions are generally for control of d.c. machines in drive systems as discussed in Chapter 19 (Section 19.3.3.2, 'Step-down d.c.–d.c. converters'). The load is the armature circuit of the machine. The filtering action at the output is often inherent in the inductance of the armature winding and the inertia of the mechanical system (that keeps the speed and back-EMF constant over the short term). In some cases, extra inductance will be added to keep the current ripple low and prevent either excessive power loss in the armature resistance or torque ripple.

The simple chopper of *Figure 18.9*(*a*) is limited as a motor drive. The output voltage can only be varied between 0 and V_{s}. In particular, the voltage cannot be made negative and, therefore, reverse rotation cannot be supported. Further, there is no path for reverse armature current so it is not possible to develop reverse torque to decelerate the mechanical system (through regeneration). This circuit is known as a one-quadrant chopper because it can operate the drive only in the first quadrant of the torque–speed plane, *Figure 18.10*.

The circuit of *Figure 18.9*(*b*) allows controlled negative current. The switches are operated in anti-phase, i.e. T_1 is on for period δT and T_2 for period $(1-\delta)T$. This forces the voltage V_{A} to become a pulse train of amplitude V_{S} and duty-cycle δ regardless of the direction of I_{A}. If the direction of I_{A} is such that it cannot flow in the switch that is on, then it will flow in the anti-parallel diode. The circuit is still constrained to positive voltage. It can therefore operate a drive in the first and fourth quadrants.

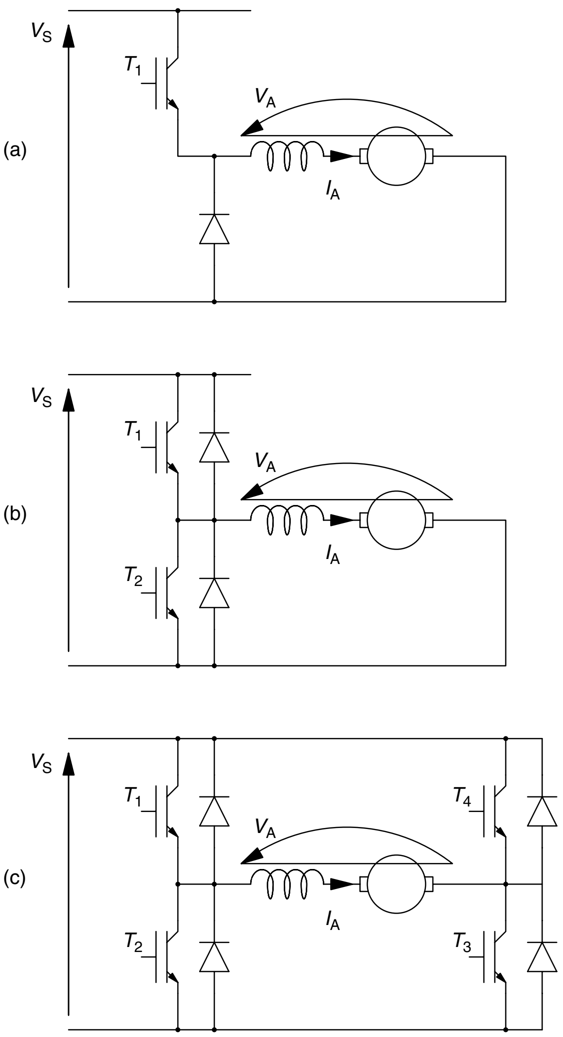


Figure 18.9 One, two and four quadrant choppers

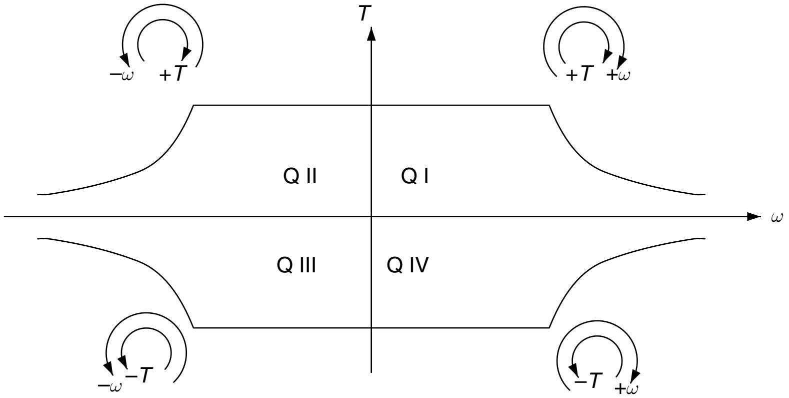


Figure 18.10 Quadrants of the torque–speed plane

The circuit of *Figure 18.9*(*c*) is variously known as a four-quadrant chopper, a full bridge or an *H*-bridge. There are several ways to operate the circuit. The transistor pair T_1 and T_2 can be operated in anti-phase as in the two-quadrant chopper. Then T_3 can be switched with T_1 and T_4 switched with T_2. Alternatively, T_3 and T_4 are not switched but simply used to select the polarity of the voltage. The circuit allows both directions of current flow and both polarities of output voltage. Therefore, the drive can be operated in all four quadrants.

18.2.2 Boost SMPS

The Boost circuit is a simple re-arrangement of the components of the Buck circuit, *Figure 18.11*.

As in the case of the Buck SMPS, energy storage in the inductor is the key to its operation. If the switch were left off for a long time, then the output capacitor would charge up via the inductor and diode to a voltage equal to the input voltage. The switching action then raises the output voltage above the input voltage hence the name Boost.

Operation of the circuit can be summarised as:

On-state (*Figure 18.12*(*a*)):

- The input voltage is imposed across the inductor:

$$\begin{aligned} V_L &= V_I - V_{DS(on)\Leftarrow} \\ &\approx V_I \end{aligned}$$

- The current in the inductor increases as:

$$\frac{di_L}{dt} = \Leftarrow \frac{V_I}{L}$$

- The energy stored in the inductor increases.
- The diode is reverse biased and the load is supplied by the capacitor alone.

Off-state (*Figure 18.12*(*b*)):

- The current flow through the inductor is maintained by its stored energy.
- The diode is forced into conduction to provide a current path.
- A voltage is imposed across the inductor of:

$$\begin{aligned} V_L &= V_I - V_{D(on)} - V_O \\ &\approx V_I - V_O \end{aligned}$$

- The current in the inductor decreases because the output voltage is higher than the input voltage:

$$\frac{di_L}{dt} = \Leftarrow \frac{V_I - V_O}{L}$$

The energy stored in the inductor decreases as it is transferred to the capacitor and output.

The inductor current ripples up and down, and again there are two cases to examine. If the average current is large then the inductor will be in continuous conduction, *Figure 18.13*(*a*). If the average current is small then the inductor will be in discontinuous conduction, *Figure 18.13*(*b*).

Again following the principle that in steady-state the change in inductor current during the on-state must be matched by the change during the off-state, we can derive a relationship between input voltage, output voltage and the timing of the switch operation.

$$\Delta i_{L(on)} + \Delta i_{L(diode)} = 0$$

$$\Delta i_{L(on)} = \Leftarrow \frac{di_L}{dt} t_{on} = \Leftarrow \frac{V_I}{L} t_{on} \qquad (18.15)\Leftarrow$$

$$\Delta i_{L(diode)} = \Leftarrow \frac{di_L}{dt} t_{diode} = \Leftarrow \frac{V_I - V_O}{L} t_{diode}$$

$$\frac{V_O}{V_I} = \frac{t_{on} + t_{diode}}{t_{diode}} \qquad (18.16)\Leftarrow$$

For the continuous conduction case where $t_{diode} = t_{off}$ the output/input relationship is a simple, but non-linear, function of δ.

$$\frac{V_O}{V_I} = \frac{1}{1 - \delta\varsigma} \qquad (18.17)\Leftarrow$$

For the discontinuous case, the relationship again becomes dependent on the circuit conditions. The relationship is found by expressing the average inductor current (equal to the input current) in terms of the diode conduction time.

$$\frac{V_O}{V_I} = \frac{1}{1 - \frac{V_I \delta^2}{2 I_I L f}} \qquad (18.18)\Leftarrow$$

It is important to note that the Boost circuit suffers a disadvantage in terms of the output voltage ripple. This is because the capacitor and output are supplied only when

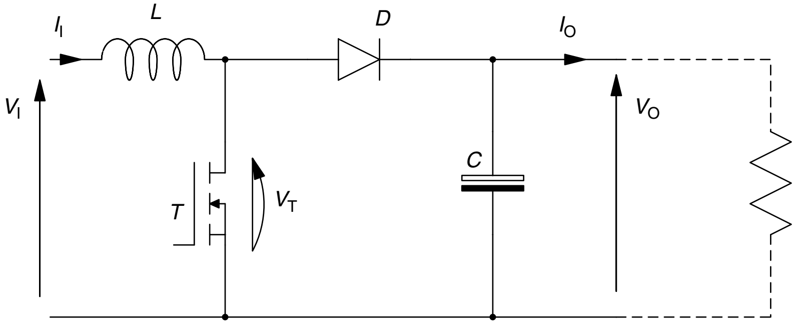


Figure 18.11 The Boost SMPS

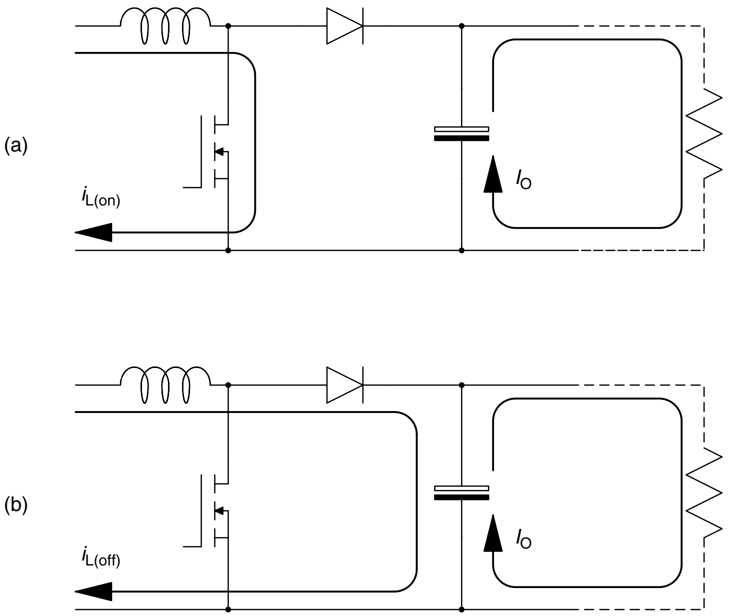


Figure 18.12 Current paths in on- and off-states of the Boost SMPS

the diode conducts, whereas in the Buck they are continually supplied via the inductor (provided that the inductor is in continuous conduction). The average component of the diode current flows onward to the load, and the ripple component flows in the capacitor. As can be seen from *Figure 18.13*, the ripple component of the diode current is larger than the ripple of the inductor current and is, in fact, at least as large as the output current.

On the other hand, the Boost SMPS draws an essentially smooth input current whereas the Buck SMPS draws a pulsed input current. This may be important if the interference generated in the input line is an issue.

18.2.3 Flyback SMPS

The Flyback SMPS is another rearrangement of the basic components as shown in *Figure 18.14*. This circuit is also known as the Buck-Boost SMPS. Some texts, such as,[3] reserve the name Flyback for the transformer isolated version of the circuit to be described in Section 18.2.5.

Operation of the circuit is as follows: the transistor is switched on to impose the input voltage across the inductor and store energy in that inductance. *Figure 18.15* shows the current paths in on- and off-states. It is clear that the release of the inductor energy by current flow through the diode charges the capacitor such that its lower plate becomes the more positive. In other words, the output voltage is negative.

The negation of the voltage is also apparent from the transfer characteristic which is derived from the assumption of steady-state in the normal way:

$$\begin{aligned} &\Delta i_{L(on)} + \Delta i_{L(diode)} = 0 \\ &\Delta i_{L(on)} = \frac{di_L}{dt} t_{on} = \frac{V_I}{L} t_{on} \qquad (18.19)\Leftarrow \\ &\Delta i_{L(diode)} = \frac{di_L}{dt} t_{diode} = \frac{V_O}{L} t_{diode} \end{aligned}$$

$$\frac{V_O}{V_I} = -\frac{t_{on}}{t_{diode}} \qquad (18.20)\Leftarrow$$

If the circuit is in continuous conduction, the transfer characteristic becomes:

$$\frac{V_O}{V_I} = -\frac{\delta\varsigma}{1-\delta\varsigma} \qquad (18.21)\Leftarrow$$

Thus, the circuit is capable of step-down (for $\delta < 1/2$) or step-up (for $\delta > 1/2$) but with a voltage negation as well.

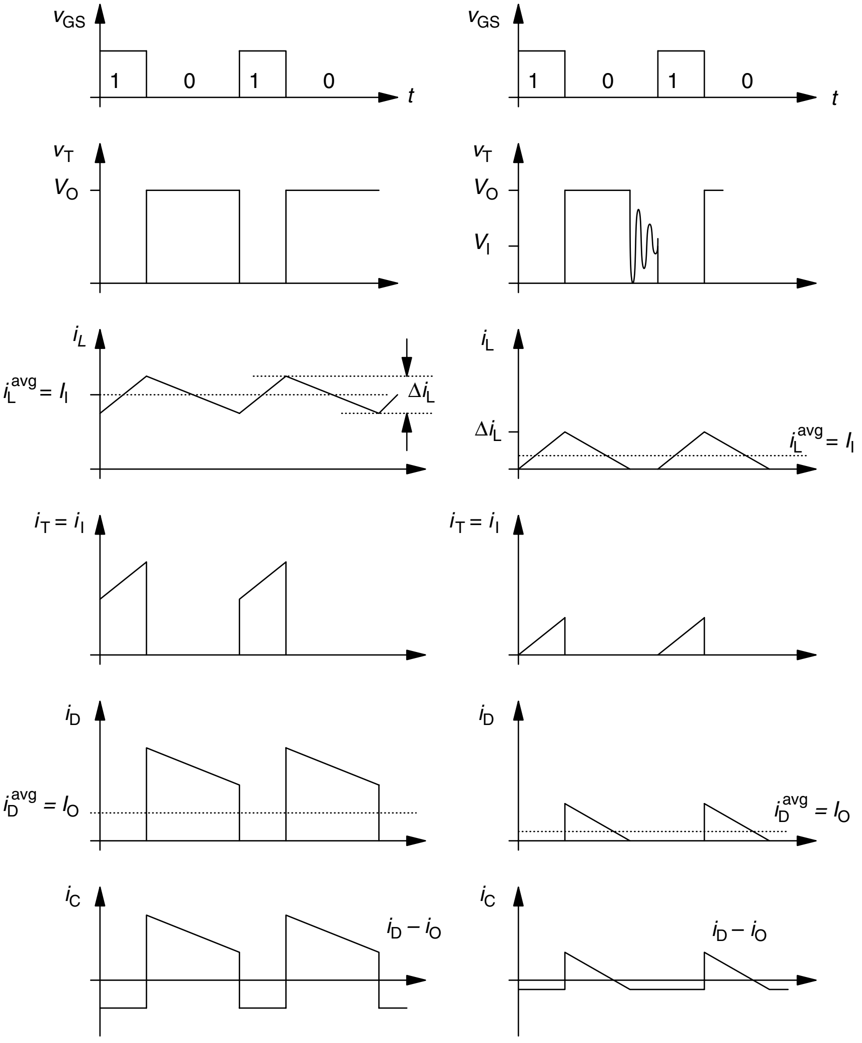
For discontinuous conduction the transfer characteristic is:

$$\frac{V_O}{V_I} = -\frac{V_O\delta^2}{2I_{I(avg)}Lf} \qquad (18.22)\Leftarrow$$

The principal current and voltage waveforms in continuous and discontinuous conduction are shown in *Figure 18.16*. It can be seen that both the input current and the capacitor current are pulsed waveforms and, as a consequence, there can be large current ripple at both the input and output.

18.2.4 Capacitor coupled SMPS

A new family of SMPS can be formed by adding an extra capacitor and an extra inductor to the basic elements, *Figure 18.17*. The capacitor is part of the energy transfer process and is (partially) charged and discharged during the cycle of operation. They have voltage transfer relationships similar to that of the Flyback (with or without the negation) and are therefore capable of step-up or step-down depending on the duty-cycle. The Čuk SMPS, *Figure 18.17(a)*, (named after its inventor) provides an identical transfer characteristic to the Flyback SMPS but has inductors in both the input connection and the charging path of the



(a)

(b)

Figure 18.13 Voltage and current waveforms of the Boost SMPS for (a) continuous and (b) discontinuous inductor current

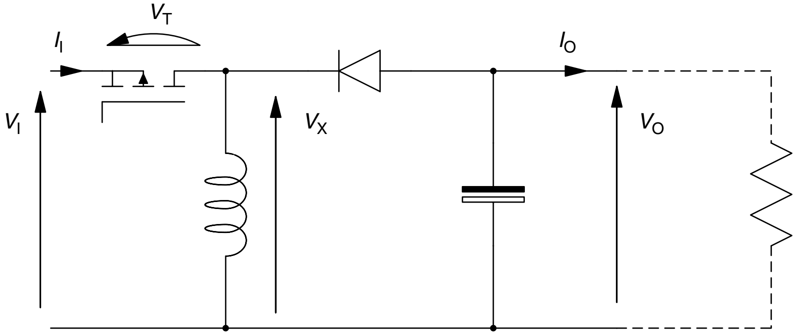


Figure 18.14 The Flyback SMPS

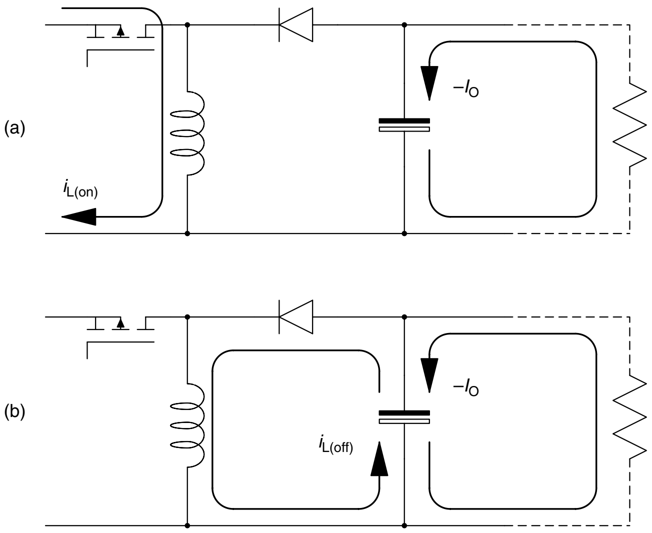
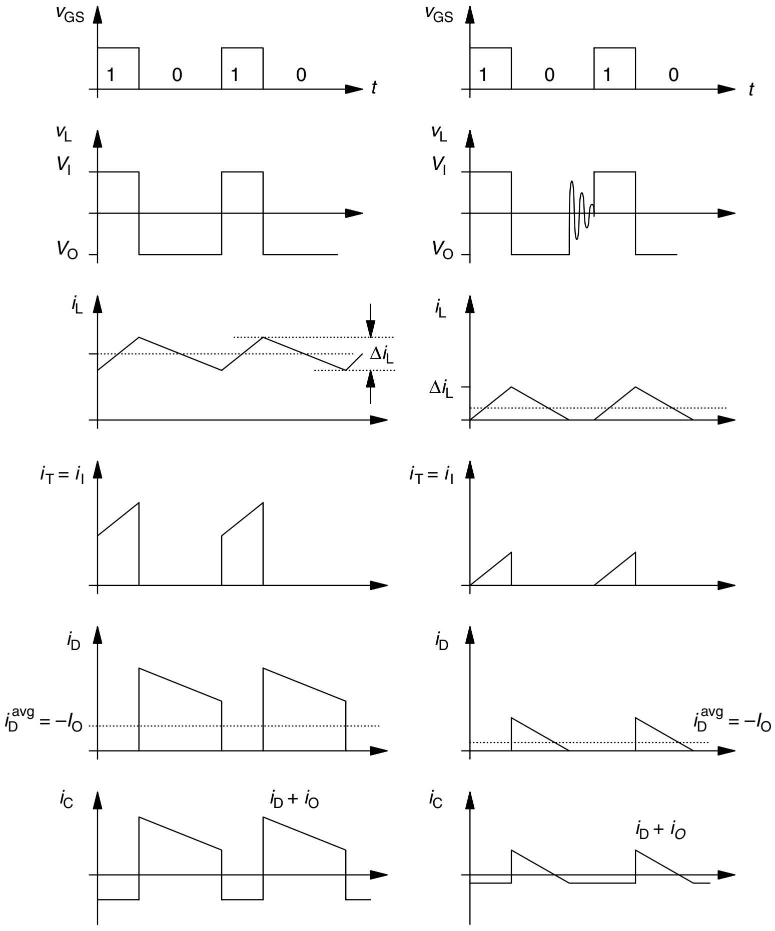


Figure 18.15 Current paths in on- and off-states of the Flyback SMPS



(a) (b)

Figure 18.16 Voltage and current waveforms of the Flyback SMPS for (a) continuous and (b) discontinuous inductor current

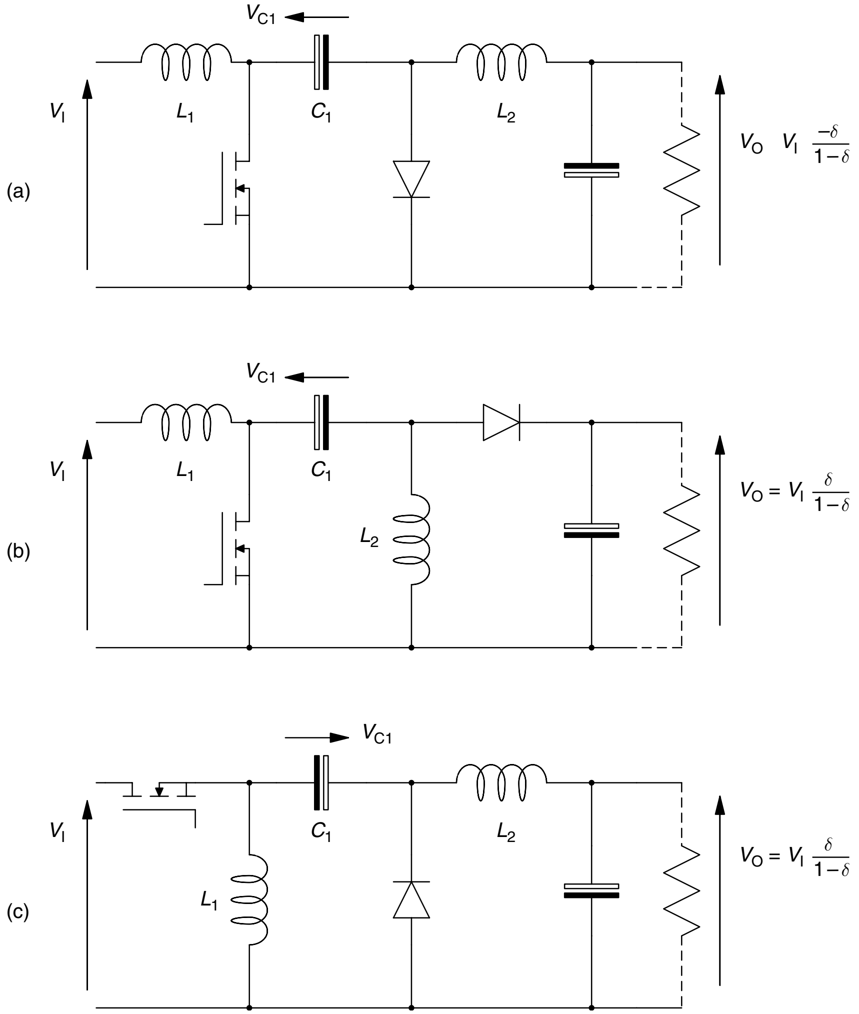

(a)
V_{C1}
V_I
L_1
C_1
L_2
V_O V_I $\frac{-\delta}{1-\delta}$
(b)
V_{C1}
V_I
L_1
C_1
L_2
$V_O = V_I \frac{\delta}{1-\delta}$
(c)
V_{C1}
V_I
L_1
C_1
L_2
$V_O = V_I \frac{\delta}{1-\delta}$

Figure 18.17 Čuk, SEPIC and Zeta SMPS

output capacitor. This provides low input current ripple and low output voltage ripple[3]. The SEPIC SMPS, *Figure 18.17*(*b*), (single-ended primary inductor converter) and Zeta SMPS, *Figure 18.17*(*c*), have an non-inverted output voltage and are capable of step-up or step-down. These two circuits differ in whether the input or output current is smoothed by an inductor.

18.2.5 Isolated flyback SMPS

Energy storage in the inductor is the key to SMPS operation. Energy can be built up by imposing one voltage and released into another. The energy is stored temporarily in the magnetic field of the inductor. If we use mutually coupled inductors then we can go one step further. We can build up stored energy in the magnetic field using one circuit and release the energy using an entirely separate circuit. Thus, we can achieve galvanic isolation between the input and output. This is useful for transferring energy from a ground-referenced supply to a non ground-referenced supply.

Figure 18.18 shows a Flyback SMPS incorporating a mutually coupled pair of inductors such that the input and output circuits do not have any direct electrical connection.

The current paths illustrated in *Figure 18.19* are very similar to those of the non-isolated Flyback circuit, *Figure 18.15*.

In analysing the operation of the circuit, it is helpful to discuss the flux, ϕ in the inductor core rather than the current. These two quantities are closely related, Equation 18.23 (which is written to include the possibility of more than two coupled inductors), but flux is the more useful quantity because it more directly relates to the stored energy in the magnetic field and is the quantity that must remain continuous from instant to instant (i.e. have a finite derivative).

$$\phi = \frac{N_1}{\mathscr{R}} i_1 + \frac{N_2}{\mathscr{R}} i_2 + \cdots$$

or (18.23)

$$\phi = \frac{L_1}{N_1} i_1 + \frac{L_2}{N_2} i_2 + \cdots$$

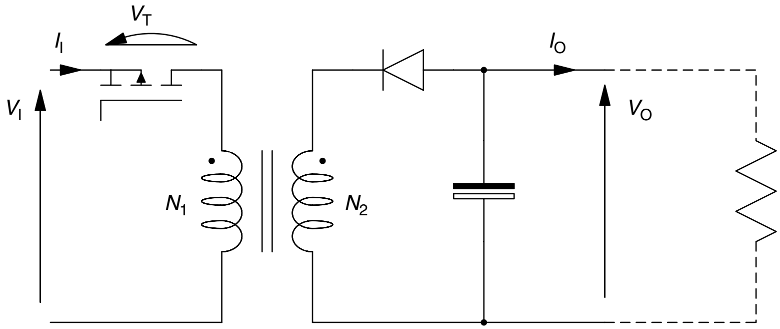


Figure 18.18 An isolated Flyback SMPS using a mutually coupled pair of inductors

where N is the number of turns of a winding, L is its inductance and $\mathcal{R}$ is the reluctance of the magnetic path.

The voltage induced across any of the mutually coupled inductors is given by Faraday's Law and for the first inductor is expressed as:

$$E_1 = N_1 \frac{d\phi}{dt}$$

$$= \frac{N_1^2}{\mathcal{R}} \frac{di_1}{dt} + \frac{N_1 N_2}{\mathcal{R}} \frac{di_2}{dt} + \cdots \qquad (18.24)$$

or

$$= L_1 \frac{di_1}{dt} + M_{12} \frac{di_2}{dt} + \cdots$$

With these relations in mind, we can draw the principal voltage and current waveforms as shown in *Figure 18.19*. There are two cases: continuous conduction (defined as the flux in the core not reaching zero by the end of the cycle) and discontinuous conduction (defined as flux reaching and holding zero by the end of the cycle). The flux rises linearly while the transistor is on because of the voltage imposed across the primary. When the transistor switches off, the primary current must stop and the diode is forced into conduction so that secondary current can flow to maintain the flux. The primary and secondary side currents are scaled versions of the flux according to the number of turns. When the diode is conducting, a voltage is imposed across the secondary and this is reflected on the primary according to the turns-ratio. Thus, the transistor must support $V_I + V_O\, N_1/N_2$ while it is off.

The transfer characteristic can be evaluated from rise and fall of flux in steady-state.

$$\Delta\phi_{(on)} + \Delta\phi_{(diode)} = 0$$

$$\Delta\phi_{(on)} = \frac{d\phi}{dt} t_{on} = \frac{V_I}{N_1} t_{on}$$

$$\Delta\phi_{(diode)} = \frac{d\phi}{dt} t_{diode} = \frac{V_O}{N_2} t_{diode}$$

$$\frac{V_O}{V_I} = -\frac{N_2}{N_1} \frac{t_{on}}{t_{diode}} \qquad (18.25)$$

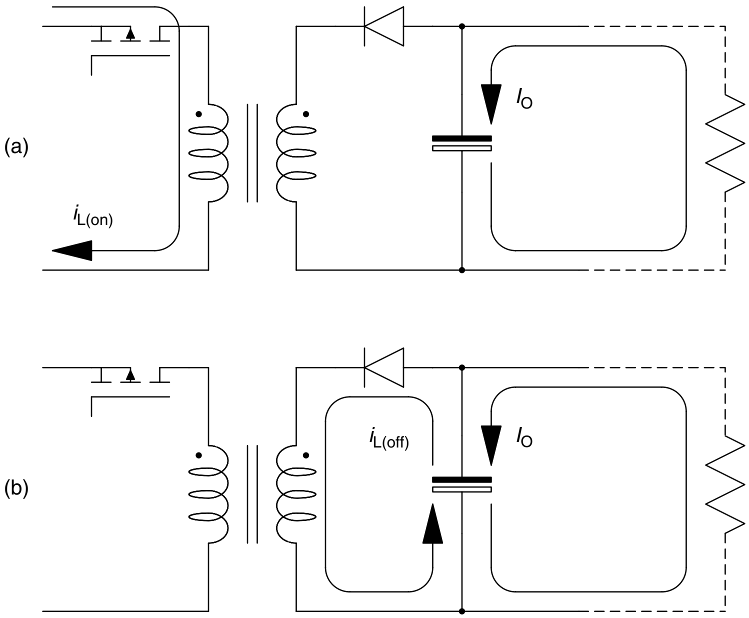


Figure 18.19 Current paths in on- and off-states of the isolated Flyback SMPS

Because the output is isolated, we can choose to take the output voltage in either sense and it is more normal to see the circuit drawn as in *Figure 18.21*. In this circuit the primary has also been rearranged to give a ground referenced MOSFET connection. The transfer characteristic for continuous conduction is:

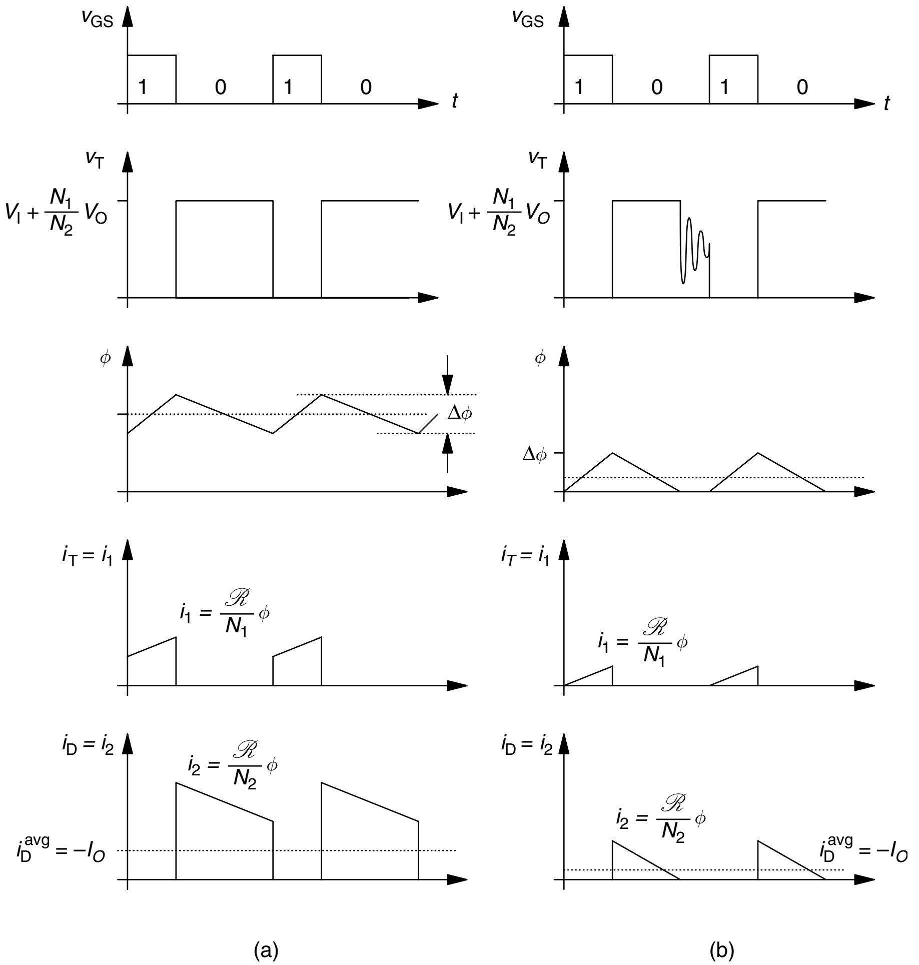
$$\frac{V_O}{V_I}=\frac{N_2}{N_1}\frac{\delta\varsigma}{1-\delta\varsigma} \qquad (18.26)\Leftarrow$$

Discontinuous operation is sometimes favoured because it is easy to ensure that the core does not saturate. It is quite useful from a control standpoint because a fixed amount of energy is delivered to the output in each cycle.

The waveforms of *Figure 18.20* were drawn assuming perfect coupling of the flux to each winding. In practice, this will not be the case and each winding will have a leakage field. The consequence of this is that the currents in the primary and secondary cannot rise and fall instantaneously. In particular, when the transistor is switched off the primary current will not cease until the energy in the primary leakage field has been removed. The current will overcharge the parasitic capacitance of the transistor, as illustrated in *Figure 18.22*.

The overshoot can be estimated by assuming that all of the energy previously stored in the leakage field is transferred to the drain-source capacitance. The voltage overshoot has important implications for the rating of the transistor. Using a resistor-capacitor-diode snubber can reduce the over-voltage[6] but power loss in the snubber is problematic and it is difficult to design it for a wide range of operating currents. More sophisticated clamp-like snubbers are difficult to apply because of the lack of a suitable voltage rail to which to clamp.

The transistor over-voltage problem can be solved, at the expense of another transistor, by switching both ends of the primary winding as shown in *Figure 18.23*. Each transistor can be rated for V_I rather than $V_I+V_O N_1/N_2+V_{\text{overshoot}}$ which would be required for the single transistor. The energy stored in the leakage field of the primary is returned to the supply via the two diodes. It is important that these diodes do not conduct in place of the output-side diode during normal operation and so the reflected secondary voltage $V_O N_1/N_2$ must be kept less than V_I.



(a) (b)

Figure 18.20 Flux, voltage and current waveforms of the isolated Flyback SMPS for (a) continuous and (b) discontinuous inductor flux

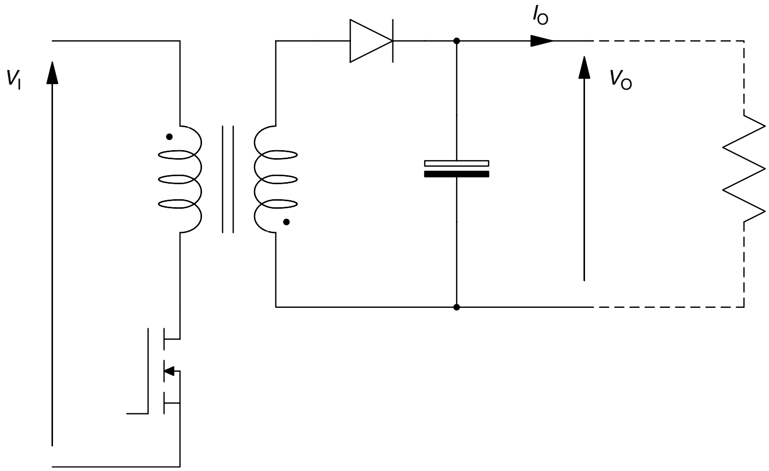
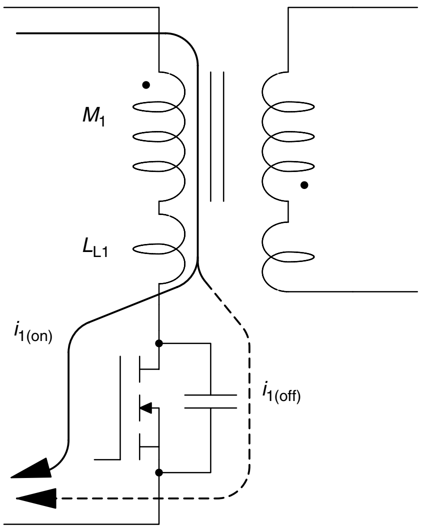


Figure 18.21 Normal arrangement of isolated Flyback SMPS



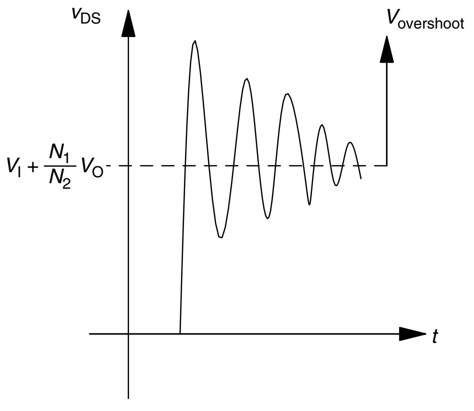


Figure 18.22 Overshoot of transistor voltage at turn-off

18.2.6 Transformer isolated buck SMPS

Some SMPS designs do not lend themselves to direct incorporation of a mutually coupled inductor. Instead, the energy storage function occurs in a single inductor and a further mutually coupled pair of inductors is introduced to provide isolation. This is illustrated by the isolated Buck SMPS circuits of *Figure 18.24*. All three versions of the circuit apply an isolated pulse train to the standard LC output stage of a Buck circuit. The transfer characteristic in each case is similar to the Buck SMPS with a scaling factor of the turns ratio of the transformer:

$$\frac{V_O}{V_I} = \frac{N_S}{N_P}\delta \tag{18.27}$$

The differences between the circuits are in the ratings of the transistors, the control of leakage energy and the utilisation of the magnetic core.

The circuit (*Figure 18.24*(*a*)) has a well-defined path in which to recover the primary leakage energy. It also causes the transformer flux to swing both positive and negative. Thus, the core can be utilised from its saturation limit in one direction to its saturation limit in the other. The transistors in the circuit can be rated for the input voltage with no significant overshoot. These three advantages are gained at the expense of using a 4-transistor full bridge. The duty-cycle should be defined as $(t_{on1} + t_{on2})/T$ for use in Equation 18.27 (where t_{on1} and t_{on2} are the times for which positive and negative voltage are applied to the primary; for the remaining time the voltage should be zero). It is essential to keep $t_{on1} = t_{on2}$ in order to avoid 'stair-case' saturation of the transformer core (small net increments of flux for each cycle of operation).

The circuit (*Figure 18.24*(*b*)) uses a bifilar primary winding to achieve bi-directional flux swings with only two transistors. However, the voltages reflected from one primary to the other require the transistors to support twice the input voltage during their off-state. Further, there is no path to discharge the primary leakage energy and voltage overshoot occurs. Again, the duty-cycle is defined by the sum of the two conduction periods for use in Equation 18.27.

The circuit (*Figure 18.24*(*c*)) (often known as the forward converter) uses only one transistor and can only create a unidirectional flux swing (thus requiring twice the core

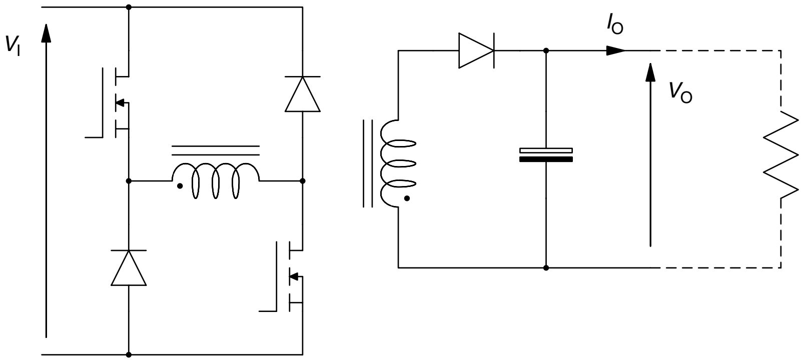


Figure 18.23 Half-bridge Flyback SMPS

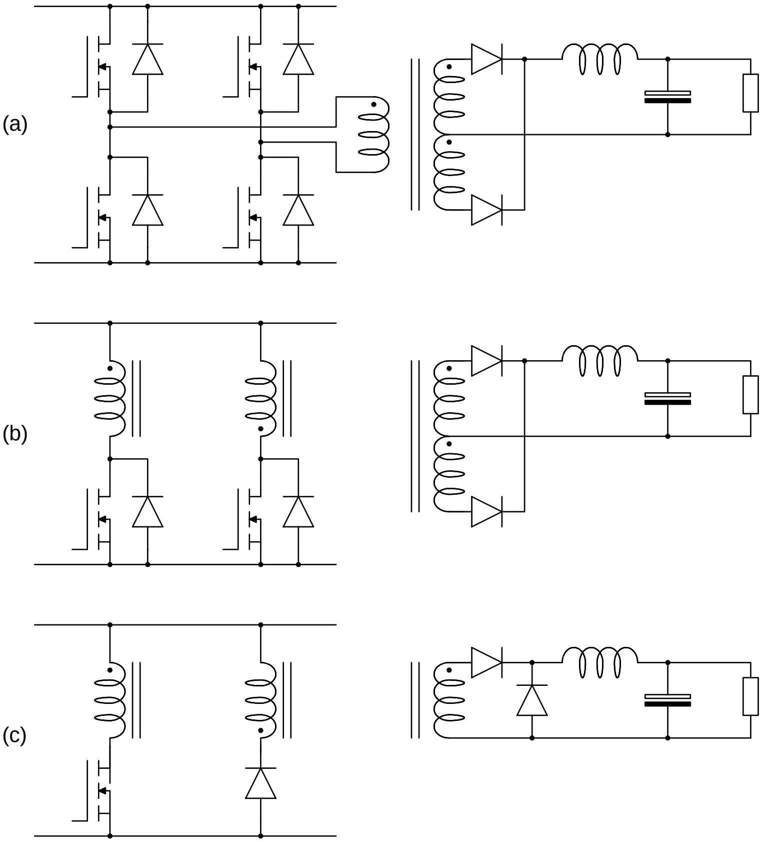


Figure 18.24 Isolated Buck SMPS: (a) full bridge; (b) push-pull and (c) forward

volume to hold off saturation for the same voltage and frequency conditions as the other circuits). The bifilar primary winding is there only to discharge the magnetising current of the transformer. There is no discharge path for the primary leakage energy and so voltage overshoot occurs.

18.2.7 SMPS control

It is rare to see a SMPS run in open loop. Closed loop operation is essential to ensure that the output voltage is accurately maintained at the desired value. The output voltage can be affected by changes in the output current or input voltage. There are several imperfections in practical SMPS that were not incorporated in the analysis of the previous sections such as voltage drops across the semiconductors and across the resistance of the inductor. The change from discontinuous to continuous mode will also affect the output voltage. All of these factors can be mitigated by feedback control. To achieve this we require a circuit that can vary, or modulate, the pulse-width applied to the transistor under the influence of some other signal. The usual approach is to use a triangle wave or saw-tooth

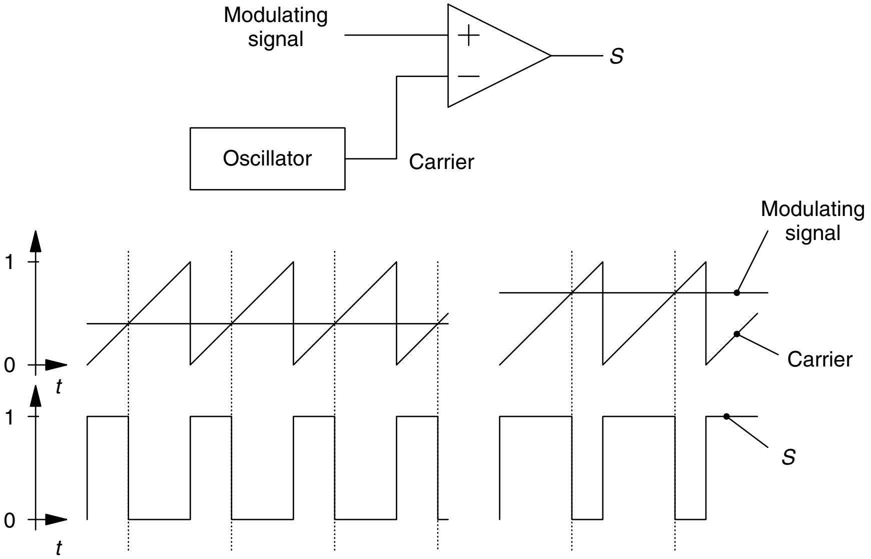


Figure 18.25 A pulse-width modulator with saw-tooth carrier

wave carrier. The carrier is compared with the modulating signal as shown in *Figure 18.25*. The comparator gives a high output whenever the modulating signal is greater than the carrier and thus produces pulses with a width proportional to the modulating signal.

A schematic of the closed loop system is shown in *Figure 18.26*. The forward transfer function, *F*(*s*) of the system is complex. The modulator is linear over its normal range but saturates outside that range. The SMPS circuit is time varying because it has two operating modes. Even in continuous mode, most SMPS have a non-linear transfer function. The parasitic components of the circuit, particularly the ESR of the output capacitor, can form significant terms in the transfer function of the circuit. The output voltage is normally found to decrease as output current is drawn (because of the increased voltage drop across resistances and semiconductors in the circuit). Variation of the input voltage forms a further disturbance input to the system.

On a simple view, the application of feedback solves many of these problems. The closed loop transfer function, taking into account the two disturbance terms, is:

$$\begin{aligned} v_O(s) &= \Leftarrow\frac{C(s)F(s)\Leftarrow}{1+C(s)F(s)G(s)}\cdot v_R(s)\Leftarrow \\ &\quad+\frac{D_1(s)}{1+C(s)F(s)G(s)}\cdot v_I(s) \qquad (18.28)\Leftarrow \\ &\quad+\frac{D_2(s)}{1+C(s)F(s)G(s)\Leftarrow}\cdot i_O(s)\Leftarrow \end{aligned}$$

Provided that the loop gain of the system, *C*(*s*)*F*(*s*)*G*(*s*), is large, then the disturbances are largely rejected. Further, the transfer function (from reference to output) is dominated by the feedback attenuator, *G*(*s*). The feedback attenuator is normally a potential divider and has a simple linear gain.

$$\begin{aligned} v_o(s) &\approx \Leftarrow\frac{1}{G(s)\Leftarrow}\cdot v_R(s)\Leftarrow \\ &\approx k v_R(s) \qquad (18.29)\Leftarrow \end{aligned}$$

where *k* is the reciprocal of the feedback gain.

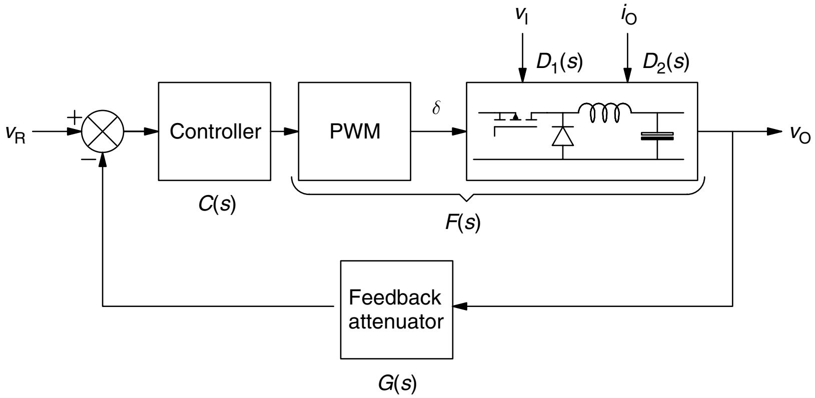


Figure 18.26 A closed loop controller schematic for an SMPS

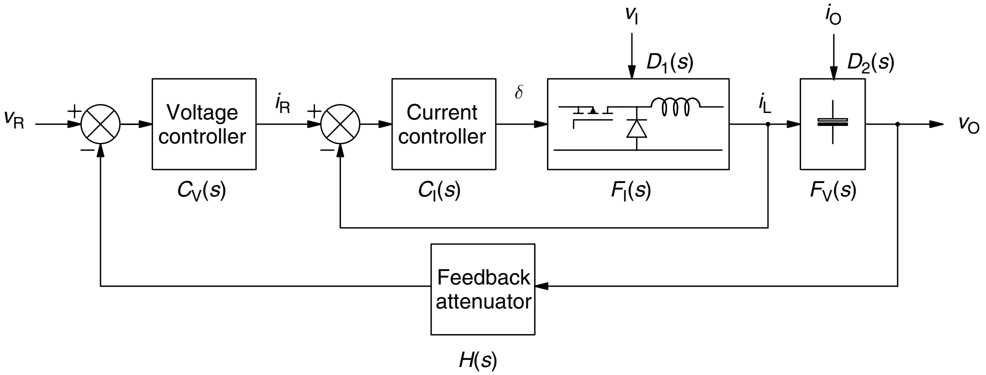


Figure 18.27 Current mode control using nested control loops

The control element $C(s)$ must be chosen with care to ensure stability. Techniques such as state-space averaging[3] can be used to derive linearised dynamic models of the SMPS circuits to facilitate such design. (Note that the transfer characteristics, as functions of δ, derived elsewhere in Section 18.2 express only the steady-state properties of the circuits.)

The usual elements of a controller for an SMPS (that is, a pulse-width modulator, error amplifier, an adjustable gain and a gate drive circuit) are available as standard integrated circuits. Often included is some form of current limit. A low value resistor placed in the ground line of the supply is used to monitor the current. Above some threshold value, the pulse-width of the gate-drive is reduced to reduce the output voltage and limit the current.

The non-linear, time-varying nature of the SMPS makes controller design problematic and it is often difficult to ensure stability across a wide range of operating conditions. A technique that can alleviate the problems is known as current-mode control. It is based on the feedback of two signals, in this case through two nested control loops, *Figure 18.27*.

The choice of nested loops follows from recognising that the rate of change of inductor current is much higher than the rate of change of output voltage. Therefore, the inductor current can be controlled in a fast inner loop and the output voltage is controlled in a slower outer loop. If the inner loop tracks the current reference accurately then the outer loop becomes first-order dominated and is relatively easy to design.

The generation of pulses to drive the transistor is somewhat different from the voltage control case. The current demand set by the outer loop forms a reference. The transistor is turned on at regular intervals and stays on until the current ramps up to the reference level. The transistor then remains off until the next of the regular turn-on events. Thus, the inductor current is controlled within each switching cycle.

A further advantage of current mode control is that feedforward of input voltage disturbances can be applied readily. There is a stability issue to be addressed at duty-cycles of greater than 1/2, however, this can be overcome by a method known as slope-compensation.[3]

18.3 D.c./a.c. conversion

18.3.1 Single phase bridge

The full-bridge circuit was introduced in Section 18.2.1.1 as a way of converting d.c. to ±d.c. It was used again in Section 18.2.6 to create a bi-directional voltage on the primary of a transformer. We can use the same circuit (repeated but re-labelled here as *Figure 18.28*) to convert d.c. to a.c. We can control the duty-cycle of the switches to produce a waveform with an average that gradually increases in a positive

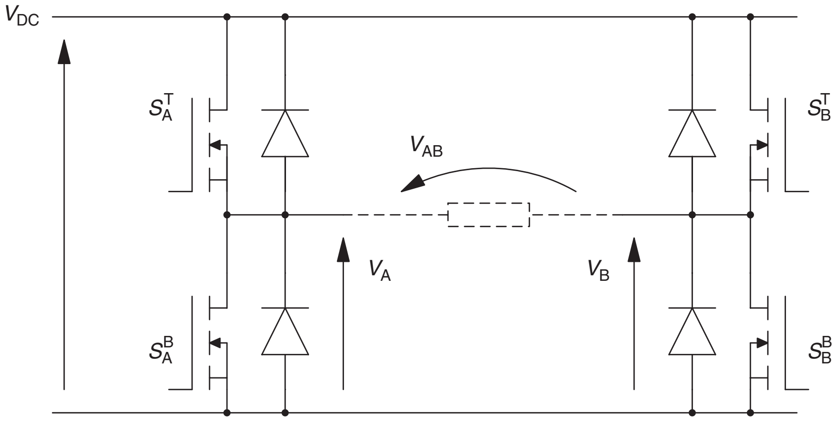


Figure 18.28 Full bridge with switching signal assignment of *Figure 18.30*

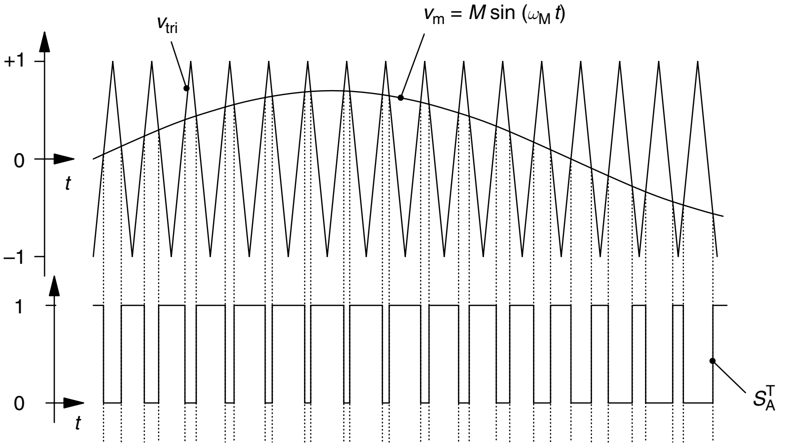


Figure 18.29 Sinusoidally modulated pulse-widths

sense and then decreases and turns negative, *Figure 18.29*. The duty-cycle is modulated as:

$$\delta = \tfrac{1}{2} + \tfrac{1}{2}M \sin(\omega_M t) \quad \text{where } 0 \le M \le 1 \tag{18.30}$$

where M is known as the depth of modulation and will set the magnitude of the a.c. voltage being synthesised.

Top and bottom switches, such as S_A^T and S_A^B, will be operated in anti-phase but not strictly so. Because each switch takes time to turn off and time to turn on, it is important to initiate turn-off of one switch before turn-on of the other. This avoids both switches being partially on at the same time and providing a shoot-through path between the supply rails. This is known as providing dead-time or under-lap.

As indicated in *Figure 18.29*, the modulation of the duty-cycle can be performed through a comparison of the desired sinewave with a triangular carrier. An analogue circuit similar to that described in Section 18.2.7 could be used. It is more common for an inverter (d.c./a.c. converter) to be controlled with a microcontroller or digital signal processor, DSP. The transistors are switched using the timer channels of the processor. The sinewave modulation is achieved by regularly updating the timer values with times calculated using a look-up table of sinewave samples. Scaling the sinewave data controls the magnitude of the output voltage. The rate at which progress is made through the sinewave look-up table determines the frequency of the sinewave synthesised at the output.

It is common to phase shift the carrier waveform of the two sides of the bridge as shown in *Figure 18.30*.

The output voltage waveforms are also shown in *Figure 18.30*. Each side of the bridge can be switched between $+V_{DC}$ and 0. The voltage appearing across the output can be $+V_{DC}$, 0 or $-V_{DC}$. These waveforms are known as two-level and three-level pwm respectively. It can be seen that in the case of a phase-shifted carrier the 3-level waveform has twice the frequency of the carrier.

The low-frequency components of these voltages are:

$$\begin{aligned} v_{A0} &= \tfrac{1}{2}V_{DC} + \tfrac{1}{2}V_{DC}M\sin(\omega_M t) \\ v_{B0} &= \tfrac{1}{2}V_{DC} - \tfrac{1}{2}V_{DC}M\sin(\omega_M t) \\ v_{AB} &= V_{DC}M\sin(\omega_M t) \end{aligned} \tag{18.31}$$

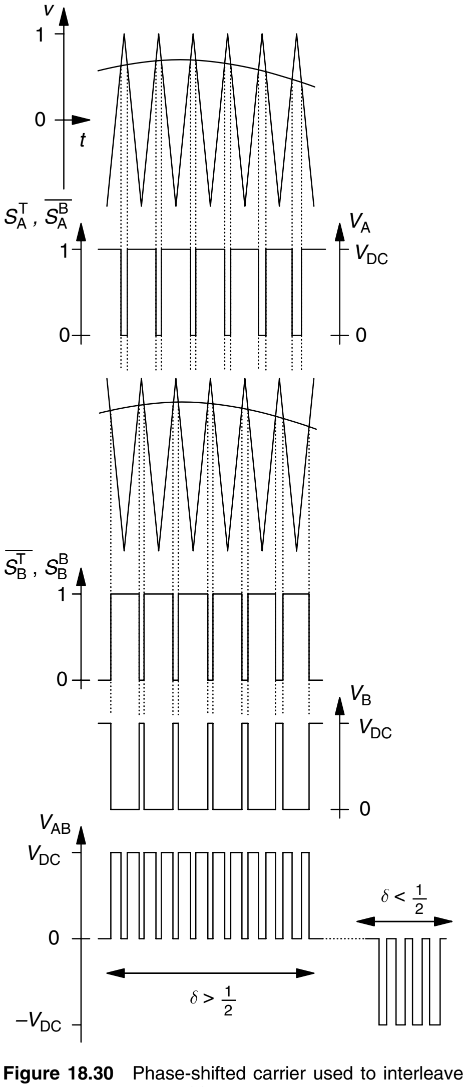


Figure 18.30 Phase-shifted carrier used to interleave PWM of left- and right-hand sides of full bridge.

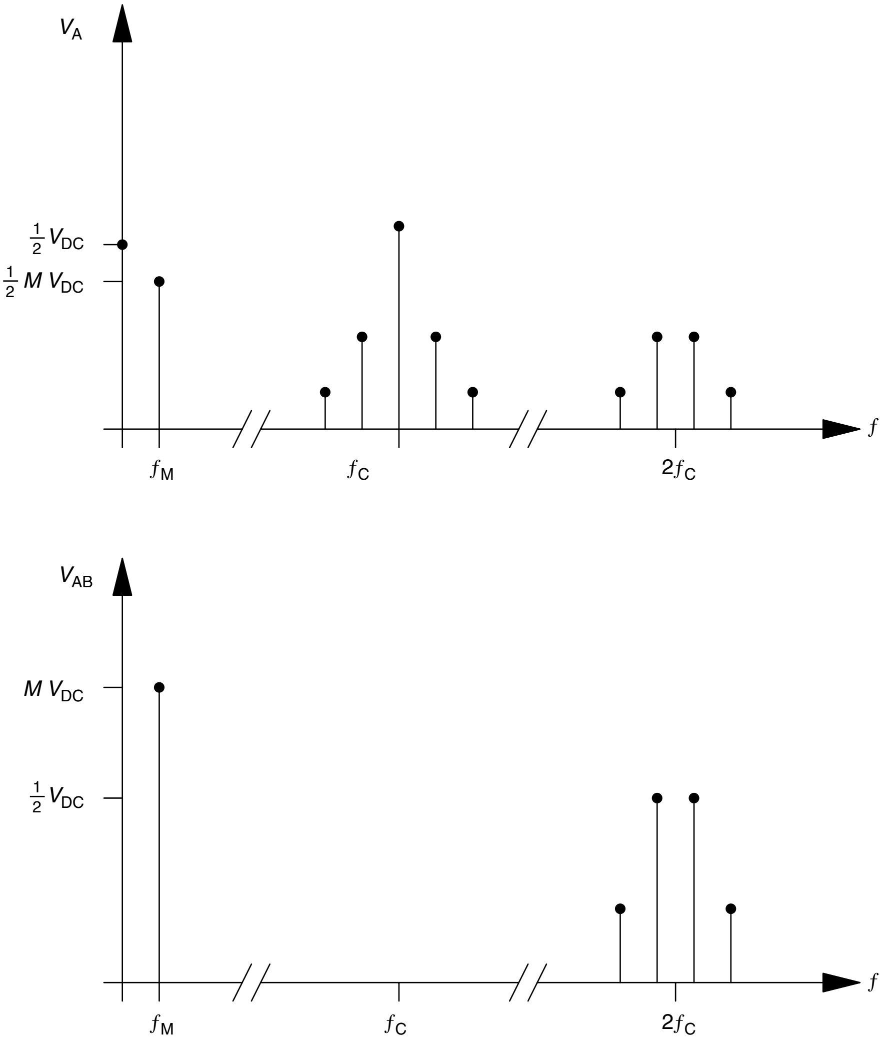


Figure 18.31 Frequency spectra of a half bridge voltage and a full bridge voltage with phase-shifted carrier

The d.c. portions of the voltage at each half of the bridge are common mode and do not appear in the output voltage.

The objective of this circuit is to produce sinusoidal voltage but what we have is a sinusoidally modulated rectangular wave. The difference is illustrated in the frequency spectra of *Figure 18.31*. The details of the spectra depend on how the modulation was implemented, but the basic form shown here is typical.

18.3.2 Three phase bridge

For most three-phase applications, the requirement is to produce a balanced set of voltages for a three-wire system. In this case, a pair of transistors is used to produce an a.c. waveform for each phase voltage. The common mode (or zero-sequence) d.c. component is unimportant in a three-wire system.

The three phase voltages V_A, V_B and V_C are modulated to give a balanced three-phase set:

$$\begin{aligned}
v_A &= \tfrac{1}{2}V_{DC} + \tfrac{1}{2}V_{DC}M\sin(\omega t)\\
v_B &= \tfrac{1}{2}V_{DC} + \tfrac{1}{2}V_{DC}M\sin\left(\omega t - \tfrac{2\pi}{3}\right)\\
v_C &= \tfrac{1}{2}V_{DC} + \tfrac{1}{2}V_{DC}M\sin\left(\omega t + \frac{2\pi}{3}\right)
\end{aligned} \tag{18.32}$$

$$\begin{aligned}
v_{AB} &= \frac{\sqrt{3}}{2}V_{DC}M\sin\left(\omega t + \frac{\pi}{6}\right)\\
v_{BC} &= \frac{\sqrt{3}}{2}V_{DC}M\sin\left(\omega t - \frac{\pi}{2}\right)\\
v_{CA} &= \frac{\sqrt{3}}{2}V_{DC}M\sin\left(\omega t + \frac{5\pi}{6}\right)
\end{aligned} \tag{18.33}$$

All three phases are modulated using the same carrier and so there is no doubling of the effective switching rate as in

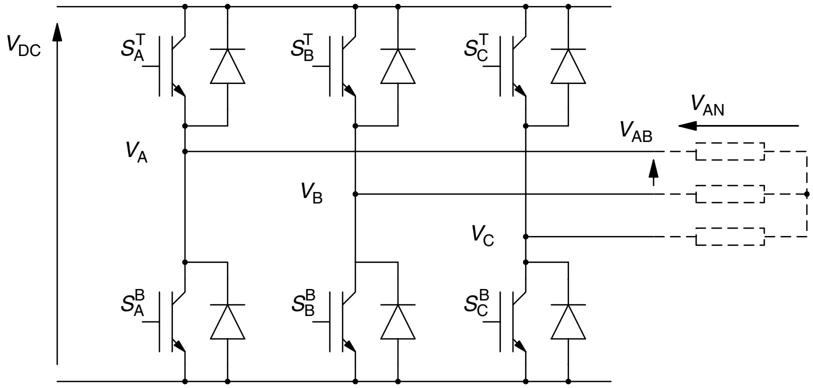

V_{DC}
S_A^T
S_B^T
S_C^T
V_{AN}
V_{AB}
V_A
V_B
V_C
S_A^B
S_B^B
S_C^B

Figure 18.32 A three-phase inverter bridge

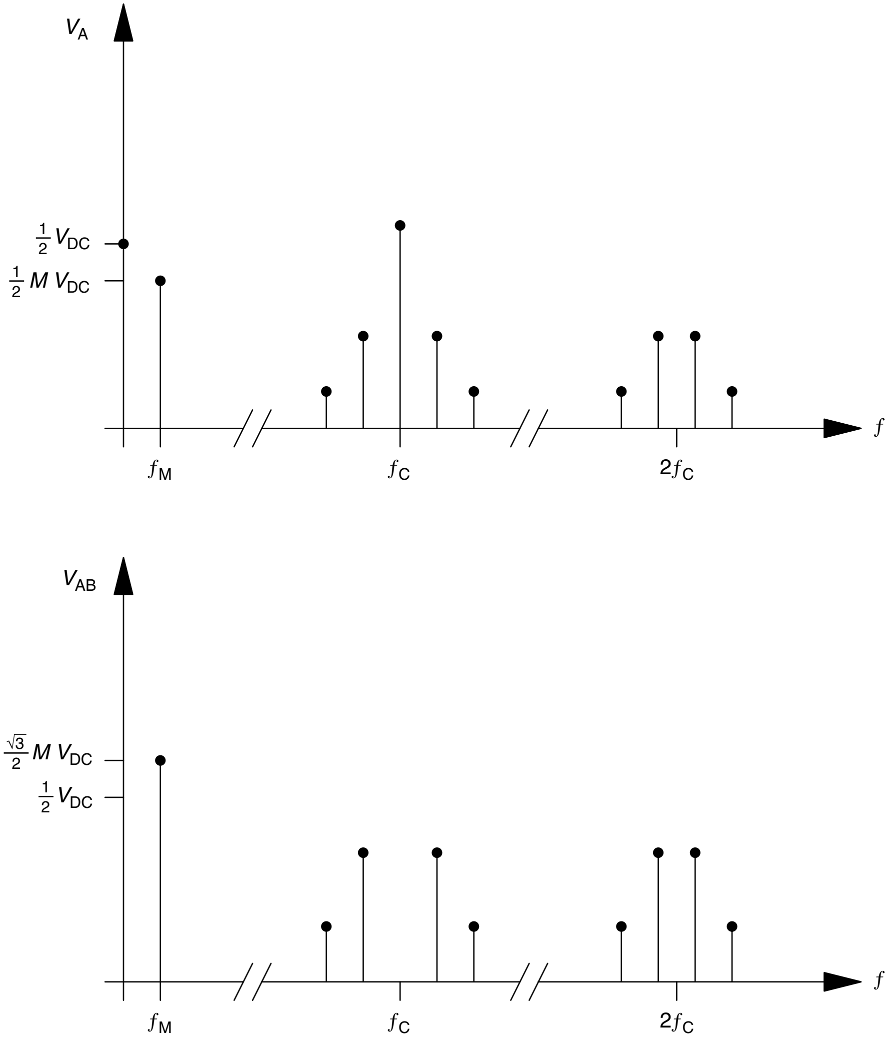


Figure 18.33 Frequency spectra of the phase and line voltages of the three-phase inverter of *Figure 18.32*

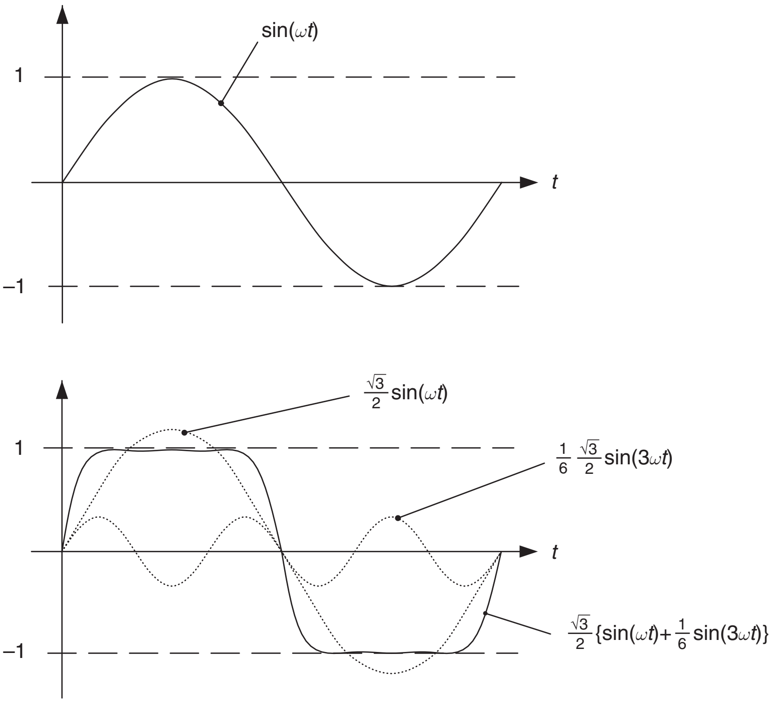


Figure 18.34 Increase of maximum fundamental by addition of third harmonic

the single-phase bridge. However, the carrier component is zero-sequence and will not appear in the line voltages to the phase voltages at the load. The frequency spectra of the phase and line voltages are shown in *Figure 18.33*.

The fact that zero-sequence components can be inserted into the inverter phase voltage waveforms without affecting the line voltage waveforms can be useful. The peak line voltage indicated above is $\sqrt{3}/2V_{\mathrm{DC}}$. It should be possible to achieve a peak line voltage of V_{DC} by switching on the top switch of one phase and the bottom switch of another. This can be realised for a sinewave pattern by including a zero-sequence third harmonic set in the waveforms as shown in *Figure 18.34*. The fundamental component can be boosted by $2/\sqrt{3}$ without exceeding a duty cycle of unity if the third harmonic is 1/6th of amplitude of the fundamental.

There is an alternative way of forming the switching waveforms of the inverter known as space-voltage vector modulation, SVM. It is based on the fact that there are only eight useful states of the inverter as indicated in *Table 18.1*. When a bottom switch is on, the voltage of that phase will be zero; and when a top switch is on, the voltage of that phase will be V_{DC}. We discount states where neither or both of the switches in a phase are on. The zero sequence voltage is the average of the three phase voltages. The overall voltage output of the inverter is represented as a space-voltage vector. That is, the voltage of each phase is taken to have a spatial position matching the spatial positioning of the phase windings in an electrical machine. This is equivalent to a Clarke transform[4] to $\alpha\beta\gamma\varsigma$co-ordinates:

Table 18.1 Inverter states and resultant space-voltage vector

State	*Switches on*			*Phase voltages*			*Zero sequence* $V_{\gamma\varsigma}=\Leftarrow V_{\mathrm{N}}$	*Space-voltage vector* $\boldsymbol{V}$
	A	*B*	*C*	V_{A}	V_{B}	V_{C}		
0	*B*	*B*	*B*	0	0	0	0	0
1	*T*	*B*	*B*	V_{DC}	0	0	$\frac{1}{3}V_{\mathrm{DC}}$	$V_{\mathrm{DC}}\angle 0^{\circ\Leftarrow}$
2	*T*	*T*	*B*	V_{DC}	V_{DC}	0	$\frac{2}{3}V_{\mathrm{DC}}$	$V_{\mathrm{DC}}\angle 60^{\circ\Leftarrow}$
3	*B*	*T*	*B*	0	V_{DC}	0	$\frac{1}{3}V_{\mathrm{DC}}$	$V_{\mathrm{DC}}\angle 120^{\circ\Leftarrow}$
4	*B*	*T*	*T*	0	V_{DC}	V_{DC}	$\frac{2}{3}V_{\mathrm{DC}}$	$V_{\mathrm{DC}}\angle 180^{\circ\Leftarrow}$
5	*B*	*B*	*T*	0	0	V_{DC}	$\frac{1}{3}V_{\mathrm{DC}}$	$V_{\mathrm{DC}}\angle 240^{\circ\Leftarrow}$
6	*T*	*B*	*T*	V_{DC}	0	V_{DC}	$\frac{2}{3}V_{\mathrm{DC}}$	$V_{\mathrm{DC}}\angle 300^{\circ\Leftarrow}$
7	*T*	*T*	*T*	V_{DC}	V_{DC}	V_{DC}	V_{DC}	0

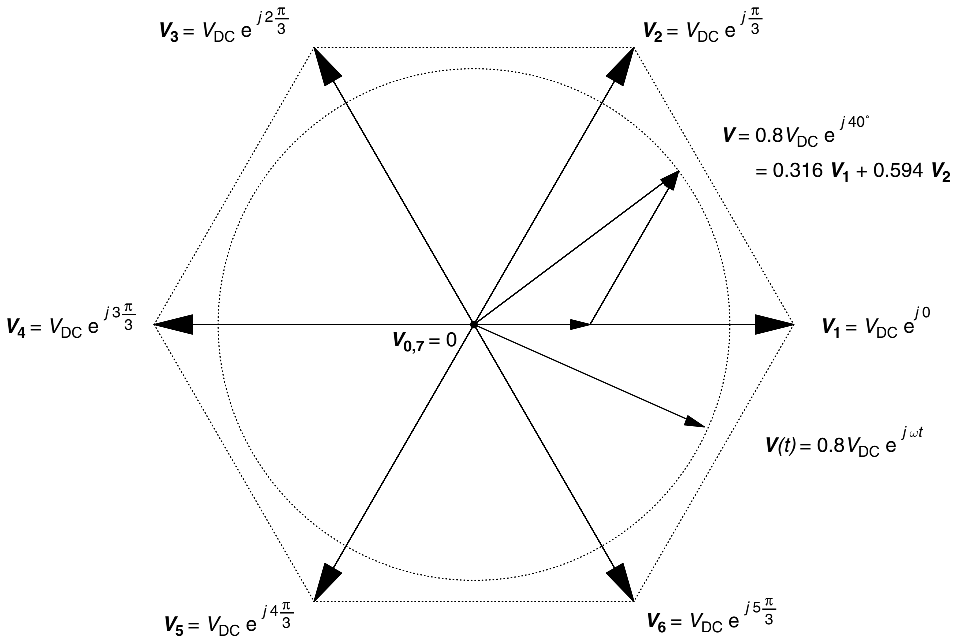

$V_3 = V_{DC}\,e^{j2\frac{\pi}{3}}$
$V_2 = V_{DC}\,e^{j\frac{\pi}{3}}$
$V = 0.8V_{DC}\,e^{j40^\circ}$
$= 0.316\ V_1 + 0.594\ V_2$
$V_4 = V_{DC}\,e^{j3\frac{\pi}{3}}$
$V_{0,7} = 0$
$V_1 = V_{DC}\,e^{j0}$
$V(t) = 0.8V_{DC}\,e^{j\omega t}$
$V_5 = V_{DC}\,e^{j4\frac{\pi}{3}}$
$V_6 = V_{DC}\,e^{j5\frac{\pi}{3}}$

Figure 18.35 Space-voltage vectors of a three-phase inverter and synthesis of a voltage trajectory by space-voltage vector modulation

$$V = V_{AN} + V_{BN}e^{j\frac{2\pi}{3}} + V_{CN}e^{-j\frac{2\pi}{3}}$$

or

$$V = V_{\alpha} + jV_{\beta}$$

$$\begin{bmatrix} V_{\alpha} \\ V_{\beta} \\ V_{\gamma} \end{bmatrix} = \begin{bmatrix} 1 & -\frac{1}{2} & -\frac{1}{2} \\ 0 & \frac{\sqrt{3}}{2} & -\frac{\sqrt{3}}{2} \\ \frac{1}{3} & \frac{1}{3} & \frac{1}{3} \end{bmatrix} \begin{bmatrix} V_A \\ V_B \\ V_C \end{bmatrix} \tag{18.34}$$

The normal objective is to produce a smoothly rotating space-voltage vector and, by implication, produce line voltages that are a balanced three-phase set. *Figure 18.35* shows the desired trajectory of the space-voltage vector. To achieve the point on the trajectory shown (a circle of radius 0.8 V_{DC} in this example) we combine the nearest two non-zero space-voltage vectors with the two zero space-voltage vectors.

For example, to achieve $V = 0.8\ V_{DC}\angle 40^\circ$ we use states 0, 1, 2, 7.

The duty-cycles of the two non-zero space-vectors must satisfy:

$$\delta_2 V_{DC}\sin(60^\circ) = 0.8 V_{DC}\sin(40^\circ)$$

$$\delta_1 V_{DC} + \delta_2 V_{DC}\cos(60^\circ) = 0.8 V_{DC}\cos(40^\circ)$$

Thus, in this case, $\delta_2 = 0.594$ and $\delta_1 = 0.316$. The remaining time within the period, $\delta_0 = (1 - \delta_1 - \delta_2)$ is split equally between the two zero-voltage states; states 0 and 7. A useful switching sequence, because it minimises the number of switch transitions and returns to the original state, is:

State 0 for $T\delta_0/4$
State 1 for $T\delta_1/2$
State 2 for $T\delta_2/2$
State 7 for $T\delta_0/2$
State 2 for $T\delta_2/2$
State 1 for $T\delta_1/2$
State 0 for $T\delta_0/4$

The sum of the duty-cycles of the two non-zero states is constrained to less than or equal to unity. This means that the synthesised space-voltage vector must lie within the hexagonal limit shown in *Figure 18.35*. The maximum radius circle that can be accommodated is $\frac{\sqrt{3}}{2}V_{DC}$ which is equivalent to the maximum amplitude sinewave that can be synthesised with simple PWM employing third-harmonic injection.

In high-power applications, particularly large machine drives, the switching frequency may be very low compared to the required sinewave to be synthesised. Special techniques are required involving synchronising the carrier to the sinewave or optimising individual switching times to reduce harmonic distortion. For very large inverters the devices might only be switched at line frequency in what is known as six-step or quasi-square operation. This is illustrated in Section 19.3.4.1 ‘Six-step square-wave inverter’.

18.3.3 Current source inverters

There are occasions on which a three-phase current set is required rather than a three-phase voltage set. A class of circuits, known as current source inverters,[6] exist which are duals of the circuits discussed in Section 18.3.2. The d.c.-side is supplied with a current via a smoothing inductor. The output can be pulse-width modulated or simply switched as a low frequency rectangular wave. Such circuits are sometimes used in large thyristor based drive systems as discussed in Section 19.3.4.2 ‘Current source inverters’. There are difficulties in designing a d.c.-side inductor with low resistance and high saturation limit. To achieve a fast-response controlled current it is normally better to use a voltage source inverter with a local current control-loop.

18.4 A.c./d.c. conversion

A.c./d.c. conversion is commonly referred to as rectification. There are well-established rectifier circuits based on diodes but these circuits suffer limitations. They have poor output voltage regulation characteristics and offer no adjustment of output voltage. They also generate harmonically distorted current in the a.c. system. Standards, such as IEC 1000 (adopted as EN 61000 in the EU), place limits on such distortion and lead to the adoption of more sophisticated circuits. The tightest limits imposed by these standards are on even-order harmonic current. This all but prohibits the use of half-wave circuits that draw current during only one half-cycle of the a.c. waveform.

18.4.1 Line-frequency-switched rectifiers

The single-phase full-wave diode bridge rectifier is shown in *Figure 18.36*. The diodes can be considered in pairs: one pair (D_A^T and D_B^B) applies the positive half-cycle of the a.c. to the load and the other (D_B^T and D_A^B) applies the negative half-cycle. The three-phase bridge rectifier of *Figure 18.37* is an extension of the single-phase bridge with three pairs of diodes. The line voltage that is highest in magnitude (whether positive or negative) is applied to the load. Each diode conducts for a third of a cycle and shares that conduction period with opposite side diodes from the two other phases.

The general form of load on the d.c.-side of a bridge rectifier is inductive-resistive impedance and a d.c. voltage. This is a reasonable representation of both the armature of a d.c. machine or large valued 'reservoir' capacitor (with either significant ESR and ESL or added inductance). *Figure 18.38* shows a single-phase example. The diodes of this circuit will begin to conduct at angle α when the magnitude of the a.c. voltage exceeds the voltage on the d.c.-side. Once in conduction the current is dictated by a first order differential equation.

$$L\frac{\mathrm{d}i}{\mathrm{d}t} = |v_{\mathrm{I}}| - iR - E \qquad (18.35)$$

The current rises while the a.c. voltage exceeds the sum of E and iR and falls thereafter. When the current reaches zero, the diodes fall out of conduction. The current will re-establish at $\pi + \alpha$ in the next half cycle. The extinction angle, β can be found from the 'equal area' criterion as shown in *Figure 18.38*. In other words, β can be found by solving:

$$\int_{\alpha}^{\beta} \frac{\hat{V}_{\mathrm{I}}\sin(\omega t) - iR - E}{L} \cdot \mathrm{d}t = 0 \qquad (18.36)$$

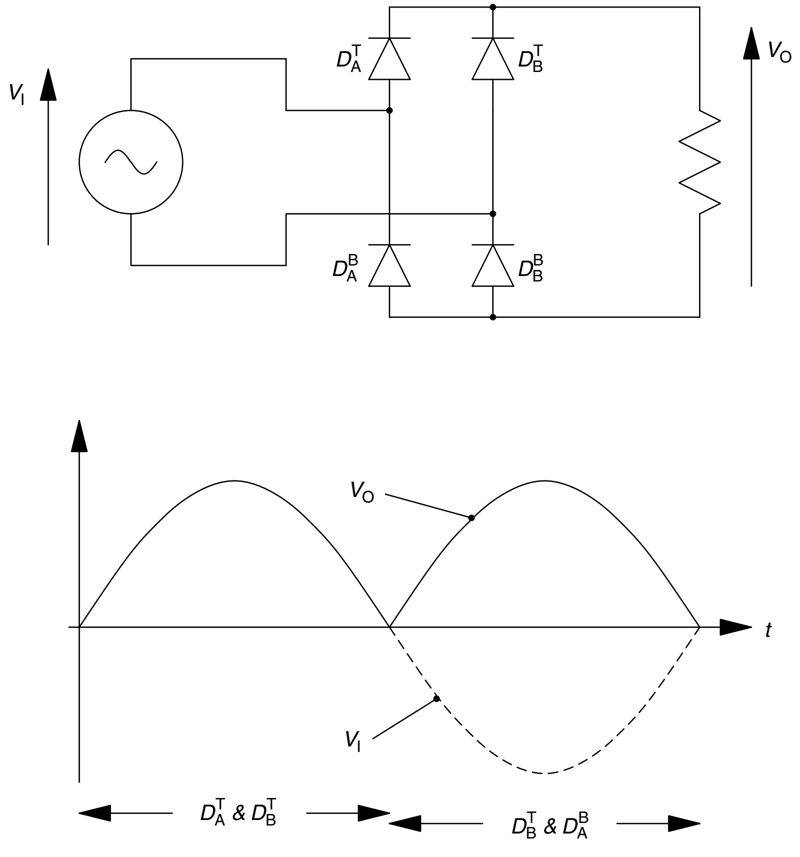


Figure 18.36 Single-phase rectifier with resistive load

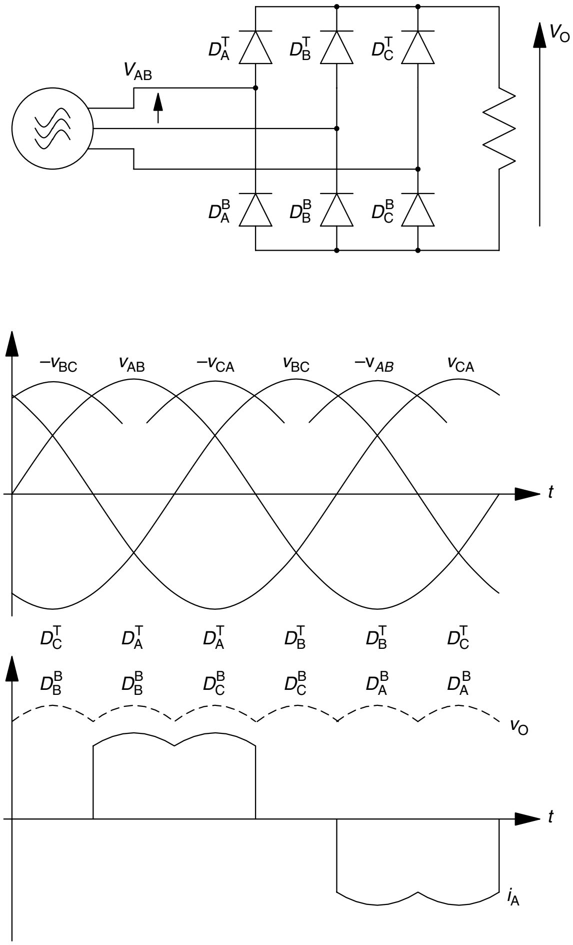


Figure 18.37 Three-phase rectifier with resistive load

The solution must be found numerically. Using $\beta\varsigma$we can also find the average voltage applied to the load:

$$V_{\mathrm{O(avg)}\Leftarrow} = \frac{\hat{V}_{\mathrm{I}}}{\pi}(\cancel{c}\mathrm{os}(\alpha) - \cos(\beta)) + E\left(1 - \frac{\beta - \alpha\varsigma}{\pi\varsigma}\right)\Bigg] \qquad (18.37)\Leftarrow$$

If the current becomes continuous then the first pair of diodes conducts from $\alpha\varsigma$to $\pi + \alpha$. The output voltage is found by replacing β with $\pi + \alpha$.

$$V_{\mathrm{O(avg)}\Leftarrow} = \frac{2\,\hat{V}_{\mathrm{I}}}{\pi\varsigma}\cos(\alpha) \ \Leftarrow \qquad (18.38)\Leftarrow$$

A variant of this circuit is obtained by substituting thyristors for the diodes, *Figure 18.39*. It is now possible to delay the onset of conduction, α_1 beyond the point α at which conduction is first possible.

$$\left.\begin{aligned} &\alpha = \sin^{-1}\left(\frac{E}{\hat{V}_{\mathrm{I}}}\right) \\ &\alpha \le \alpha_1 < \pi - \alpha\varsigma \end{aligned}\right] \qquad (18.39)\Leftarrow$$

The analysis of the diode circuit can be applied again and it is seen that the average voltage appearing across the load can now be varied.

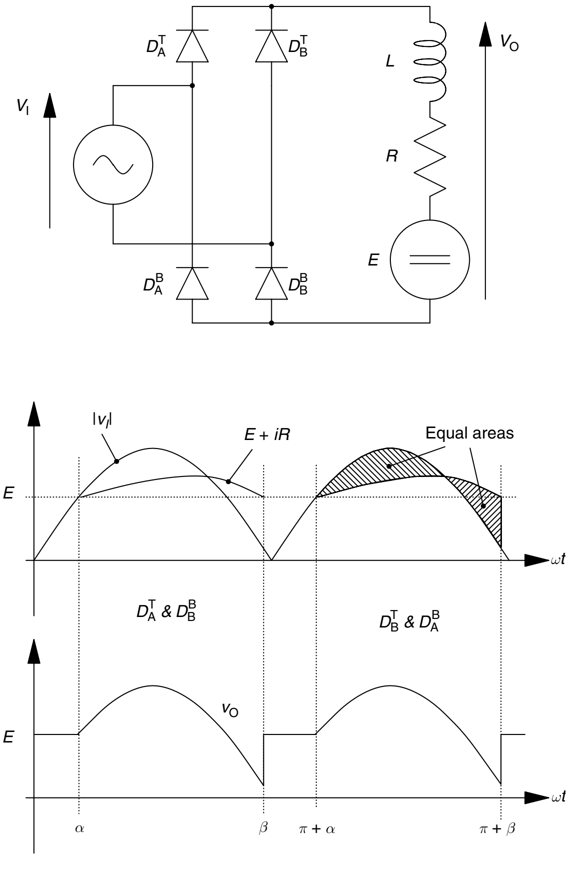


Figure 18.38 Single-phase rectifier with ERL load

For some loads, notably d.c. machines, it is possible to make E negative and it then becomes possible to transfer power from the d.c.-side to the a.c.-side. This can provide regenerative braking of a machine as discussed in Section 19.2.1 'D.c. Motor Drive Systems'. Initiation of conduction is delayed such that current flow is predominantly in the following half-cycle. The power on the d.c.-side is negative since the current is positive while the voltage E is negative. The power on the a.c.-side is also negative (on average) since, for instance, positive current flows during the negative half-cycle. The negative power indicates a reversal of flow from the assumed direction.

The three-phase versions of the bridge exist such as *Figure 18.40*. Many variations of the rectifier bridge exist based on full or partial replacement of diodes with thyristors. Such circuits are well known for use in high power drives as discussed in Section 19.3.1 'A.c. to d.c. Power Conversion'.

The circuit of *Figure 18.40* has an average output voltage (in continuous conduction) of:

$$V_{\mathrm{O(avg)}\Leftarrow} = \frac{3\hat{V}_{\mathrm{I}}}{\pi\varsigma}\cos(\alpha) \ \Leftarrow \qquad (18.40)\Leftarrow$$

where $\hat{V}_{\mathrm{I}}$ is the peak line voltage and α is measured from the cross-over of two line voltages.

These circuits have been popular choices for the supply of d.c. machines from a.c. sources because of the control possibilities they offer. Their use is waning because d.c. machines

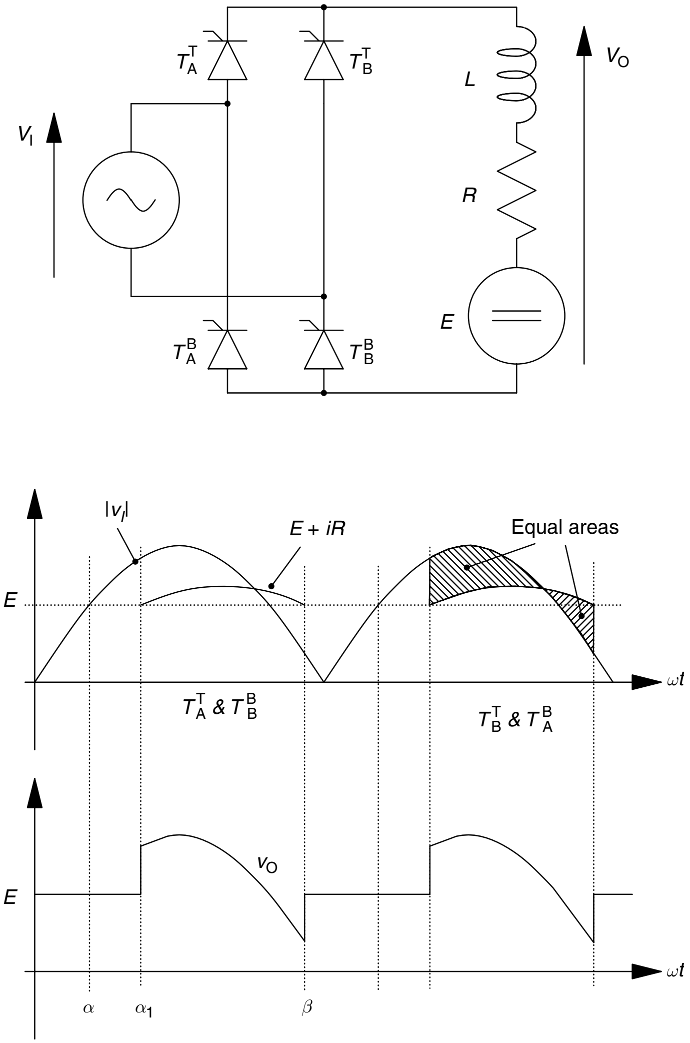

T_A^T
T_B^T
V_O
L
V_I
R
E
T_A^B
T_B^B
$|v_I|$
$E + iR$
Equal areas
E
ωt
T_A^T & T_B^B
T_B^T & T_A^B
v_O
E
ωt
α
α_1
β

Figure 18.39 Single-phase thyristor rectifier

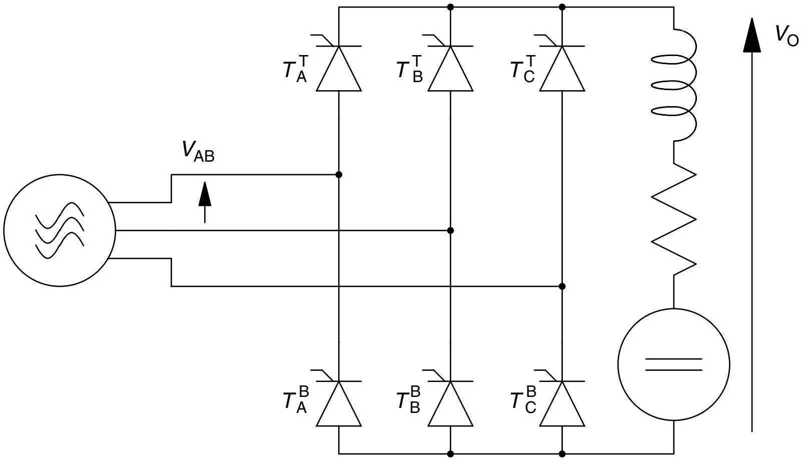

V_O
T_A^T
T_B^T
T_C^T
V_{AB}
T_A^B
T_B^B
T_C^B

Figure 18.40 Three-phase fully-controlled thyristor rectifier

are losing ground to a.c. machines. Even where d.c. machines are used, other power conversion circuits may be sought because the non-sinusoidal a.c.-side current of these line-frequency-switched rectifiers are problematic. Phase-angle controlled thyristor rectifiers will continue to be used at very high powers where thyristors are the only viable devices and the costs of harmonic mitigation can be met.

18.4.1.1 *Non-sinusoidal a.c.-side current*

Figure 18.41 shows the results of a circuit simulation of a rectifier like that of *Figure 18.40* operated such that the thyristors turn-on 18°⇐after the cross-over of the line voltages. The line voltages were 415 V at 50 Hz. The frequency spectrum shows that there are no even harmonics (as expected of a full wave rectifier with symmetric current wave-shape in the two half cycles). There are also no triplen harmonics (i.e. harmonics of order $3k$ with k any integer) as there is no path for these zero-sequence harmonics. The harmonics present are of order $6k \pm 1$ and this is characteristic of a six-pulse rectifier.

Table 18.2 compares the harmonic amplitudes found in this example circuit with the amplitudes allowed by the standard IEC 1000-3-3.

The fundamental amplitude is recorded as 15.3 A (peak) and so falls within the scope of class-A of IEC 1000-3-3, which applies up to 16 A (RMS). As expected, the largest amplitude emissions are the 5th, 7th and 11th harmonics and, in fact, the converter exceeds the Class-A limits at these points. Harmonic mitigation in the form of filtering or multi-pulse operation (Section 18.7.2 and Paice) may be sufficient to meet IEC 1000. If not, then a rectifier with active waveshape control (Section 18.4.2) will be needed.

Supply waveforms such as those in *Figure 18.41* need to be treated carefully when discussed in terms of the power factor. The original definition of power factor, as the ratio of real power to apparent power, must be used and any derived terms, based on sinusoidal conditions, must be ignored. A useful set of definitions follows if the a.c. voltage can be assumed to be perfectly sinusoidal even if the current is not. It then follows that only the in-phase fundamental frequency component of the current transfers real power. All harmonic currents cause oscillating instantaneous power flow that average to zero.

$$P = V_{\mathrm{RMS}} I_{1,\mathrm{RMS}} \cos(\theta) \Leftarrow \qquad (18.41)$$

where $I_{1,\mathrm{RMS}}$ is the RMS value of the fundamental frequency component of current and $\theta\varsigma$is the displacement angle between that component of current and the voltage.

This leads to the following definition of power factor:

$$\text{Power factor} = \text{Distortion factor} \times \text{Displacement factor} \qquad (18.42)\Leftarrow$$

$$\text{where}\quad \text{Distortion factor } \mu\varsigma\!=\!\Leftarrow\frac{I_{1,\mathrm{RMS}}}{I_{\mathrm{RMS}}}$$

$$\text{and}\quad \text{Displacement factor} = \cos(\theta)\Leftarrow$$

$$P = V_{\mathrm{RMS}}\ I_{\mathrm{RMS}}\mu\varsigma\cos(\theta) \qquad (18.43)\Leftarrow$$

18.4.2 Wave-shape controlled rectifiers

Just as switch-mode techniques are used to create sinewave supplies from d.c. sources, so they can be used to transform sinewave input currents into d.c. current. Various topologies of power converter exist but the one that has gained greatest acceptance in practice is derived from the Boost SMPS. *Figure 18.42* shows the d.c./d.c., 1-phase/d.c. and 3-phase/d.c. variants of the Boost converter. In each circuit, a switch (or switches) can be turned on to provide current path-A that imposes a positive voltage across the inductor and increases the current. Power is taken from the input and is stored in the inductor. The switch can be turned off and current path-B is established. A negative voltage is imposed across the inductor and the current reduces. Energy is transferred from both the input and the inductor to the capacitor, and the capacitor voltage rises.

Two control tasks exist; one to regulate the output voltage and the other to control the increase and decrease of current in order to form a sinewave. Circuits operated in this fashion are often referred to as power factor correctors. They produce a distortion factor (and displacement factor) close to unity in a rectifier that would otherwise have a low power factor.

Table 18.2 Comparison of simulated circuit with harmonic limits

Harmonic order	*Frequency* (Hz)	*Current amplitude* (A)	*IEC 1000 class-A* (A)	*Pass/fail*
1	50	15.330	—	—
2	100	0.013	1.080	Pass
3	150	0.014	2.300	Pass
4	200	0.009	0.043	Pass
5	250	9.472	1.140	Fail
6	300	0.011	0.300	Pass
7	350	5.350	0.770	Fail
8	400	0.010	0.230	Pass
9	450	0.017	0.400	Pass
10	500	0.010	0.180	Pass
11	550	0.454	0.300	Fail
. . .				
38	1900	0.008	0.048	Pass
39	1950	0.016	0.058	Pass
40	2000	0.007	0.046	Pass

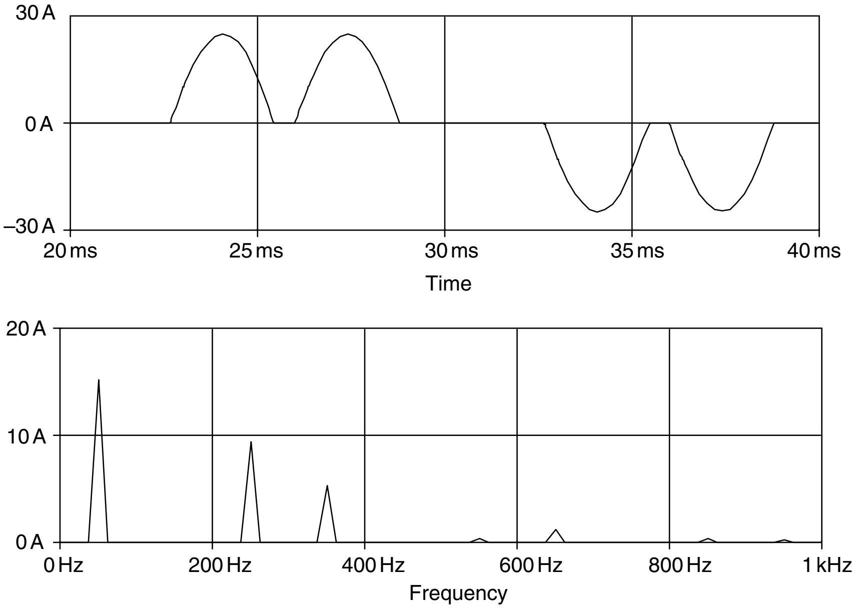


Figure 18.41 A.c.-side current of a thyristor bridge in time and frequency domain

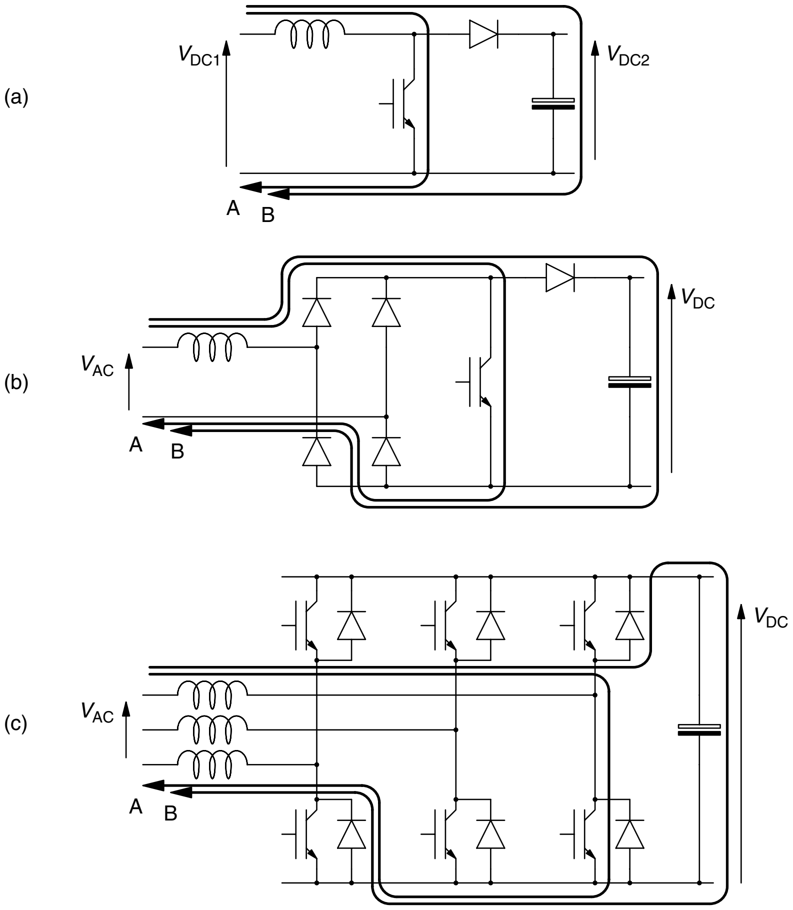

(a)
V_{DC1}
V_{DC2}
A
B
(b)
V_{AC}
V_{DC}
A
B
(c)
V_{AC}
V_{DC}
A
B

Figure 18.42 Boost SMPS in d.c./d.c., 1-phase/d.c. and 3-phase/d.c. forms

18.4.2.1 Single-phase

The Boost topology is advantageous because it places an inductor in series with input. Thus, the input current can be continuous and composed of a sinusoid with a small ripple superimposed. The step-up of voltage is not normally desired so a second stage step-down converter (such as a Buck or Flyback) is often added. Clearly, this adds expense and the double processing of power means it is also less efficient than a single stage converter[1]. However, the Boost circuit plus second-stage is considered the best solution where good a.c.-side waveform quality is required. The circuit of *Figure 18.43* is similar to that of *Figure 18.42*(*b*) except that the inductor has been relocated to allow all the required control signals to be referenced to the same potential.

Control of the converter is split into two sections. The first measures the output voltage error and sets a current demand that will correct it. The second stage takes the current demand, creates from it a sinusoidal current reference and modulates the switch to obtain such a current. The circuit is similar to a standard Boost SMPS but with a full-wave rectified input voltage. The duty-cycle is modulated to keep the output voltage constant while the input voltage varies. The sinusoidal reference shape is obtained by measuring the (full-wave rectified) instantaneous voltage.

It is inevitable that there is some distortion of the current immediately following the zero-crossing of the voltage because the input voltage is insufficient to increase the inductor current quickly enough to follow the sinusoidal reference. Careful design can reduce this distortion to the point where the harmonic content is small and standards such as IEC 1000 can be met with ease.

This has become a popular topology and several manufactures produce ICs that include these control functions. Two variants exist, as illustrated in *Figure 18.44*. The first uses a hysteresis controller for the current loop and results in fast response but a varying switching frequency. The second uses fixed-frequency PWM. The spread spectrum resulting from a varying switching frequency can make filter design difficult but it does have the advantage of spreading emissions thinly over a range rather than producing high level emissions concentrated at particular frequencies.

The control loop for the output voltage must be designed with a limited bandwidth typically around 5 Hz. The energy drawn from a single-phase supply will vary at twice line frequency even if sinusoidal current wave-shape is achieved. Thus, there will be an unavoidable output voltage ripple at twice line frequency which the control loop should not attempt to reject. The output capacitor is chosen to be large to keep this low-frequency ripple of the output-voltage small.

18.4.2.2 Three-phase

The three-phase boost topology, *Figure 18.46*, is very similar to a standard three-phase inverter, *Figure 18.32*. PWM of the inverter switches can be used to generate three voltages, V'_{A}⇐ V'_{B}, V'_{C}⇐similar in phase and magnitude to V_{A}, V_{B}, V_{C} once a voltage has been established on the d.c. side. (Pre-charge of the capacitor through the diodes of the circuit is problematic and needs some form of in-rush protection.)

Power flow through the Boost converter is controlled in the same manner as a standard synchronous machine. The phase difference (or load angle), δςcan be set by the PWM system and used to control the real power flow,

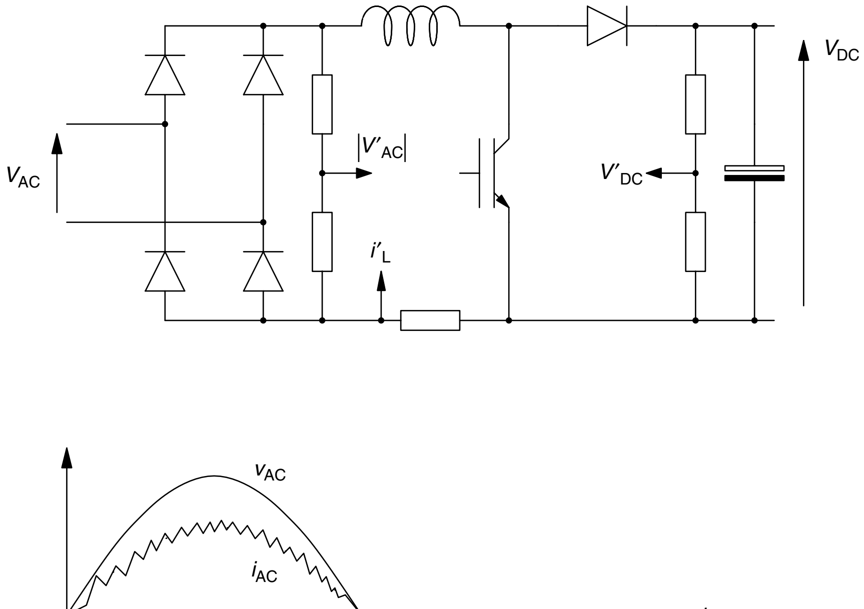


Figure 18.43 Single-phase Boost a.c./d.c. converter including its measurement points and example waveforms

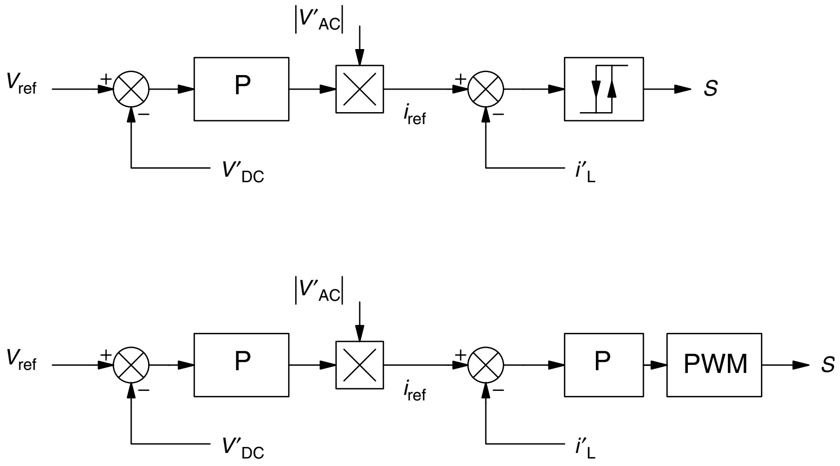


Figure 18.44 Hysteresis and PWM methods of current control

Figure 18.47. The voltage magnitude could also be used to control the reactive power flow. The Boost converter is able to operate in all four quadrants of the $S = P + jQ$ plane. This is useful for regenerative systems since power can be returned to the a.c. system. It is employed in some drive systems where sustained regeneration warrants the extra cost of this converter over a diode rectifier plus 'dump' resistor (Sections 19.3.4.1 'Voltage Source Inverters' and

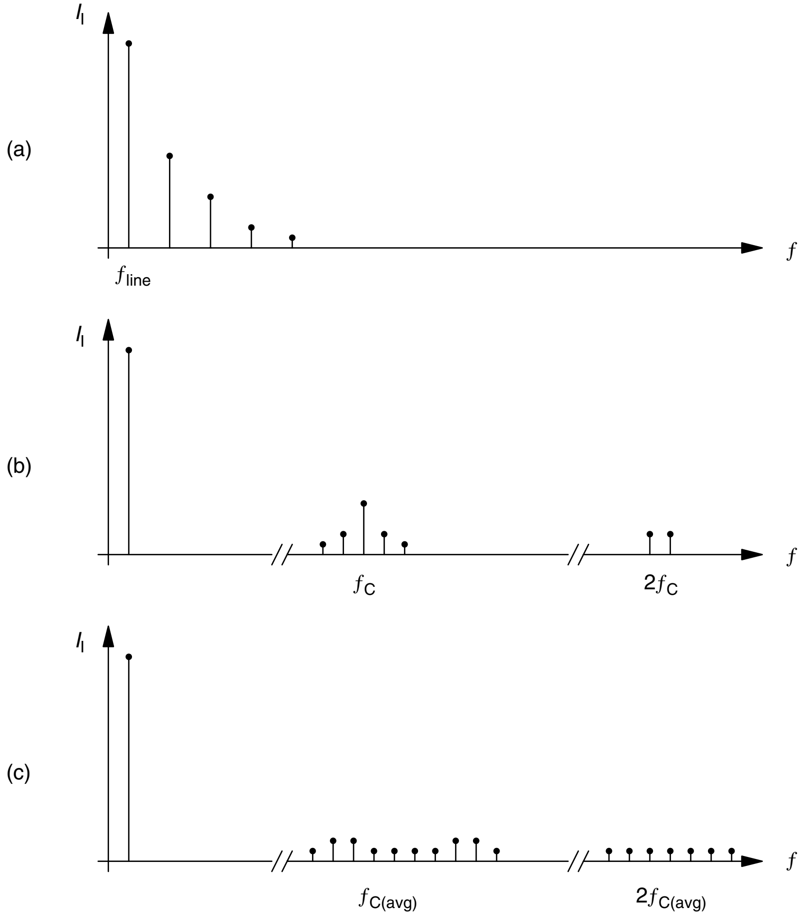


Figure 18.45 Frequency spectra of (a) diode rectifier, (b) PWM controlled a.c./d.c. converter and (c) hysteresis controlled a.c./d.c. converter

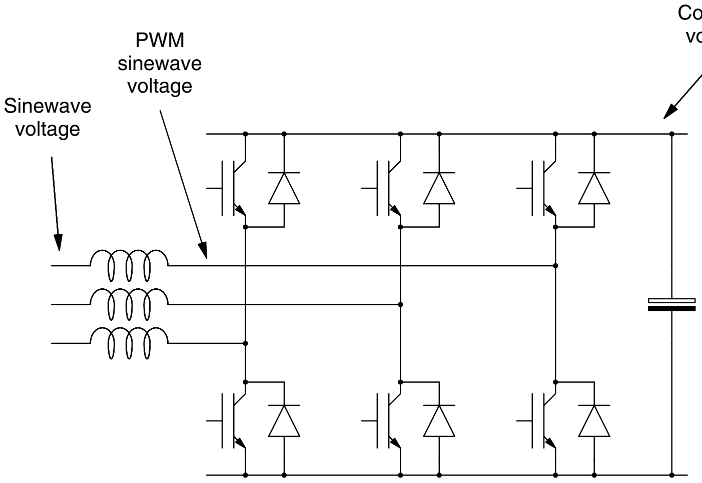

Constant voltage
PWM sinewave voltage
Sinewave voltage

Figure 18.46 Three-phase boost a.c./d.c. converter

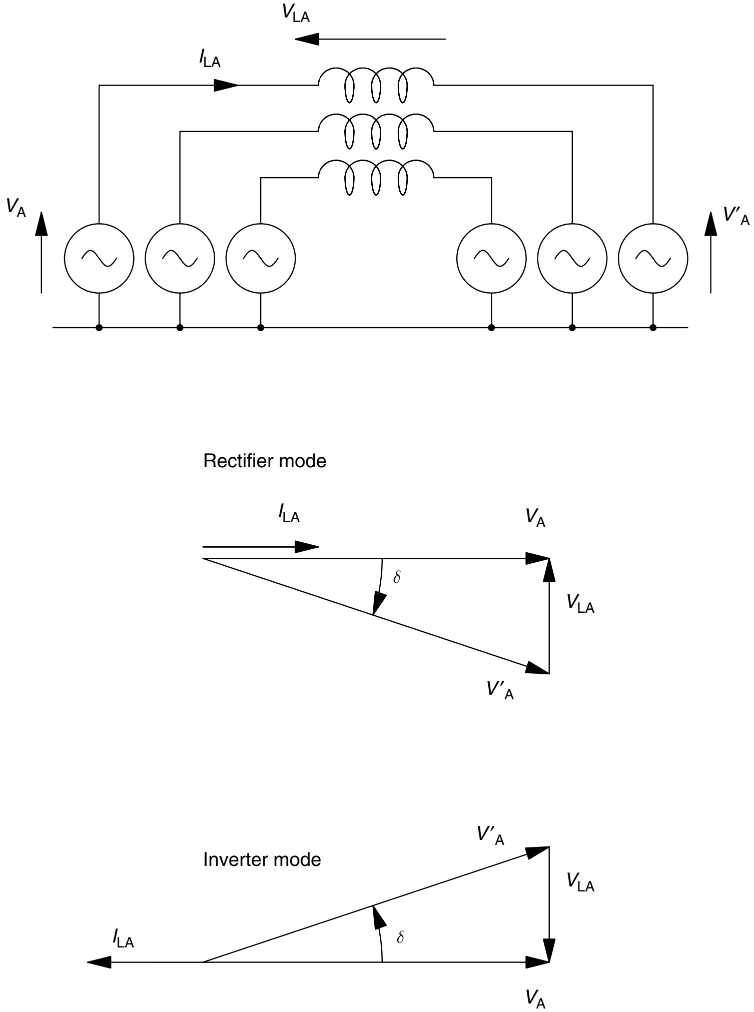

V_{LA}
I_{LA}
V_A
V'_A
Rectifier mode
I_{LA}
V_A
δ
V_{LA}
V'_A
V'_A
Inverter mode
V_{LA}
I_{LA}
δ
V_A

Figure 18.47 Bi-directional power flow controlled through phase-shift of voltage

19.4.2.4 ‘Four Quadrant Operation’). It is common to implement current control in this class of converter and separately control the in-phase (real power) and in-quadrature (reactive power) components of current. This might well be performed in the *dq*-domain following Clark and Park transforms of the variables.

18.5 A.c./a.c. conversion

There are circuits that will convert directly from one a.c. frequency and voltage to another a.c. frequency and voltage. These are known as cyclo-converters and matrix converters. However, they are not as attractive as might at first appear because they use a greater number of semiconductor switches than indirect methods of a.c./a.c. conversion.

18.5.1 A.c. voltage regulator

Where control of voltage amplitude only is required, then the techniques of phase-angle control described in Section 18.4.1 can be used. *Figure 18.48* shows a pair of thyristors in anti-parallel used as a phase-angle controlled switch to regulate the voltage applied to a single-phase load. Conduction of each thyristor can be initiated at any point in the appropriate half-cycle. With an inductive load, conduction will continue part way into the following half-cycle. In low power applications, such as incandescent lamp dimmers and a.c. commutator motors in domestic white-goods, a TRIAC would replace the pair of thyristors. Three-phase versions of this circuit are sometimes used as soft-start circuits for induction machines as discussed in Section 19.3.5.1 ‘Soft-Starter/Voltage-Regulator’.

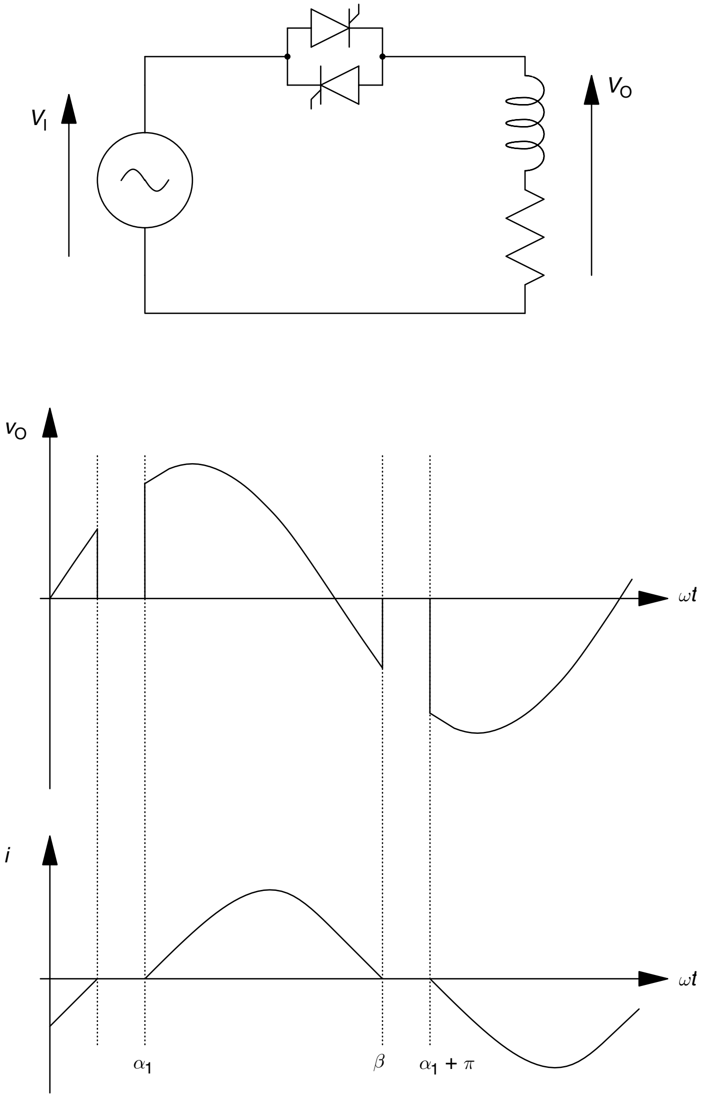


Figure 18.48 Phase-angle controlled a.c. voltage regulator

18.5.2 Direct frequency converter

18.5.2.1 Cyclo-converter

The cyclo-converter is able to provide direct a.c. to a.c. transformation with control over both the output voltage magnitude and output frequency. It uses phase-angle-controlled bridges to synthesise voltages. The most common form is the three-phase to three-phase converter of *Figure 18.49*. Each output requires two three-phase thyristor bridges of the same form as the rectifiers of *Figure 18.40*. One bridge is used to synthesise positive voltages at the output; the other is connected in reverse and synthesises negative voltages. The phase-angles of the bridges are gradually altered to give an approximately sinusoidal variation of the voltage at the output. Both the output voltage and the input current are rich in harmonics and some form of filtering is normally required. The frequency of the output voltage is constrained to less than about 1/3 of the input frequency by the allowable distortion of the input current. To prevent shoot-through paths between the phases while thyristors are commutating, the input phases are isolated using transformers. Cyclo-converters are employed for very high power, low speed machine drives where the complexity of the circuit is less of a problem and where line-frequency thyristors are the only viable device (Section 19.3.5.2 ‘Cycloconverter’). Cylco-converters are able to transfer power in either direction and can form the basis of a four-quadrant drive (c.f. the choppers of Section 18.2.1.1).

18.5.2.2 Matrix converter

The matrix converter performs a similar function to the cyclo-converter but employs semiconductors that can be commutated off (in contrast to thyristors). The switches are operated at high frequency to synthesise high-quality waveforms free of low-order harmonics. In its three-phase to three-phase form, *Figure 18.50*, the matrix converter requires 9 bi-directional switches to provide connections between every input phase and every output phase. There are no suitable intrinsically bi-directional semiconductor switches so each switch is made up of two uni-directional semiconductor switches. There are several possible ways to configure the uni-directional switches to achieve this. Like the cyclo-converter, the matrix converter is able to transfer power in either direction and can form the basis of a four-quadrant drive.

The circuit is normally placed between a voltage–source input and a current–sink (or inductive) output. It is imperative that there is always a path for each output current to flow and that there is never a short circuit path between two input connections. Given the finite switching times of devices this requires care. There are commutation sequences that meet these constraints provided that the two uni-directional halves of the switches are individually controlled in keeping with the prevailing voltages and currents.

18.5.3 Indirect frequency converter

We can convert a.c. to a.c. via d.c. using a pair of converters and a d.c.-link, *Figure 18.51*. The a.c. to d.c. converter will require 6 semiconductor switches and the d.c. to a.c. converter a further 6 switches. The total, 12 switches, is less than the 18 (uni-directional) switches required for the matrix converter or the 36 required for the cyclo-converter. However, the indirect conversion requires an energy storage component in the d.c.-link.

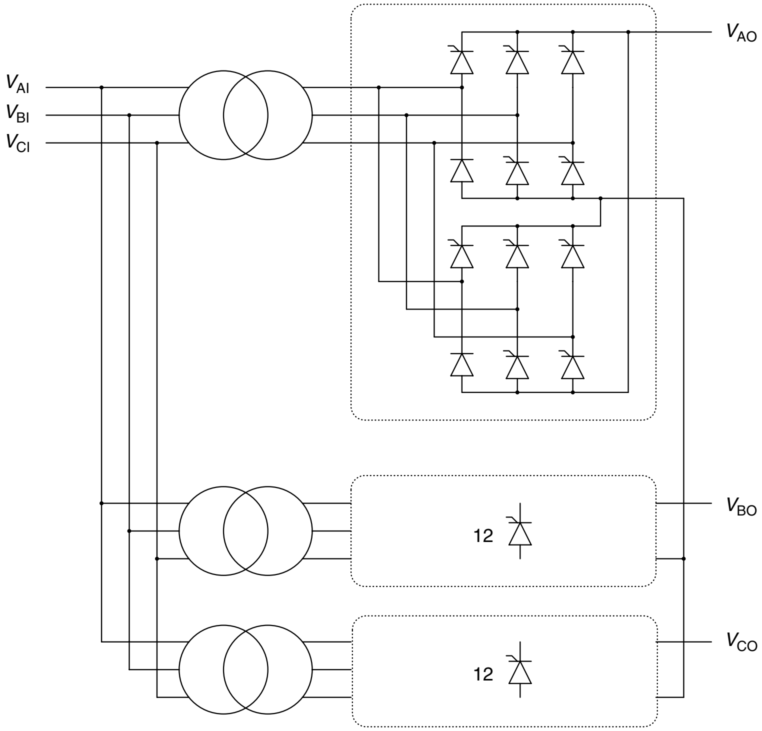


Figure 18.49 A three-phase to three-phase cyclo-converter

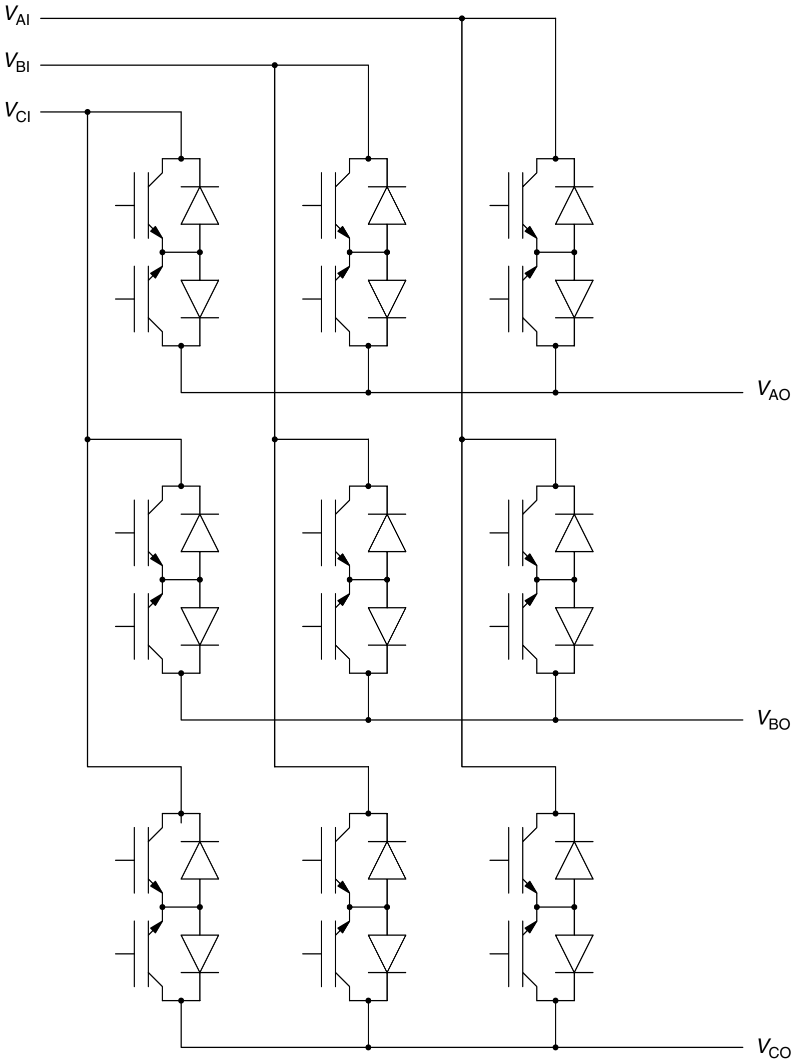


Figure 18.50 A three-phase to three-phase matrix converter

The inverter on the output-side can naturally transfer power in either direction. The rectifier on the input-side could be diode based on *Figure 18.37* and therefore capable of only uni-directional power flow. This would suit a two-quadrant drive system. If a four-quadrant drive system is required then a bi-directional rectifier based on *Figure 18.46* should be used. There is, of course, a cost penalty in using the more complex bi-directional circuit. If the drive system is re-generative for only brief periods (short periods of deceleration for instance) then a uni-directional rectifier and a resistive energy dump might be used as discussed in Section 19.4.2.4 'Four Quadrant Operation'.

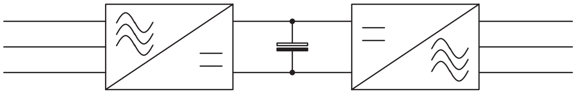


Figure 18.51 Indirect frequency conversion using a d.c.-link

The most common form of indirect conversion is using voltage source converters and a capacitive link. The d.c.-link capacitor is both a strength and a weakness of this circuit. It must exist to provide the short-term (switching cycle) energy storage to attenuate voltage ripple. In practice the capacitor is much larger than this and can provide sufficient energy storage to enable a system to operate through a momentary loss of supply. Its energy storage can also prevent low frequency variation of power flow on one side feeding through to the other (provided that the controller is so designed). This is useful for avoiding 'flicker' problems, i.e. voltage disturbances at a point of common coupling that would cause flicker of light output at frequencies (circa 12 Hz) that are irritating to the human eye.

The weakness of the d.c.-link capacitor is that it will normally be an electrolytic capacitor that has a relatively short lifetime. The lifetime is degraded at elevated temperatures and this can be a severe limitation in some applications.

18.6 Resonant techniques

Section 18.1.2 discussed the power dissipation in a semiconductor switch when a transition is made between on- and off-states. It was shown that in commutating the current in an inductive load, the switch is subjected to a coincidence of voltage drop and current flow and as a consequence the instantaneous power dissipation is high. Because there is a certain energy loss per commutation, the operating frequency (commutations per second) must be limited. However, there are ways to break this limit. The switching trajectory has to be modified so that either the current or the voltage or both are held at zero while commutation is performed. This is illustrated in *Figure 18.52*. The hard-switched trajectory shown is idealised. The trajectory will be modified, and energy loss increased, both by diode recovery at turn-on and voltage overshoot at turn-off.

There are two approaches to modifying the switching trajectory. One is to add a snubber circuit to the switch and control the rise of current at turn-on and rise of voltage at turn-off. Snubbers are discussed in Section 17.2.1 and several circuits are described.[6,3] The other approach is to modify the load circuit so that the load seen by the switch naturally has periods of zero voltage or zero current that present opportunities to turn on or off without incurring power dissipation. This second method covers techniques known as soft-switching and resonant power converters. As will be seen in this section, additional current flow is necessary to ensure that the correct switching conditions are achieved. Thus, there is a penalty to be paid in conduction losses for the reduction in switching losses.

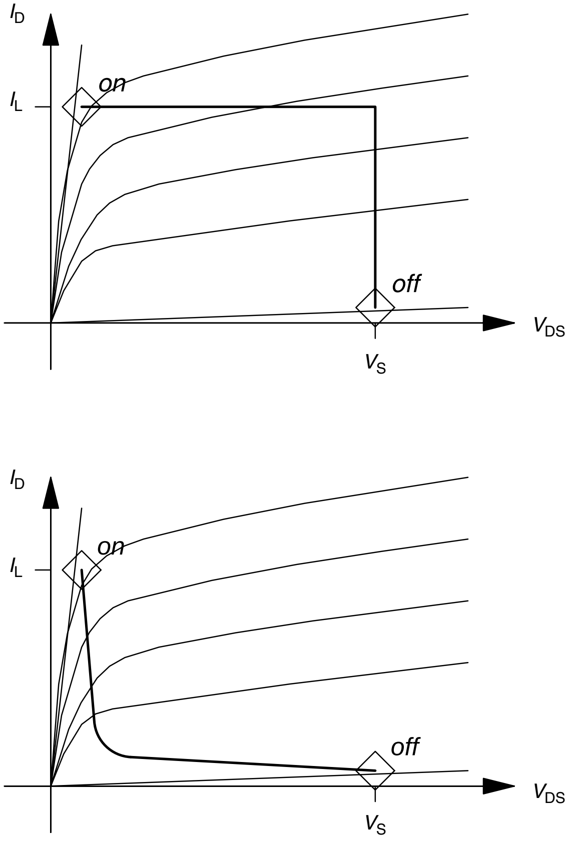


Figure 18.52 Hard-switched and soft-switched trajectories on the voltage/current plane

Two examples will be given here of the resonant circuit technique. The technique has been applied to almost every power converter circuit configuration.

18.6.1 Quasi-resonant SMPS

The term quasi-resonant describes a circuit that has one stable condition (one of either on- or off-states) and a resonant condition initiated when the switch changes state. The circuit is operated such that one cycle of resonant action occurs and then the circuit resumes its stable state.

18.6.1.1 Zero-current switched Buck SMPS

The circuit of *Figure 18.53* is a standard Buck SMPS with a resonant inductor–capacitor pair, $L_R C_R$, and an extra diode added. The circuit has a stable state when the transistor is off and the current in L_O flows through main diode D. For the purpose of the following analysis, the current in L_O is assumed to be constant and equal to the output current, I_O. A resonant cycle is initiated by turning the transistor on. As shown in *Figure 18.54*, this creates a pulse of voltage across C_R that is applied to the LC filter at the output. The duration of the pulse is fixed by the resonant frequency of $L_R C_R$, but the period between the pulses can be controlled by setting the off-time of the transistor. Thus the output voltage can be regulated.

The stages of operation are:

(I) The transistor is off and I_O flows in D. This state is stable.
(II) The transistor is turned on and the input voltage is imposed across L_R. As a consequence i_{LR} rises linearly. D stays in conduction (and v_{CR} is held close to zero) because i_{LR} is less than I_O. This period lasts until $i_{LR} = I_O$.
(III) D falls out of conduction and $L_R C_R$ form a resonant circuit governed by:

$$\begin{aligned} V_i &= v_{C_r} + L_r \frac{d}{dt}(i_{C_R} + I_O) \Leftarrow \\ &= v_{C_r} + L_r C_r \frac{d^2 v_{C_r}}{dt^2} \end{aligned} \qquad (18.44)\Leftarrow$$

with initial conditions of $i_{LR} = I_O$ and $v_{CR} = 0$. Solving this differential equation yields:

$$\begin{aligned} I_{L_R} &= I_O + \frac{V_I}{\omega_R L_R}\sin(\omega_R t) \Leftarrow \\ V_{C_R} &= V_I(1 - \cos(\omega_R t)) \Leftarrow \end{aligned} \qquad (18.45)\Leftarrow$$

where $\omega_R = \frac{1}{\sqrt{L_R C_R}}$

It is important that i_{LR} swings negative, period IIIb. When this happens, i_{LR} is carried by D_R and the transistor carries no current. The transistor can be switched off any time during period IIIb without incurring power dissipation. The circuit must be designed

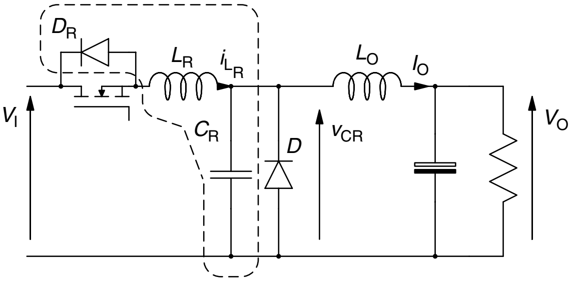


Figure 18.53 Zero-current switched Buck SMPS

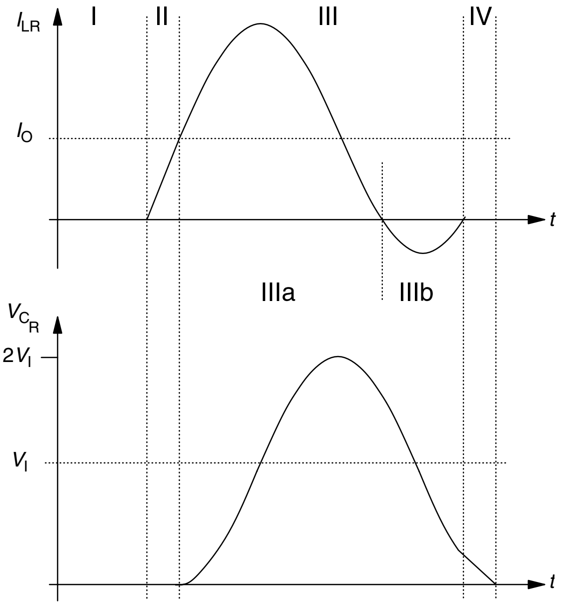


Figure 18.54 Waveforms of the resonant components of a ZCS QR Buck SMPS

such that $\frac{V_I}{\omega_R L_R}$ safely exceeds I_O so that the zero-voltage condition is ensured. This can lead to substantial extra current flow and raised conduction loss.

Period III ends when the i_{LR} rises again to zero but is blocked from becoming positive because the transistor is off.

(IV) When period III ends there is residual charge on C_R that is discharged (linearly) by the continued flow of I_O. This period ends when v_{CR} reaches zero and D is brought into conduction. The circuit then re-enters the stable state I.

If period IV is short then the average voltage during period III is V_I and the output voltage is given by:

$$\frac{V_O}{V_I} \approx \frac{t_R}{T} \approx \frac{2\pi f}{\omega_R} \qquad (18.46)\Leftarrow$$

18.6.1.2 Zero-voltage switched Buck SMPS

The zero-voltage switched Buck SMPS, *Figure 18.55* is the dual of the zero-current switched circuit. Its stable state is with the transistor on, and resonance is initiated by turning the transistor off. As shown in *Figure 18.56*, this produces a dip in the voltage applied to the output LC filter. The output voltage can be regulated by controlling the on-time of the switch.

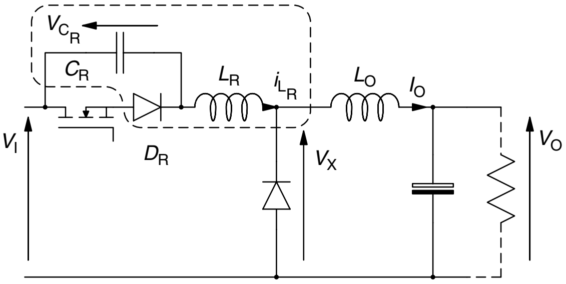


Figure 18.55 Zero-voltage switched Buck SMPS

The stages of operation are:

(I) The transistor is on and I_O flows through the transistors, D_R and L_R. This state is stable.

(II) The transistor is turned off and the current I_O $(= i_{LR})$ is diverted into C_R. C_R holds the voltage across the transistor low while it switches. The diversion of I_O causes v_{CR} to rises linearly. This continues until $v_{CR} = V_I$ at which point the main diode D becomes forward biased and provides an extra current path.

(III) With D in conduction, i_{LR} is free to change and $L_R C_R$ form a resonant circuit governed by:

$$V_i = v_{C_r} + L_r C_r \frac{d^2 v_{C_r}}{dt^2} \qquad (18.47)\Leftarrow$$

with initial conditions of $i_{LR} = I_O$ and $v_{CR} = V_I$.

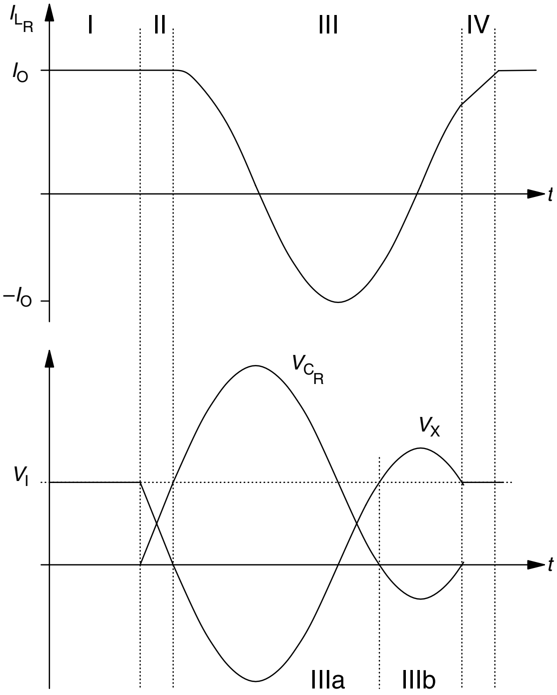


Figure 18.56 Waveforms of the resonant components of a ZVS QR Buck SMPS

Solving this differential equation yields:

$$I_{\mathrm{LR}} = I_{\mathrm{O}} \cos(\omega_{\mathrm{R}} t)$$
$$V_{C_{\mathrm{R}}} = V_{\mathrm{I}} + \frac{I_{\mathrm{O}}}{\omega_{\mathrm{R}} C_{\mathrm{R}}} \sin(\omega_{\mathrm{r}} t) \tag{18.48}$$

It is important that v_{CR} swings negative, period IIIb. When this happens, D_{R} is reverse biased and the transistor no longer has to support voltage. The transistor can be switched on at any time during period IIIb without incurring power dissipation. The circuit must be designed such that $\frac{I_{\mathrm{O}}}{\omega_{\mathrm{R}} C_{\mathrm{R}}}$ safely exceeds V_{I} so that the zero-voltage condition is ensured. Period III ends when the v_{LR} rises again to zero but is prevented from becoming positive because the transistor is on.

(IV) When period III ends, the current i_{LR} is less than I_{O}. L_{R} has the input voltage imposed across it (because D must still be in conduction to carry the short fall in I_{O}) and so the current rises. This period ends when i_{LR} reaches I_{O} and D is released from conduction. The circuit then re-enters the stable state I.

18.6.2 Resonant SMPS

The difference between quasi-resonant and resonant circuits is not distinct. The terms are used here to differentiate between circuits that are operated at or slightly off resonant frequency with essentially continuous resonant action and those of the previous section in which the resonant action is paused for a period.

Figure 18.57 is an example of a half-bridge used to excite a series combination of inductance and capacitance. A load, comprising a full-wave rectifier and reservoir capacitor, is connected in series with the resonant circuit. This is known as a series-loaded resonant converter.

The parallel-loaded variant (that is, a load connected in parallel with the capacitor) is shown in *Figure 18.58*. This circuit also incorporates transformer isolation.

There are various operating modes of these circuits depending on the ratio of the resonant frequency to the switching frequency of the switches. The modes are defined by whether more than one, less than one or less than a half of a resonant cycle is completed between each commutation of the switches.[3] The output of the circuit is controlled by varying the operating frequency. In the case of the series-loaded circuit this varies the current delivered to the output filter; in the case of the parallel-loaded circuit, this varies the voltage applied to the output filter.

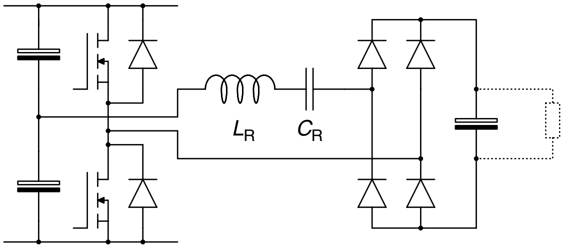


Figure 18.57 A series-loaded resonant SMPS

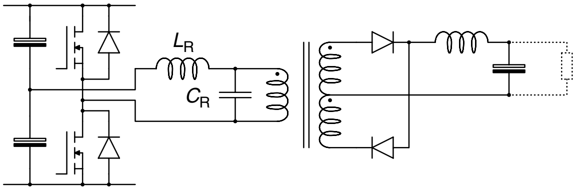


Figure 18.58 A transformer isolated, parallel-loaded resonant SMPS

An advantage of these circuits is that the parasitic capacitance of the semiconductors can be utilised in the normal operation of the circuit and may be supplemented with additional parallel capacitance. Thus, the voltages across the devices have limited rates-of-change.

Many forms of resonant SMPS exist. In addition, circuits without the full-wave rectifier are used to supply high frequency a.c. for applications such as induction heating furnaces.

18.7 Modular systems

For some applications power converters are built from modules. There are several reasons why this might be done.

First, device ratings are limited and for power converters of very high rating no single device can match the required rating. This is sometimes solved by series or parallel connection of devices to form composite switches (often known as valves in the power industry). An alternative is to form lower rated power converter modules from individual

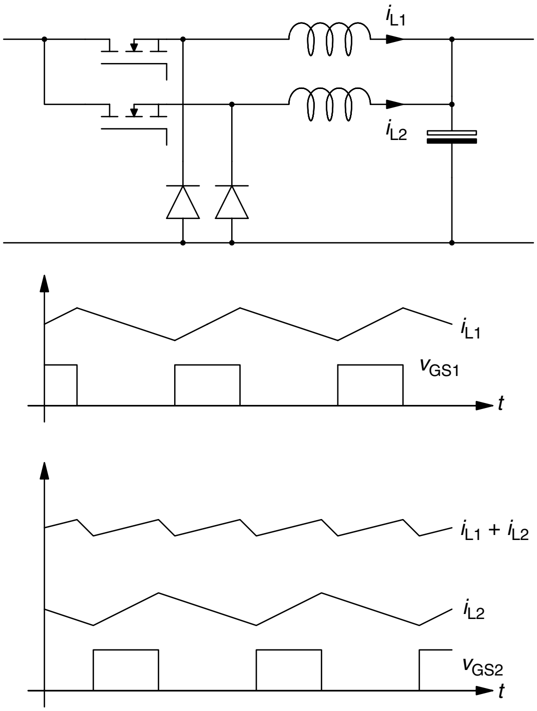


Figure 18.59 A pair of interleaved Buck SMPS showing a reduction of amplitude and increase in the effective frequency of the current ripple

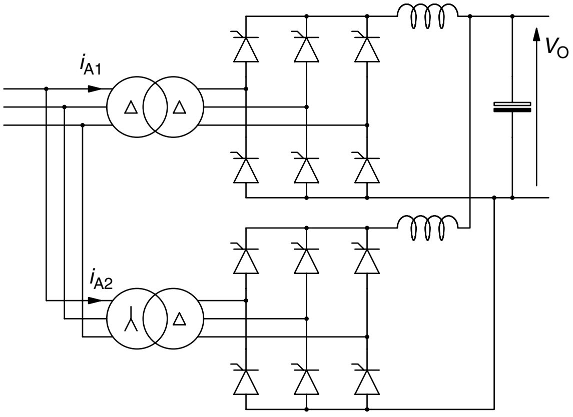


Figure 18.60 A 12-pulse rectifier

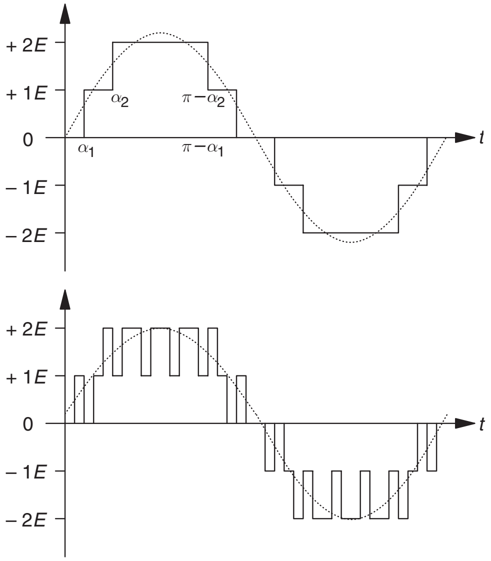
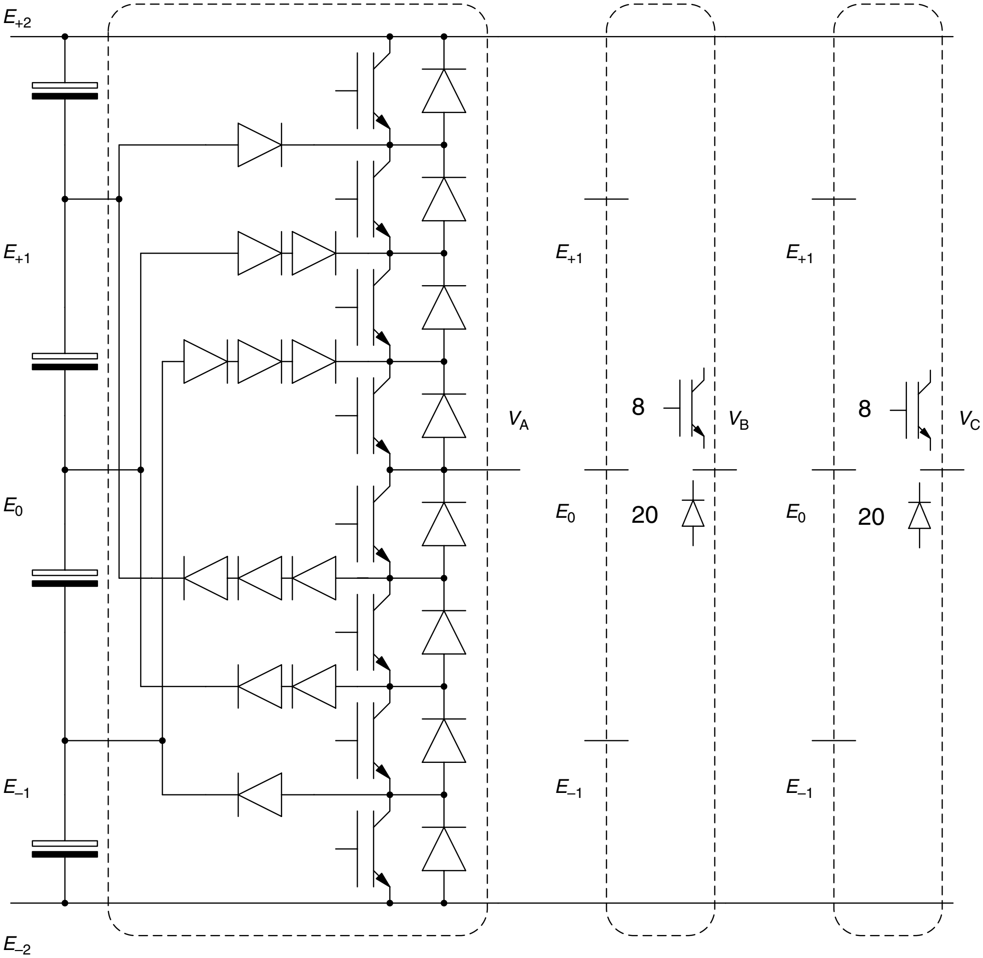


Figure 18.61 5-level staircase and PWM waveforms



E_{+2}
E_{+1}
E_0
E_{-1}
E_{-2}
V_A
V_B
V_C
8
20

Figure 18.62 A 5-level multi-point clamped inverter

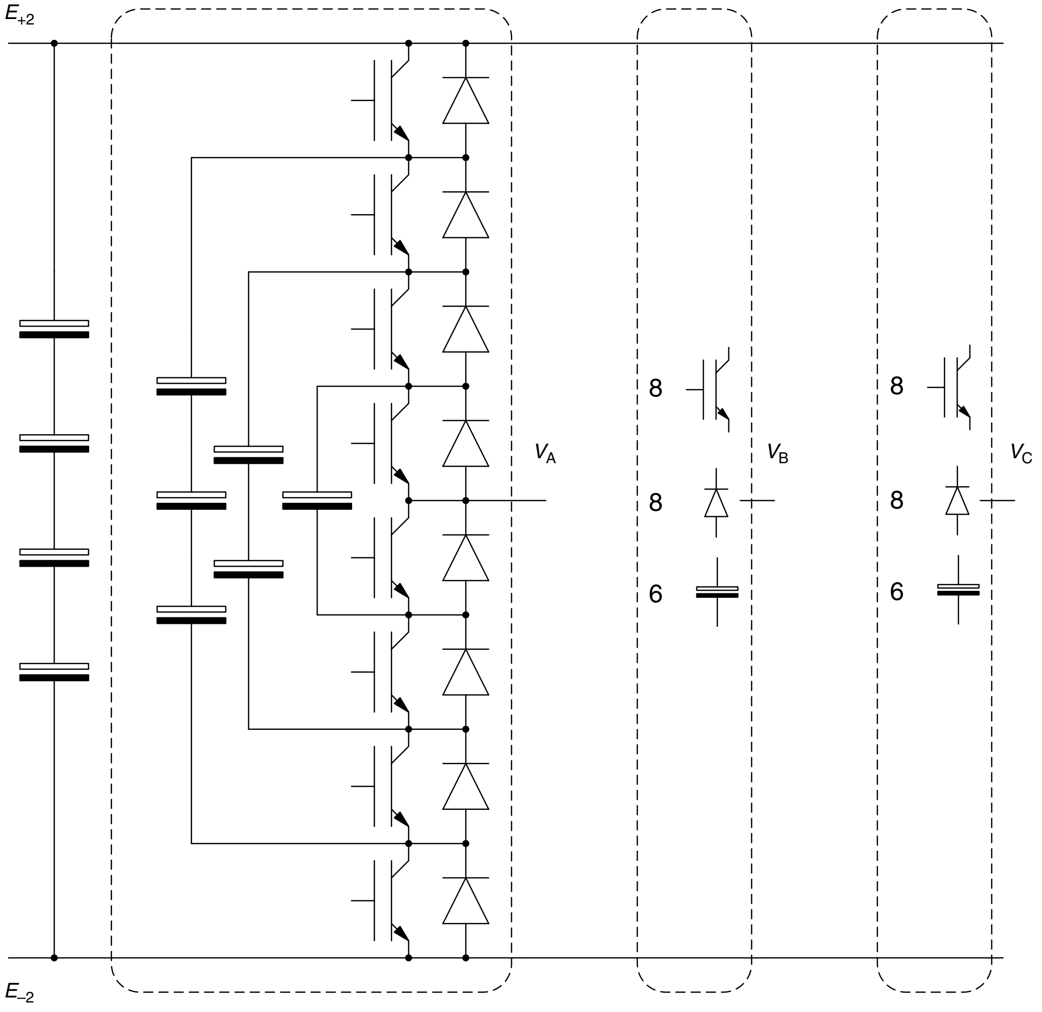


Figure 18.63 A 5-level nested cell inverter

switches and then connect the modules in series or parallel to achieve the desired rating.

Second, switching power loss limits the maximum frequency at which a power converter may operate. If several modules are operated with interleaved switching, then the overall circuit has an effective switching frequency equal to the switching frequency of each individual module multiplied by the number of modules. Each module is rated for a fraction of the voltage (series connection) or fraction of the current (parallel connection).

Third, in applications where reliability is a major concern, redundancy may be provided through modular construction with a greater number of modules than necessary to meet the power rating. The 'more-electric aircraft' and 'more-electric ship' are example application areas.

Some application areas combine all of three reasons. The STATCOM (a d.c./a.c. converter used for reactive power support) presently being commissioned by Alstom for National Grid PLC in the UK is a modular, chain-cell design. In this case:

(i) the voltage ratings are such that no single switch is suitable,
(ii) the switching frequency of high-rating GTO thyristors is low and good waveform quality can only be achieved through independent operation of multiple devices and
(iii) the availability requirement specified by the user is such that the equipment must continue to operate after the failure of one or possibly two devices and, therefore, each chain must contain redundant elements.

18.7.1 Interleaved SMPS

There is a strong desire in SMPS design to reduce output voltage ripple and input current ripple (especially in mains connected equipment subject to EMC constraints). *Figure 18.59* shows an example of parallel connected output stages with interleaved operation of the switches. It can be seen that the resultant current ripple is at twice the frequency of operation of the switches and that each switch has processed half of the current. Viewed in the frequency domain, we would see ripple current components at all of the even multiples of the switching frequency, whereas those at odd multiples created by one module will have been cancelled by those created (in anti-phase) by the other module. The same principle was used in the single-phase d.c./a.c. converter of Section 18.3.1 in which the two halves of the bridge were operated with phase-shifted carriers.

This idea can be extended to series connection of modules and to any number of modules. It can be applied to the input connection, output connection or both. In general, an n-module system has an effective switching rate (ripple frequency) of $n f_s$. In all cases it is important to match components and operating conditions in the modules in order to achieve the desired cancellation of ripple components.

18.7.2 Multi-pulse rectifiers

The diode and thyristor rectifiers of Section 18.4.1 were seen to create distorted a.c.-side currents with harmonic distortion of order $6k \pm 1$ for three-phase systems. This situation can be improved by operating rectifier modules with phase-shifts between them such that some of the harmonics created by one module are cancelled by those of another. These rectifiers are described by their pulse number. The standard rectifier creates a voltage (or current) waveform on the d.c.-side with 6 pulses per cycle. Adding a second unit with appropriate phase shift creates 12 pulses. Any number of 6-pulse modules can be combined to produce higher pulse number rectifiers. Combining n modules creates a $6n$-pulse rectifier that generates a.c.-side harmonics of order $6nk \pm 1$. That is, the first $(n-1)$ pairs of characteristic harmonics (and their multiples) are cancelled.

Figure 18.60 shows an example of a 12-pulse rectifier with parallel connection on the d.c.-side. One 6-pulse module is supplied through a delta–delta transformer (with no phase shift) and the second is supplied through a star–delta transformer providing 30° of phase advance. The thyristors of the second module are fired ahead of those of the first module by a time-shift equivalent to 30° of a mains cycle. This gives a 12-pulse pattern on the d.c.-side. The current drawn by the second module is also time-shifted forward but the fundamental is re-aligned with the first module by the 30° phase-shift of the transformer. Each harmonic is affected differently by the time-shift created by the thyristor firing and the phase-shift (which is phase-sequence dependent) of the transformer. For example, the 5th and 7th harmonics created by the two modules are in anti-phase and cancel whereas the 11th and 13th harmonics are in phase and add.

Zig-zag or differential delta transformers can be used to give the small phase-shifts necessary to construct high pulse-number rectifiers (n increments of $60°/n$ for a $6n$-pulse system).[5]

18.7.3 Multi-level inverters

The multi-pulse techniques of Section 18.7.2 have also been used for inverters and reactive power compensators. However, the complexity and bulk of the transformer system is a considerable disadvantage. Instead, inverters can be constructed in which the output voltages can be synthesised by switching between many voltage levels rather than the two levels of the circuit in *Figure 18.32*. In high-power inverters, the switches are operated at line frequency and the waveform synthesised as a staircase, *Figure 18.61*. In medium-power inverters, PWM or SVM can be used to operate the switches at high frequency. In the line-frequency-switched case, the angle of each transition can be optimised to meet harmonic distortion criteria.

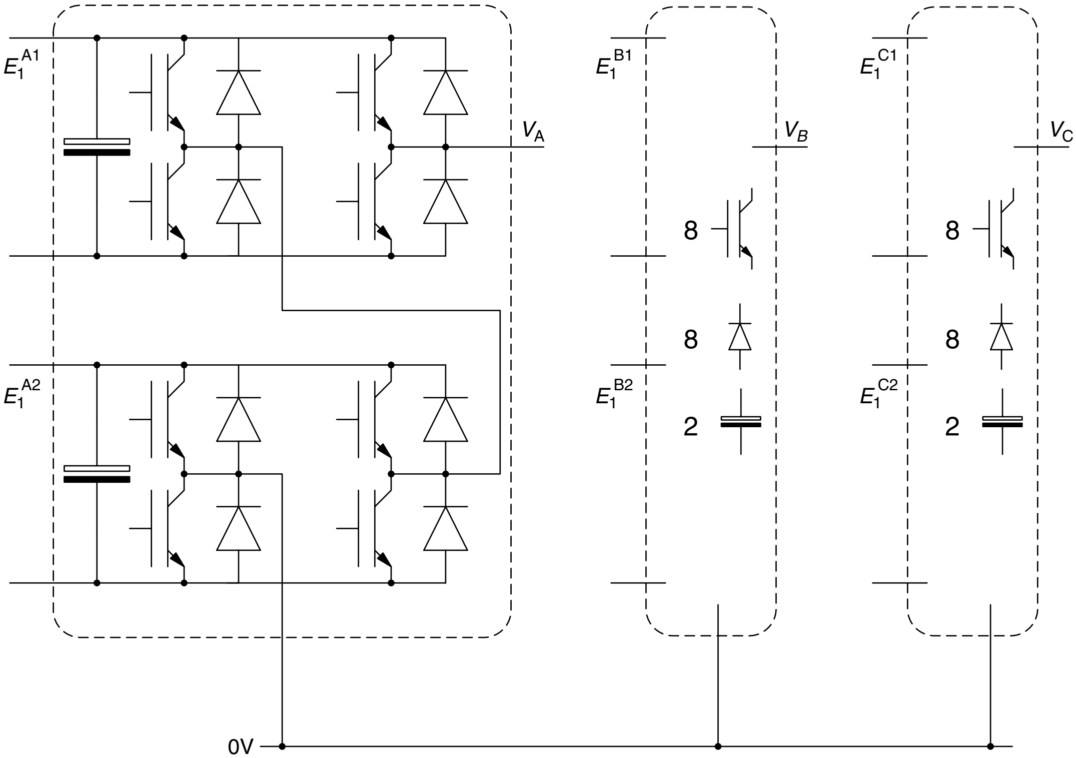


Figure 18.64 A 5-level chain cell inverter

There are several types of circuit for implementing multi-level inverters. Three examples of 5-level inverters are shown in *Figure 18.62*, *Figure 18.63* and *Figure 18.64*. Note that each circuit uses 8 transistors per phase and each transistor has a voltage rating of one level and a current rating of the full phase current. The differences between the circuits lie in how the numbers of other components (diodes and capacitors) increase with the number of levels.

The multi-point clamped circuit, *Figure 18.62*, has the advantage that the capacitor chain is shared between the three phases. This reduces the ripple in the capacitors (for balanced conditions) and allows smaller volume capacitors to be used than with the other implementations. Its principal disadvantage is the number of clamping diodes required (and that some of these must be series connections to span multiple voltage levels). Charge drawn from each capacitor must be managed to ensure that the capacitors remain balanced. To some extent this can be achieved in SVM by utilising redundant states. This cannot be achieved under all conditions (where suitable states do not exist) and auxiliary balancing circuits are required.

The nested-cell circuit, *Figure 18.63*, (also known as the flying capacitor inverter) avoids high voltage clamp diodes but requires capacitors to set the voltage levels. Operation of the circuit must be managed to ensure that the correct voltage is maintained on each capacitor. Only one of the capacitor chains is common to all phases. The volume of capacitance required is greater than for the other designs.

The chain-cell converter, *Figure 18.64*, avoids both a larger number of diodes or a large volume of capacitance. It is the only circuit in which the component numbers increase linearly with the number of modules. However, each module requires an isolated d.c.-link. This makes the circuit difficult to apply where a.c./d.c./a.c. conversion is required, for instance, in a drive system. The chain-cell inverter is suitable for floating (i.e. supply-less) applications. It has been used for reactive power compensation as mentioned in the introduction to this section.

18.8 Further reading

This chapter has given a flavour of the opportunities afforded by power electronics in power conversion. It has also introduced some of the analysis necessary for design. However, there are several topics that need further treatment before practical design could be attempted.

In all but the simplest circuits, the semiconductor devices need to be operated in the correct environment. Texts such as[6,3] give details of the gate drive circuits (including electrical isolation of signal and power), protection against over-voltage and over-current, snubber circuits and thermal management. Power electronic circuits put unusual voltage and current stresses on passive components too and so these components need to be chosen with care.

As indicated in the main text, there is a huge variety of SMPS circuits and almost any combination of voltage transfer characteristics, isolation and resonant action can be found.[2,3,6] There is a similarly large variety of single-phase a.c./d.c. converter circuits and a good review of these has been provided by[1]. This review also discusses the usefulness of circuits that fall short of perfectly sinusoidal current but still meet distortion standards. Control of SMPS is a large subject in itself.[3]

Line-frequency-switched converters can cause serious distortion to a.c. systems. The severity of the distortion needs to be analysed and harmonic mitigation will often be required.[5]

Control of power converters with a 3-phase a.c. connection is often best achieved in a rotating co-ordinate system in which the sinusoidal variation of the signals is removed. This is discussed in several texts in terms of control of a.c. machines, for instance.[4]

References

1 Garcia, O., Cobos, J. A., Prieto, R., Alou, P. and Uceda, J., *Power Factor Correction: A Survey*, IEEE Power Electronics Specialist Conference (PESC2001), Vancouver, Canada, June (2001)
2 Kassakian, J. G., Schlecht, M. F. and Verghese, G. C., *Principles of Power Electronics*, Addison Wesley, New York (1991)
3 Mohan, N., Undeland, T. and Robbins, W. P., *Power Electronics: Converters, Applications and Design*, Wiley, New York (1995)
4 Novotny, D. W. and Lipo, T. A., *Vector Control and Dynamics of A.C. Drives*, Oxford Science Publications, Oxford (1996)
5 Paice, D. A., *Power Converter Harmonics*, IEEE Press, New York (1995)
6 Williams, B.W., *Power Electronics: Devices, Drivers, Applications and Passive Components*, 2nd edition, McMillan, London (1995)

19

Electrical Machine Drives

W Drury BSc, PhD, CEng, FIEE
Control Techniques

Contents

19.1 Introduction 19/3

19.2 Fundamental control requirements for electrical machines 19/3
  19.2.1 D.c. motor control 19/3
  19.2.2 A.c. induction motor control 19/4
  19.2.3 A.c. synchronous motors 19/7
  19.2.4 Brushless servomotors 19/7
  19.2.5 Reluctance motors 19/8
  19.2.6 Switched reluctance motors 19/8

19.3 Drive power circuits 19/9
  19.3.1 A.c. to d.c. power conversion 19/9
  19.3.2 D.c. motor drive systems 19/9
  19.3.3 D.c. to d.c. power conversion 19/12
  19.3.4 A.c. to a.c. power converters with intermediate d.c. link 19/16
  19.3.5 Direct a.c. to a.c. power converters 19/21

19.4 Drive control 19/22
  19.4.1 D.c. drive control 19/22
  19.4.2 A.c. drive control 19/23

19.5 Applications and drive selection 19/28
  19.5.1 General 19/28

19.6 Electromagnetic compatibility 19/33

19.1 Introduction

It is now generally considered that 25–50% of all electrical power passes through semiconductor conversion. The importance of this technology is therefore self evident though the selection of appropriate equipment is often less clear. It would be a difficult, if not impossible, task to detail every power conversion circuit available, concentration will be given to those of greatest practical importance of the broad base of industry, particularly in the area of electrical variable speed drives.

In some respects semiconductor based power converters perform rather badly the functions of voltage and frequency changing which rotating machine based converters perform rather well. Economic and certain other factors are, in other than a few specialised applications, now so heavily in favour of electronic converters that only their performance will be considered.

Control philosophy and choice of semiconductor switching device are important aspects of any system but are sadly (from the standpoint of review) heavily dependent upon manufacturer. It is therefore useful to first consider the basic power conversion circuits, together with their inherent characteristics and then look at the available control strategies, their features and how they apply and impact upon converter systems. Basic power conversion circuits are considered in the Power Electronics chapter, however, it is helpful to consider salient circuits and characteristics here in the specific context of drives.

The existence of so many power conversion technologies reflects the diverse requirements industry demands. No single technology is the ‘optimum’ for all applications (despite some manufacturers’ claims). Good drive selection demands a good understanding of the application in regard torque/speed and dynamic characteristics together with the site environment. Simple guidelines are given.

Aspects of electrical variable speed drives vital for trouble free installation and ease of operation are also discussed. These include supply distortion/harmonics and communication systems the latter of growing importance for drives employed in an automated environment.

19.2 Fundamental control requirements for electrical machines

19.2.1 D.c. motor control

History will recognise the vital role played by d.c. motors in the development of industrial power transmission systems. The d.c. machine was the first practical device to convert electrical power into mechanical power, and vice versa in its generator form. Inherently straightforward operating characteristics, flexible performance and high efficiency encouraged the widespread use of d.c. motors in many types of industrial drive application.

The majority of standard d.c. motors, both wound-field and permanent-magnet, are now designed specifically to take advantage of rectified a.c. power supplies. Square, fully laminated frame construction allows minimal shaft centre height for a given power rating, and affords reduced magnetic losses, which in turn greatly improves commutating ability.

Over the last few years the use of permanent magnet motors, usually in the fractional to 3 kW range, has become commonplace in general-purpose drive applications. In this design the conventional wound field is replaced by permanent magnets bonded into the motor frame. The magnets have a curved face to offer a constant air gap to the conventional armature.

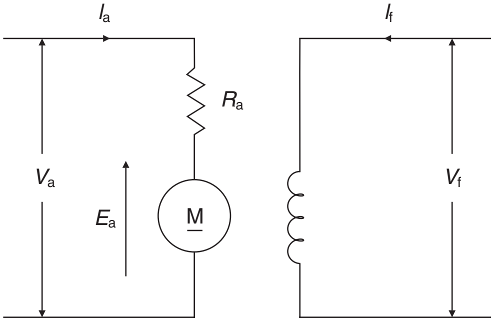


Figure 19.1 Shunt wound d.c. motor

The circuit of a shunt-wound d.c. motor, *Figure 19.1*, shows the armature, armature resistance (R_a) and field winding. The armature supply voltage V_a is supplied typically from a controlled thyristor system and the field supply voltage V_f from a separate bridge rectifier.

As the armature rotates an EMF E_a is induced in the armature circuit and is called the back-EMF since it opposes the applied voltage V_a and the flow of current produced by V_a. This back-EMF E_a is related to armature speed and main field flux by:

$$E_a = k_1 n \Phi \qquad (19.1)\Leftarrow$$

where n = speed of rotation
Φ = field flux
k_1 = motor constant

Also the applied, or terminal armature voltage V_a is given by:

$$V_a = E_a + I_a \cdot R_a \qquad (19.2)\Leftarrow$$

where V_a = applied armature voltage
I_a = armature current
R_a = armature resistance

Multiplying each side of Equation (19.2) by I_a gives:

$$V_a \cdot I_a = E_a \cdot I_a + I_a^2 \cdot R_a \qquad (19.3)$$

Total power supplied = Power output + Armature losses

Interaction of the field flux and armature flux produces an armature torque. Thus:

$$\text{Torque } M = \mathrm{k}_2 \cdot I_f \cdot I_a \qquad (19.4)\Leftarrow$$

where k_2 = constant
I_f = main field current
I_a = armature current

This confirms the straightforward and linear characteristic of the d.c. motor and consideration of these simple equations will show its controllability and inherent stability.

The speed characteristic of a motor is generally represented by curves of speed against input current or torque and its shape can be derived from Equations (19.1) and (19.2):

$$k_1 n \Phi = V_a - (I_a \cdot R_a)\Leftarrow$$

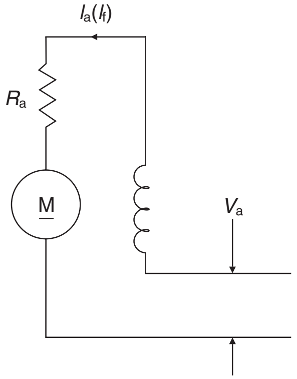


Figure 19.2 Schematic of series d.c. motor

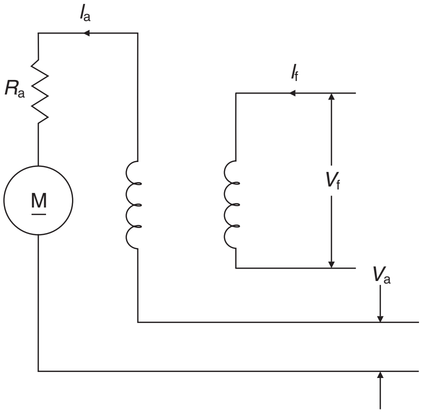


Figure 19.3 Compound d.c. motor

If the flux is held constant, which is achieved by simply holding the field current constant in a properly compensated motor then:

$$n = k_2[V_a - (I_a \cdot R_a)]$$

The circuits of Shunt wound and Series d.c. motors are shown in *Figures 19.1* and *19.2* respectively.

With the Shunt motor the field flux Φ is only slightly affected by armature current, and the value of $I_a R_a$ at full load rarely exceeds 5% of V_a, giving a torque–speed curve shown typically in *Figure 19.4(a)*, where speed remains sensibly constant over a wide range of load torque.

The compound-wound d.c. machine combines the shunt and series characteristics. The circuits of the compound d.c. motors is shown in *Figure 19.3*.

The exact shape of the torque/speed characteristic is determined by the resistance values of the shunt and series fields. The slightly drooping characteristic, *Figure 19.4(b)*, has the advantage in many applications of reducing the mechanical effects of shock loading.

The series motor curve, *Figure 19.4(c)*, shows initial flux increase in proportion to current, falling away due to magnetic saturation. In addition the armature circuit includes the resistance of the field winding and the speed becomes roughly inversely proportional to the current. If the load falls to a low value the speed increases dramatically, which may be hazardous: the series motor should not normally be used where there is a possibility of load loss; but because it produces high values of torque at low speed and its characteristic is falling speed with load increase, it is useful in applications such as traction and hoisting or some mixing duties where initial friction is dominant.

The power speed limit for d.c. machine manufacture is approximately 3×10^6 kW rev/min. This limit together with the maintenance attributable to d.c. motors can be blamed on the otherwise commendable commutator.

Under semiconductor converter control, with speed feedback from a tachogenerator, the shape of the speed/load curve is largely determined within the controller. It has become standard to use a plain shunt d.c. motor on the basis of reduced cost, even though the speed/load curve on open loop control is often slightly rising.

19.2.2 A.c. induction motor control

The a.c. squirrel cage induction motor is the basic, universal workhorse of industry, converting some 70–80% of all electrical power into mechanical energy. This type of motor nontheless exhibits some quite unattractive performance characteristics in spite of intensive development, notably instability and a non-linear load–current characteristic.

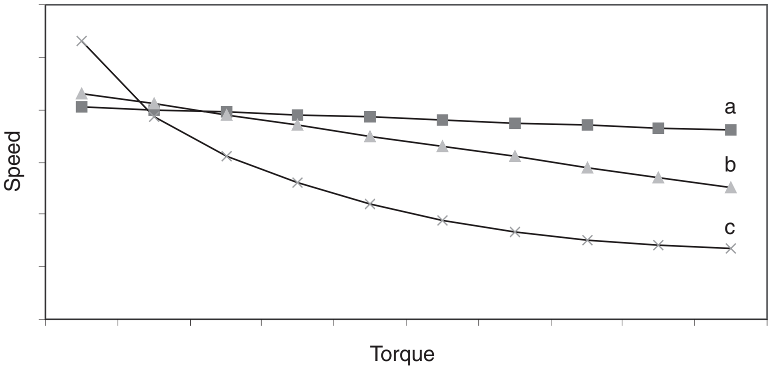


Figure 19.4 Torque/speed characteristic: (a) shunt wound d.c. motor; (b) compound d.c. motor; (c) series d.c. motor

It is invariably designed for fixed speed operation, larger ratings having such features as deep rotor bars to limit Direct on Line (DOL) starting currents. Electronic variable speed drive technology is able to provide the necessary variable voltage/current, variable frequency supply that the 3-phase a.c. machine requires for efficient, dynamic and stable variable speed control.

Modern electronic control technology is able not only to render the a.c. induction motor satisfactory for many modern drive applications but also to greatly extend its application and enable advantage to be taken of its low capital and maintenance costs. More striking still, microelectronic developments have made possible the highly dynamic operation of induction motors by the application of flux vector control. The practical effect is that it is now possible to drive an a.c. induction motor in such a way as to obtain a dynamic performance in all respects better than could be obtained with a phase controlled d.c. drive combination.

19.2.2.1 Fundamental equations and performance

Consider the stator winding of a simple three-phase 2-pole a.c. cage induction motor, each phase winding having only one slot per pole per phase as shown in *Figure 19.5*. End-connections for the winding coils are not shown, but R and R1 represent the start and finish of the red phase winding and similarly for the Y and B phase conductors. The R, Y and B phase windings are displaced 120° in space relative to one another.

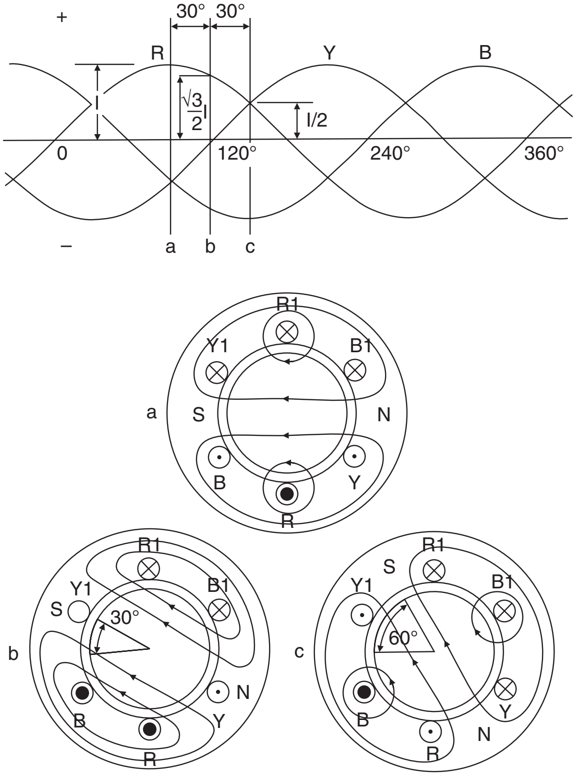


Figure 19.5 Three-phase rotating field

Assuming that when stator current is positive it is flowing inwards in conductors R, Y and B, and therefore outwards in R1, Y1, B1; that the current in phase R in *Figure 19.5*(*a*), is at its maximum positive value and that in phase windings Y and B the currents at the same instant are negative and each equal to half maximum value: then these currents produce the magnetic fluxes represented in *Figure 19.5*(*a*) and the flux axis is horizontal.

Thirty degrees later in the supply cycle, the currents in phases R and B are 0.866 ($\sqrt{3}/2$) of their maximum, and zero in the Y phase. The pattern of the flux due to this current is shown in *Figure 19.5*(*b*). It will be noted that the axis of this field is now in line with coil Y–Y1 and therefore has turned clockwise through 30° from that of *Figure 19.5*(*a*).

After a further 30° in the supply cycle the current in phase winding B has reached maximum negative value, the currents in R and Y are both positive, at half their maximum value. These currents produce the magnetic flux shown in *Figure 19.5*(*c*), the flux axis being displaced clockwise by a further 30° compared with that of *Figure 19.5*(*b*).

Thus for every time-interval corresponding to 30° in the supply cycle, the axis of the flux in a 2-pole a.c. stator rotates 30° in space. With a 2-pole stator (one pair of poles) the flux rotates through one revolution in space in one cycle of the power supply. The magnetic flux is said to rotate at synchronous speed. The rotational speed of the flux is:

f/p revolutions in one second

where $f =$ supply frequency in Hertz, and
$p =$ number of pole pairs

Note: A 2-pole motor has 1 pole pair.

It is more usual to express speed in revolutions per minute:

$60 \cdot f/p$ revolutions per minute (r.p.m.)

The e.m.f. generated in a rotor conductor by transformer action is at a maximum in the region of maximum flux density. The e.m.f. generated in the single rotor conductor produces a current, the consequence being a force is exerted on the rotor tending to move it in the direction of the flux rotation. The higher the speed of the rotor, the lower the speed of the rotating stator flux field relative to the rotor winding, and therefore the smaller is the e.m.f. generated in the rotor winding.

If the speed of the rotor became the same as that of the rotating field, i.e. synchronous speed, the rotor conductors would be stationary in relation to the rotating flux. This would produce no e.m.f. and no rotor current and therefore no torque on the rotor. Because of friction and windage, the rotor could not continue to rotate at synchronous speed. The rotor speed must therefore fall below synchronous, and as it does so rotor e.m.f. and current, and therefore torque, will increase until it matches that required by the losses and by any load on the motor shaft.

The difference in rotor speed relative to that of the rotating stator flux is known as the slip. It is usual to express slip as a percentage of the synchronous speed. Slip is closely proportional to torque from zero to full load.

The most popular squirrel cage induction motor in sizes up to about 5 kW is of 4-pole design. Its synchronous speed with a 50 Hz supply is therefore:

$60 \cdot f/p = 60 \cdot 50/2 = 1500$ (r.p.m.)

Slip accounts for about 5%, and a typical nameplate speed is 1425 r.p.m.

Voltage–frequency relationship If it is desired to convert a constant speed motor operating direct-on-line to a variable speed drive using an inverter it is necessary to consider the effect of frequency on flux and torque.An induction motor on a normal supply operates with a rotating field set up by three-phase currents in the stator winding. The magnitude of the field is controlled broadly by the voltage impressed upon the winding by the supply. This is because the resistance of the windings results in only a small voltage drop, even at full load current, and therefore in the steady-state the supply voltage must be balanced by the e.m.f. induced by the rotating field. This e.m.f. depends on the product of three factors:

(1) The total flux per pole (which is usually determined by the machine designer);
(2) The total number of turns per phase of the stator winding; and
(3) The rate of field rotation or frequency.

Exactly the same factors are valid for transformer design, except that the field is pulsating instead of rotating. For inverter operator the speed of field rotation for which maximum voltage is appropriate is known as the 'base speed'.

The consequence of reducing the supply frequency can readily be deduced from the relationship described above. For the same flux the induced e.m.f. in the stator winding will be proportional to frequency, hence the voltage supplied to the machine windings must be correspondingly reduced in order to avoid heavy saturation of the core. This is valid for changes in frequency over a wide range. The voltage/frequency relationship should therefore be linear if a constant flux is to be maintained within the machine, as the designer intended. If flux is constant so is the motor torque for a given stator current, hence the drive has a constant torque characteristic.

Although constant V/f control is an important underlying principle, it is appropriate to point out departures from it which are essential if a wide speed range is to be covered. Firstly, operation above base speed is easily achieved by increasing the output frequency of the inverter above the normal mains frequency; two or three times base speed is easily obtained. The output voltage of an inverter cannot usually be made higher than its input voltage therefore the V/f characteristic is typically like that shown in *Figure 19.6*(*a*). Since V is constant above base speed, the flux will fall as the frequency is increased after the output voltage limit is reached. The machine flux falls (*Figure 19.6*(*b*)) in direct proportion to the actual V/f ratio. Although this greatly reduces the core losses, the ability of the machine to produce torque is impaired and less mechanical load is needed to draw full load current from the inverter. The drive is said to have a constant-power characteristic above base speed. Many applications not requiring full torque at high speeds can make use of this extended speed range.

The second operating condition where departure from a constant V/f is beneficial is at very low speeds, whereby the voltage drop arising from the stator resistance becomes significantly large. This voltage drop is at the expense of flux, as shown in *Figure 19.6*(*b*). To maintain a truly constant flux within the machine the terminal voltage must be increased above the constant V/f value to compensate for the stator resistance effect. Indeed, as output frequency approaches zero, the optimum voltage becomes the voltage equal to the stator IR drop. Compensation for stator resistance is normally referred to as 'voltage boost' and almost all inverters offer some form of adjustment so that the degree of voltage boost can be matched to the actual winding resistance. It is normal for the boost to be gradually tapered to zero as the frequency progresses towards base speed. *Figure 19.6*(*c*) shows a typical scheme for tapered boost. It is important to appreciate that the level of voltage boost should increase if a high starting torque is required, since in this case the IR drop will be greater by virtue of the increased stator current. In this case automatic load-dependent

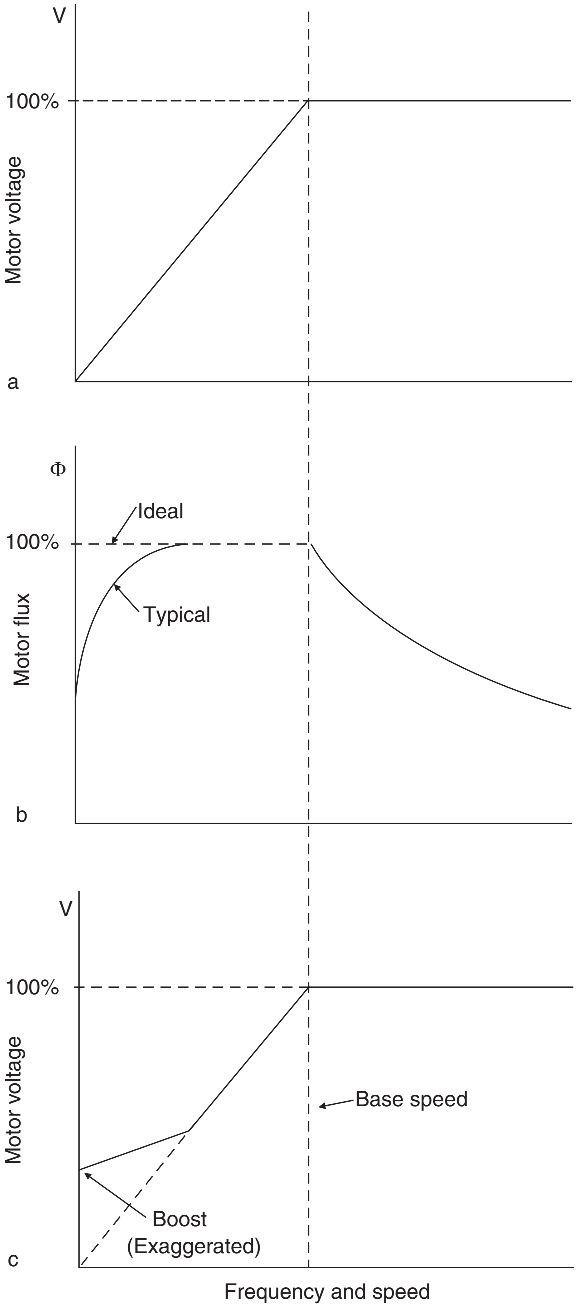


Figure 19.6 Voltage/frequency characteristics: (a) linear *V/f* below base speed; (b) typical motor flux with linear *V/f* (showing fall in flux at low frequency as well as above base speed); (c) modified *V/f* characteristic with low frequency boost (to compensate for stator resistance effects in steady state)

boost control is useful in obtaining the desired low speed characteristics. Such a strategy is referred to as constant *V*/*f* (or V/Hz) control and is a feature of most commercially available a.c. drives though more advanced open-loop strategies are now becoming more available.

So far the techniques described have been based on achieving constant flux within the air gap of the machine or, if that is not possible, then the maximum flux. Constant flux is the ideal condition if the largest capability of torque is required because the load cannot be predicted with certainty, or if the most rapid possible acceleration time is desired. A large number of inverters are used however for variable air volume applications where control of airflow is obtained by variable speed fans. The torque required by a fan follows a square law characteristic with respect to speed and reducing the speed of a fan by 50% will reduce its torque requirement to only 25% of rated. As the load is entirely predictable there is no need for full torque capability and hence flux to be maintained, and higher motor efficiency can be obtained by operating at a reduced flux level. A further benefit is that acoustic noise, a major concern in air conditioning equipment, is significantly reduced. It is therefore common for inverters to have an alternative square law *V*/*f* characteristic or, ideally, a self-optimising economy feature so that rapid acceleration to meet a new speed demand is followed by settling to a highly efficient operating point.

Slip ring induction motor The wound rotor or slip-ring a.c. machine whilst introducing the negative aspect of brushes does address some of the disadvantages of the cage induction motor but with the handicap of cost compared to the equivalent rated d.c. shunt-wound machine.

With the correct value of (usually) resistance inserted in the rotor circuit, near-unity relationship between torque and supply current at starting can be achieved, i.e. 100% full load torque, with 100% full load current, 200% FLT with 200% FLC etc., (i.e. comparable with the starting capability of the d.c. machine). Not only the high starting efficiency but also the smooth controlled acceleration historically gave the slip-ring motor great popularity for lift, hoist and crane applications. It has had similar popularity with fan engineers, to provide a limited range of air volume control, either 2:1 or 3:1 reduction, at constant load, by the use of speed regulating variable resistances in the rotor circuit. Although a fan possesses a square law torque–speed characteristic, so that motor currents fall considerably with speed, losses in the rotor regulator at lower motor speeds are still relatively high, severely limiting the useful speed range.

Rotor slip-ring systems, used with this type of motor, offer a similar service life to that of the d.c. motor commutator system.

Efficient variable speed control of slip-ring motors can be achieved by converters based upon the slip energy recovery principle first proposed by Krammer. Such schemes are based upon converting the slip frequency on the rotor to supply frequency. It is also possible to retrofit variable frequency inverters to existing slip-ring motors. This can be done simply by shorting out the slip-ring terminations (ideally on the rotor thereby eliminating the brushes) and treating the motor as a cage machine.

Variable voltage control of slip-ring motors has been used extensively, notably in crane and lift applications, though these are now largely being replaced by flux vector drives and will therefore not be considered further.

19.2.3 A.c. synchronous motors

A.c. induction motors produce shaft torque which is proportional to percentage slip, implying that with zero slip the machine produces zero torque. In a synchronous motor, torque can be produced at synchronous speed.

This is achieved by a field winding, generally wound on the rotor, and d.c. excited so that it produces a rotor flux which is stationary relative to the rotor. Torque is produced when the rotating three-phase field and the rotor field are stationary relative to each other, hence there must be physical rotation of the rotor at speed n_s in order that its field travels in step with the stator field axis. At any other speed a rotor pole flux would approach alternately a stator north pole flux, then a south pole flux, changing the resulting torque from a positive to a negative value at a frequency related to the flux speed difference, the mean torque being zero.

A typical inverter for variable speed control automatically regulates the main stator voltage to be in proportion to motor frequency. It is possible to arrange an excitation control loop which monitors the main stator voltage and increases the excitation field voltage proportionately.

The a.c. synchronous motor appears to have some attractive features for inverter variable speed drive applications, particularly at ratings of 40 MW and above. Not least is overall cost when compared with an a.c. cage motor plus inverter, or d.c. shunt wound motor and converter alternatives. In applications requiring a synchronous speed relationship between multiple drives or precise speed control of single large drives the a.c. synchronous motor plus inverter control system appears attractive: freedom from brushgear maintenance, good working efficiency and power factor are the main considerations.

19.2.4 Brushless servomotors

The synchronous machine with permanent magnets on the rotor is the heart of the modern brushless servo motor.

The synchronous motor stays in synchronism with the supply, though there is a limit to the maximum torque which can be developed before the rotor is forced out of synchronism. 'Pull out torque' will be typically between 1.5 and 4 times the continuously rated torque. The torque–speed curve is therefore simply a vertical line, which indicates if we try to force the machine above the synchronous speed it will become a generator.

The industrial application of brushless servomotors has grown significantly for several reasons: reduction of price of power conversion products; establishment of advanced control of PWM inverters; development of new more

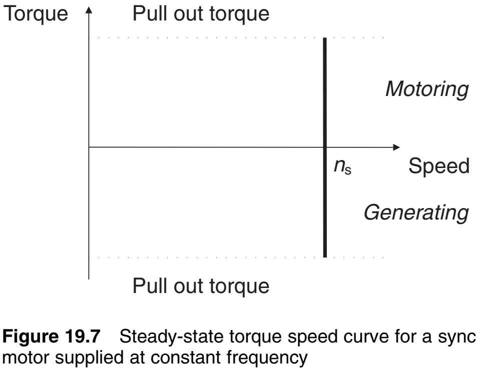


Figure 19.7 Steady-state torque speed curve for a synchronous motor supplied at constant frequency

powerful and easier to use permanent magnet materials; development of highly accurate position sensors, and the manufacture of all these components in a very compact form. They are, in principle, easy to control because the torque is generated in proportion to the current. In addition, they have high efficiency and high dynamic responses can be achieved.

Brushless servomotors are often called brushless d.c. servomotors because their structure is different from that of d.c. servomotors. Brushless servomotors rectify current by means of transistor switching within the associated drive/amplifier, instead of a commutator as used in d.c. servomotors. In order to confuse, brushless servomotors are also called a.c. servomotors because brushless servomotors of synchronous type with a permanent magnet rotor detect the position of the rotational magnetic field to control the three-phase current of the armature. It is now widely recognised that BRUSHLESS AC refers to a motor with a sinusoidal stator winding distribution which is designed for use on a sinusoidal or PWM inverter supply voltage. BRUSHLESS DC refers to a motor with a trapezoidal stator winding distribution which is designed for use on a square wave or block commutation inverter supply voltage.

The brushless servomotor lacks the commutator of the d.c. motor, and has a device (the drive sometimes referred to as the amplifier) for making the current flow according to the rotor position. In the d.c. motor, torque variation is reduced by increasing the number of commutator segments. In the brushless motor, torque variation is reduced by making the coil three-phase and, in the steady state, by controlling the current of each phase into a sine wave.

Stationary torque characteristics A motor which uses permanent magnets to supply the field flux is represented by the simple equivalent circuit of *Figure 19.8*. This is a series circuit of the armature resistance, R_a, and back e.m.f., E.

If we ignore the voltage drop across the transistors, the equation for the voltage is

$$V = R_a I_a + K_e \Omega$$

K_e is known as the back-e.m.f. constant of the motor.
The armature current I_a is

$$I_a = \frac{V - K_e \Omega}{R_a}$$

Therefore, from above, the torque T is

$$\begin{aligned} T &= K_t I_a \\ &= \frac{K_t (V - K_e \Omega)}{R_a} \end{aligned}$$

K_t is known as the torque constant of the motor.

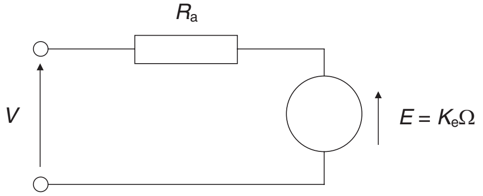


Figure 19.8 Simplified equivalent circuit

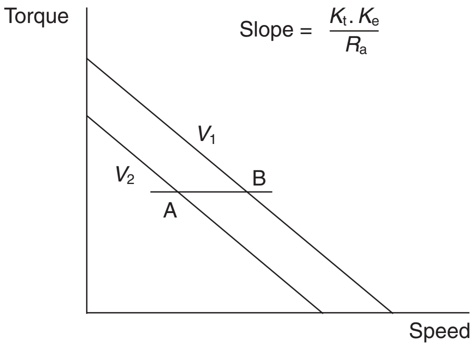


Figure 19.9 Torque-speed characteristic

Figure 19.9 shows the relation between T (torque) and Ω (rotational speed) at two different voltages. The torque decreases linearly as the speed increases. The slope of this function is a constant $K_t K_e / R_a$ and is independent of the terminal voltage and the speed. Such characteristics make the speed or position control of a d.c. motor relatively easy.

The starting torque and the no load speed (assuming no bearings friction and windage loss) are given by

$$T_s = K_t \frac{V}{R_a}$$

$$\Omega_o = \frac{V}{K_e}$$

19.2.5 Reluctance motors

The reluctance motor is arguably the simplest synchronous motor of all, the rotor consisting of a set of iron laminations shaped so that it tends to align itself with the field produced by the stator.

The stator winding is identical to that of a three phase induction motor. The rotor is different in that it contains saliency (a preferred path for the flux). This is the feature which tends to align the rotor with the rotating magnetic field making it a synchronous machine. The practical need to start the motor means that a form of 'starting cage' needs also to be incorporated into the rotor design, and the motor is started as an induction motor, which then the reluctance torque 'pulls in' the rotor to run synchronously in much the same way as a permanent magnet rotor.

Reluctance motors may be used on both fixed frequency (mains) supplies and inverter supplies. These motors tend to be one frame size larger than a similarly rated induction motor and have low power factor (perhaps as low as 0.4) and poor pull in performance. As a result of these limitations their industrial use has not been widespread except for some special applications such as textile machines where large numbers of reluctance motors may be connected to a single 'bulk' inverter and maintain synchronism. Even in this application, as the cost of inverters has reduced, bulk inverters are infrequently used and the reluctance motor is now rarely seen.

19.2.6 Switched reluctance motors

The SR motor is very different from the other polyphase machines described in that both the stator and the rotor have salient poles. The motor can only be used in conjunction with its specific power converter and control, and consequently only overall characteristics are relevant. SR drives are increasing in popularity. They are finding application in high volume appliances and industrial applications which can take good advantage of their charactristics notably high starting torque.

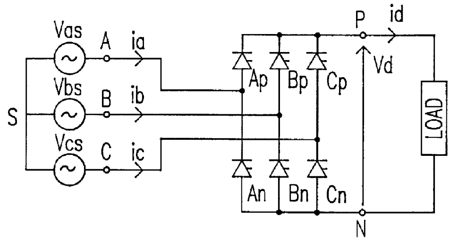
19.3 Drive power circuits

Converter circuits and their characteristics have been described in the Power Electronics chapter of this book. For convenience, the most important configurations used in drive systems are described here.

19.3.1 A.c. to d.c. power conversion

The three-phase controlled converters dominates all but the lowest powers where single-phase converters are used. Two power circuits are of practical importance:

- Fully controlled—This is by far the most important practical bridge arrangement. *Figure 19.10* shows the power circuit together with associated a.c./d.c. relationships. *Figure 19.11* shows how the d.c. voltage can be varied by adjusting the firing delay angle π. The pulse number, p, of this bridge equals 6. Energy flow can be from a.c. to d.c. or d.c. to a.c.
- Half Controlled—In this circuit either the top three devices of *Figure 19.10* (Ap, Bp and Cp) or the bottom three devices (An, Bn and Cn) are replaced by diodes. The pulse number of this bridge equals 3. Only energy flow from a.c. to d.c. is possible. The voltage ripple is much greater than in the case of the fully controlled bridge, *Figure 19.12*, but the a.c. current drawn is lower at reduced d.c. voltage.



CONDUCTING THYRISTORS AC/DC RELATIONSHIPS

TOP ROW	BOTTOM ROW	Vd	ia	ib	ic
Ap	An	0	0	0	0
	Bn	Vab	+id	−id	0
	Cn	Vac	+id	0	−id
Bp	An	Vba	−id	+id	0
	Bn	0	0	0	0
	Cn	Vbc	0	+id	−id
Cp	An	Vca	−id	0	+id
	Bn	Vcb	0	−id	+id
	Cn	0	0	0	0

Figure 19.10 Three-phase fully controlled bridge

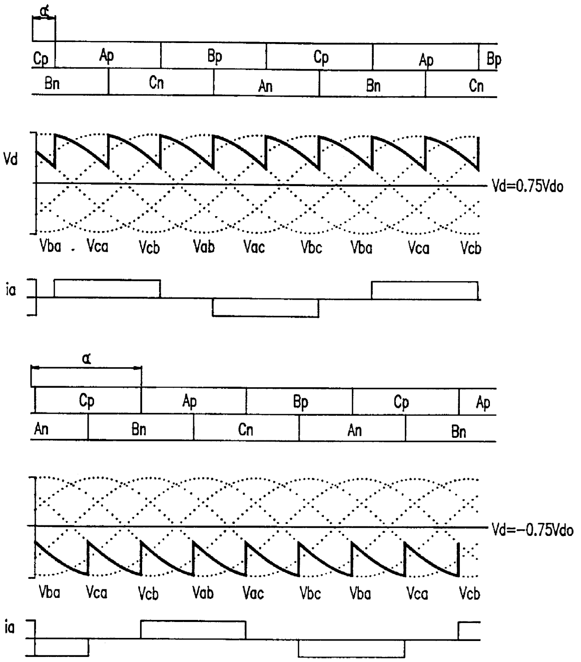


Figure 19.11 Three-phase fully controlled bridge—d.c. voltage control

Summary of characteristics *Table 19.1* shows the salient characteristics of the three-phase a.c. to d.c. converters described above.

Practical effects The characteristics in *Table 19.1* have, for the most part, been based upon idealised conditions of negligible a.c. inductance and constant d.c. current. Whilst these assumptions provide a convenient means for compassion they are not often valid in practice. It is not practicable to consider all such effects here. The effect of d.c. link current ripple on a.c. supply harmonics is of great practical industrial importance mainly in relation to 3-phase bridges (ignoring single-phase traction requirement). Practical experience has led to the adoption by many of the following values:

$I_5 = 0.25I_1$ ('Ideal' $= 0.2I_1$)⇐
$I_7 = 0.13I_1$ ('Ideal' $= 0.14I_1$)⇐
$I_{11} = 0.09I_1$ ('Ideal' $= 0.11I_1$)⇐
$I_{13} = 0.07I_1$ ('Ideal' $= 0.08I_1$)⇐

In general, the amplitudes of higher harmonics are rarely of significance, in regard to supply distortion. Under conditions of very high d.c. current ripple, the 5th harmonic can assume a considerably higher value than that quoted above. A practical example would be an application with a very capacitive d.c. load (e.g. a voltage source inverter) in such a case where no smoothing choke is used I_5 could be as high as $0.5I_1$.

19.3.2 D.c. motor drive systems

In principle little has changed since 1896 when Harry Ward Leonard presented his historic paper 'Volts versus Ohms—the speed regulation of electric motors'. In practice,

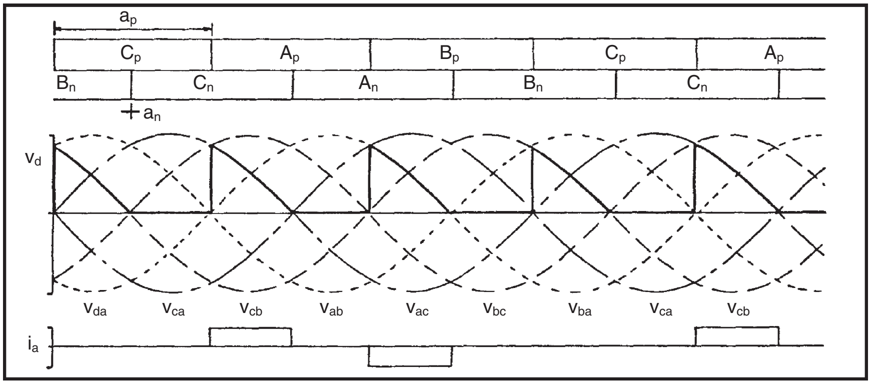


Figure 19.12 Three-phase half controlled bridge—d.c. voltage control

however, many advances have been made from auxiliary machines through mercury arc rectifiers to thyristors.

The d.c. motor is still the most versatile machine for variable-speed drive systems and is often the preferred choice when considerations such as freedom from maintenance or operation under adverse conditions are not paramount.

Earlier, it has been shown that complete control of a d.c. machine can be achieved by controlling the armature voltage, V_a and the field current, I_f. Two power converters are employed for this purpose in most variable speed drives which employ the separately excited d.c. machine. (In referring to the number of converters in a drive, it is common to ignore the field converter—this nomenclature will be adopted below). It is relatively common in simple drives for the field converter to be a single-phase uncontrolled bridge thereby applying fixed field voltage.

In applications where the variation in motor resistance with temperature, or on sites with poorly regulated supplies, which results in unacceptable variations in field current, a controlled power converter is employed with current control. Such field controllers are further discussed later as applied to field weakening control.

Table 19.1 Three-phase converter characteristics

Bridge	*Fully controlled*	*Half controlled*	
Firing angle	α	α	
		$\alpha > \pi/3$	$\alpha < \pi/3$
V_{do}	$\frac{3\sqrt{2}}{\pi} V_s$	$\frac{3\sqrt{2}}{\pi} V_s$	
P_{do}	$\frac{3\sqrt{2}}{\pi} V_s I_d$	$\frac{3\sqrt{2}}{\pi} V_s I_d$	
V_d/V_{do}	$\cos\alpha$	$0.5\,(1+\cos\alpha)$	
I_s/I_d	$\sqrt{(3/2)}$	$\sqrt{\{(\pi-\alpha)/\pi\}}$	$\sqrt{(3/2)}$
Overall power factor	$(3/\pi)\cdot\cos\alpha$	$\frac{\sqrt{3}\cdot(1+\cos\alpha)}{\sqrt{(\pi-\alpha)}}$	$(3/2\pi)\cdot(1+\cos\alpha)$
Maximum corrected power factor	$\frac{\cos\alpha}{\sqrt{\{(\pi^2/9-\sin^2\alpha)\}}}$	$\frac{1+\cos\alpha}{\sqrt{\{2\pi(\pi-\alpha)-\sin^2\alpha\}}}$	$\frac{1+\cos\alpha}{\sqrt{\{4\pi^2/9-\sin^2\alpha\}}}$
Input power/P_{do}	$\cos\alpha$	$0.5\,(1+\cos\alpha)$	
Input VARs/P_{do}	$\sin\alpha$	$0.5\sin\alpha$	
Supply current nth harmonic/I_d	0 for $n=3$, 6, 9....	0 for $n=3$, 6, 9....	
	0 for n even	$\sqrt{(3/n\pi)}\cdot\sqrt{(1-\cos n\alpha)}$ for n even	
	$\sqrt{6}/n\pi$ for n odd	$\sqrt{(3/n\pi)}\cdot\sqrt{(1+\cos n\alpha)}$ for n odd	
Phase of supply current harmonics	$n\alpha$	$n\alpha/2$	

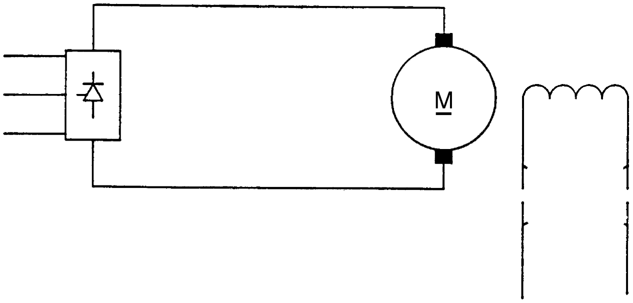


Figure 19.13 Single-phase converter d.c. drive

Single converter drives *Figure 19.13* shows a single-converter d.c. drive.

In its most basic form the motor will drive the load in one direction only without braking or reverse running. It is said to be a 'single-quadrant drive', only operating in one quadrant of the torque–speed characteristic.

Such drives have wide application from simple machine tools to fans, pumps, extruders, agitators, printing machines etc.

If the drive is required to operate in both the forward and reverse directions and/or provide regenerative braking a single fully-controlled converter can still be used however, some means of reversing either the field or armature connections, as shown in *Figure 19.14*, must be added.

Reversal of armature current can involve bulky (high current) reversing switches but due to the low inductance of the armature circuit can be completed in typically 0.2 seconds. Field current reversal takes longer, typically in the order of 1 second, however lower cost reversing switches may be used. The field reversal time can be reduced by using higher voltage field converters to force the current. Forcing voltages up to 4 per unit are used but care must be taken not to overstress the machine. Obviously this increased voltage cannot be applied continuously and requires either a switched a.c. supply or a controlled field converter.

Armature and field reversal techniques are used where torque reversals are infrequent such as hoists, presses, lathes and centrifuges.

Dual-converter drives When a four quadrant drive is required to change the direction of torque rapidly, the delays associated with reversing switches described above may be unacceptable. A dual converter comprising two fully controlled power converters connected in inverse-parallel can be used as shown in *Figure 19.15*. Bridge 1 conducts when the armature current I_a is required to be positive, bridge 2 when it is required to be negative.

There are two common forms of dual converter. In the first, both bridges are controlled simultaneously to give the same mean output voltage. However, the instantaneous

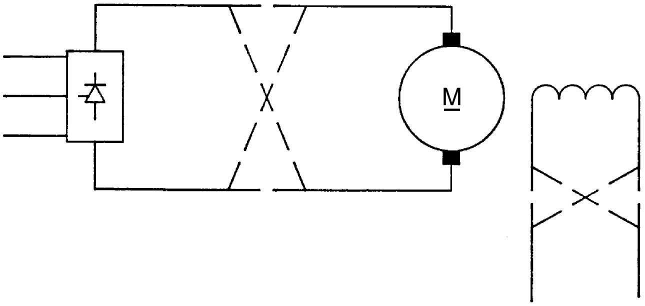


Figure 19.14 Single-phase converter reversing/regenerative d.c. drive

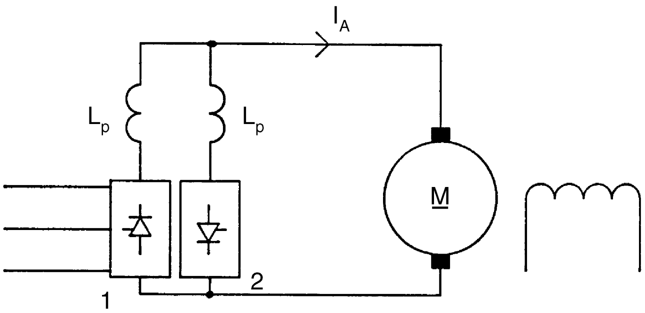


Figure 19.15 Single-phase dual-converter d.c. drive

voltages from the rectifying and inverting bridges cannot be identical, and reactors L_p are included to limit the current circulating between them. The principal advantage of this system is that when the motor torque, and hence current, is required to change direction (polarity), there need be no delay between the conduction of one bridge and the other. This is the dual converter bridge *with circulating current*.

In the other, the circulating current-free dual converter only one bridge at a time is allowed to conduct. The cost and losses associated with the L_p reactors can then be eliminated, and economies can also be made in the drive control circuits. However, the penalty is a short time delay, as the current passes through zero, while it is ensured that the thyristors in one bridge have safely turned off before those in the second are fired. This delay introduces a 'torque-free' period of typically 10 ms. Speed reversal for a 3 kW drive of this type from −1500 to +1500 rev/min can still be achieved in approx 200 ms.

This circulating current-free dual converter is by far the most common industrial four-quadrant drive and is used in many demanding applications—paper, plastics and textile machines where rapid control of tension is required, are good examples.

Field control The output power of a motor is the product of torque and speed. If torque reduces in proportion to speed increases, the motor is said to have a constant power characteristic.

In applications where material is coiled or uncoiled at constant tension, the torque required to produce that tension varies in proportion to coil diameter, whilst rotational speed required to maintain a constant peripheral speed (and therefore line speed) is inversely proportional to diameter. A motor having a constant power characteristic is well suited to this type of application, the advantage being that a smaller motor can be used than would otherwise be the case. Machine tool drives also make use of constant power operation, since loads are small at high speeds, whilst heavy work is done at low speed.

The torque produced by a d.c. motor is proportional to the product of armature current and field flux. By weakening the field as speed increases, a constant power characteristic can be achieved.

In practice there are two major techniques for field weakening both of which rely on a field controller which itself is a simple thyristor converter operating in a current control mode.

In the first method, suitable for coiler and uncoiler applications, the field current reference is arranged to be inversely proportional to coil diameter (measured directly, or calculated from the ratio of line speed to motor speed). Since flux is not strictly proportional to field current, this method does not give a true constant-power characteristic unless compensation for the non-linear part of the motor field curve is applied.

The second method is to use the field controller with an outer, voltage loop having a fixed reference, and to use motor armature voltage as the feedback signal. At low speeds, the voltage loop saturates, providing maximum field current, since armature voltage is below the set value. As speed increases, the armature voltage rises to the point where it matches the preset reference in the field controller. Above that speed, an error signal is produced by the voltage loop, which causes the field controller to weaken the motor field current and thereby restore armature voltage to the setpoint level. The resulting characteristics are shown in *Figure 19.16*.

As regenerative braking depends on the return of power from the motor to the mains, it cannot work if the mains supply fails due to a blown fuse or a power cut. Dynamic braking of 4Q drives is often encountered as a fail-safe means of stopping the motor and its load, and as the only means of (reverse) braking of single-ended drives. This involves switching in a resistor across the d.c. motor.

Since the kinetic energy of the motor and its load is converted into heat by the braking resistor, it is important to rate it correctly for the duty it is expected to perform, taking account of load inertia and the number of stops per hour.

19.3.3 D.c. to d.c. power conversion

19.3.3.1 General

D.c–d.c. power converters (often refered to as 'choppers') provide the means to change one d.c. voltage to another. It is more usual for the conversion to be to a lower voltage, although step-up converters are available and have significant potential for the future.

D.c–d.c. power converters are fed from a d.c. supply usually comprising an uncontrolled a.c. to d.c. converter or alternatively a battery supply; the controlled d.c. output can then be used to control a d.c. machine as in the case of the controlled a.c. to d.c. converters.

D.c. drives employing controlled a.c. to d.c. converters have several important limitations which are overcome by the d.c.–d.c. converter:

- The inability of a thyristor to interrupt current means that an alternating supply is necessary to commutate the converter—this precludes operation from a d.c. supply. This is a common requirement on battery vehicles and d.c. fed rail traction.
- The d.c. ripple frequency is determined by the a.c. and is, for a 50 Hz supply frequency, 100 Hz for single-phase and 300 Hz for three-phase fully-controlled bridges. This means that additional smoothing components are often required when using high speed machines, permanent magnet motors or other special motors with low armature inductance.
- As a result of the delay inherent in thyristor switching (3.3 ms in a 50 Hz 3-phase converter) the current control loop band width of the converter is limited to approximately 100 Hz, which is too low for many servo drive applications.
- Thyristor controlled a.c.–d.c. converters have an inherently poor input power factor at low output voltages. (Near unity power factor can be achieved using an uncontrolled rectifier feeding a d.c.–d.c. converter).
- Electronic short-circuit protection is not economically possible with thyristor converters. Protection is normally accomplished by fuses.

D.c.–d.c. converters are however more complex and somewhat less efficient than a.c.–d.c. converters. They find application mainly in d.c. servodrives, rail traction drives and small fractional kW drives employing permanent magnet motors.

Since step-down converters are of greatest practical importance emphasis shall be placed on their consideration. For the purpose of illustration bipolar transistors will be considered however MOSFET, IGBT and GTO's are widely used.

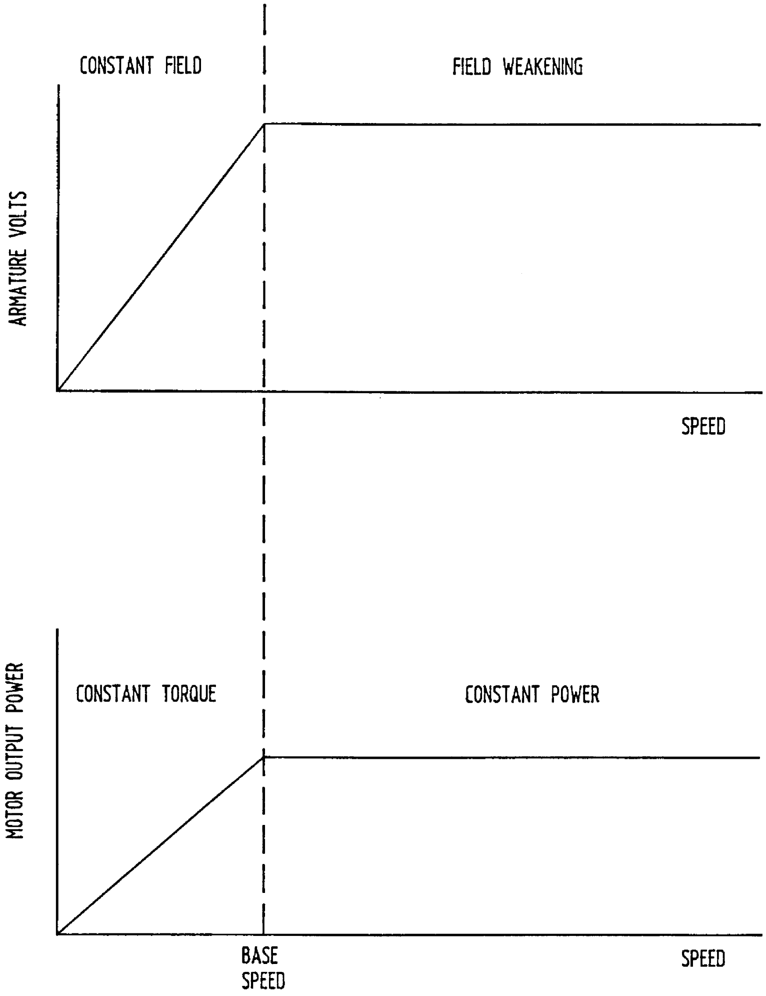


Figure 19.16 Constant power operation using a field controller

19.3.3.2 Step-down d.c.–d.c. converters

Single-quadrant d.c.–d.c. converter The most basic d.c. to d.c. converter is shown in *Figure 19.17*. The output voltage is changed by pulse-width modulation (PWM)—that is, by varying the time for which the transistor T is turned on. The voltage applied to the motor is therefore in the form of a square wave of varying periodicity. Because the motor is inductive the current waveform is smoothed, the flywheel diode D carrying the current whilst the transistor is turned off.

The basic formulae relating the variables in this circuit are as follows:

$$V_a = V_{dc} \cdot t \cdot f$$

$$\Delta I_a = \frac{V_{dc}}{4 \cdot L_a \cdot f}$$

where f = frequency of transistor 'on pulse' (Hz)
ΔI_a = maximum deviation of armature current
L = motor inductance
t = on pulse duration (s)
V_{dc} = source d.c. voltage

The circuit is only capable of supplying unidirectional current and voltage to the motor and is therefore not capable of four-quadrant operation, that is reversing or regenerating.

Applications for this circuit are normally limited to drives below 5 kW and simple variable-speed applications.

Two-quadrant d.c.–d.c. converter In order to achieve full four-quadrant operation a converter must be capable of supplying reversible voltage and current to the motor. A circuit that is capable of two-quadrant operation—that is motoring and braking in one direction only—is shown in *Figure 19.18*. This converter is able to reverse the current flow to the motor but unable to reverse the motor terminal voltage and hence the speed. During motoring, the converter operates as the basic 'chopper' with T1 and D2 carrying the current. During braking (or regeneration) T1 is inoperative and T2 controls the current. During its on-periods, motor current builds up negatively, limited by motor inductance *L*. When T2 turns off, the only path for the current is via D1 back into the supply; hence the circuit is regenerative.

Since this circuit is not capable of motor speed reversal it is normally only used in unidirectional applications. However

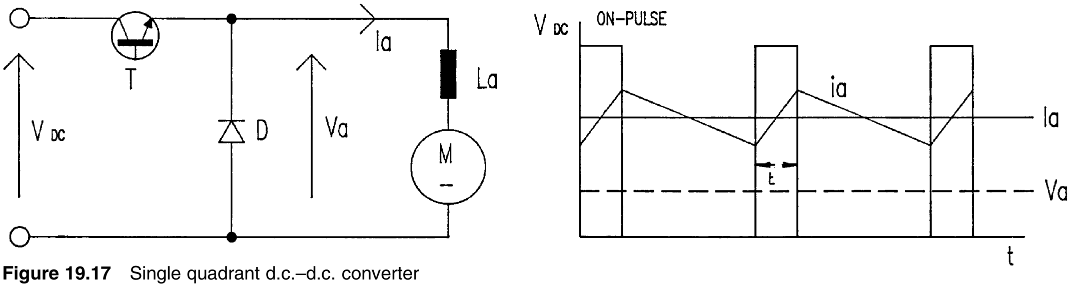

V DC
T
D
Va
Ia
La
M
ON–PULSE
ia
Ia
t
Va
t

Figure 19.17 Single quadrant d.c.–d.c. converter

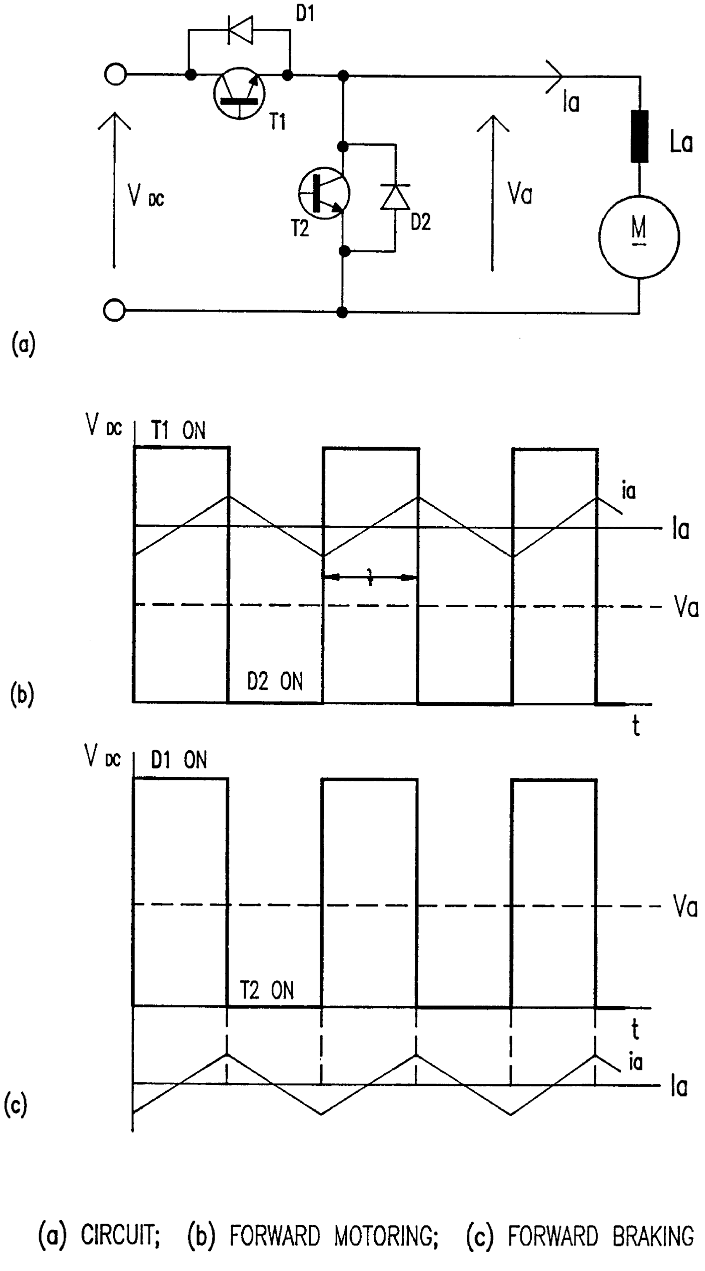

D1
T1
Ia
La
V DC
T2
D2
Va
M
(a)
V DC
T1 ON
ia
Ia
t
Va
D2 ON
(b)
t
V DC
D1 ON
Va
T2 ON
t
ia
Ia
(c)
(a) CIRCUIT; (b) FORWARD MOTORING; (c) FORWARD BRAKING

Figure 19.18 Two-quadrant d.c.–d.c. converter

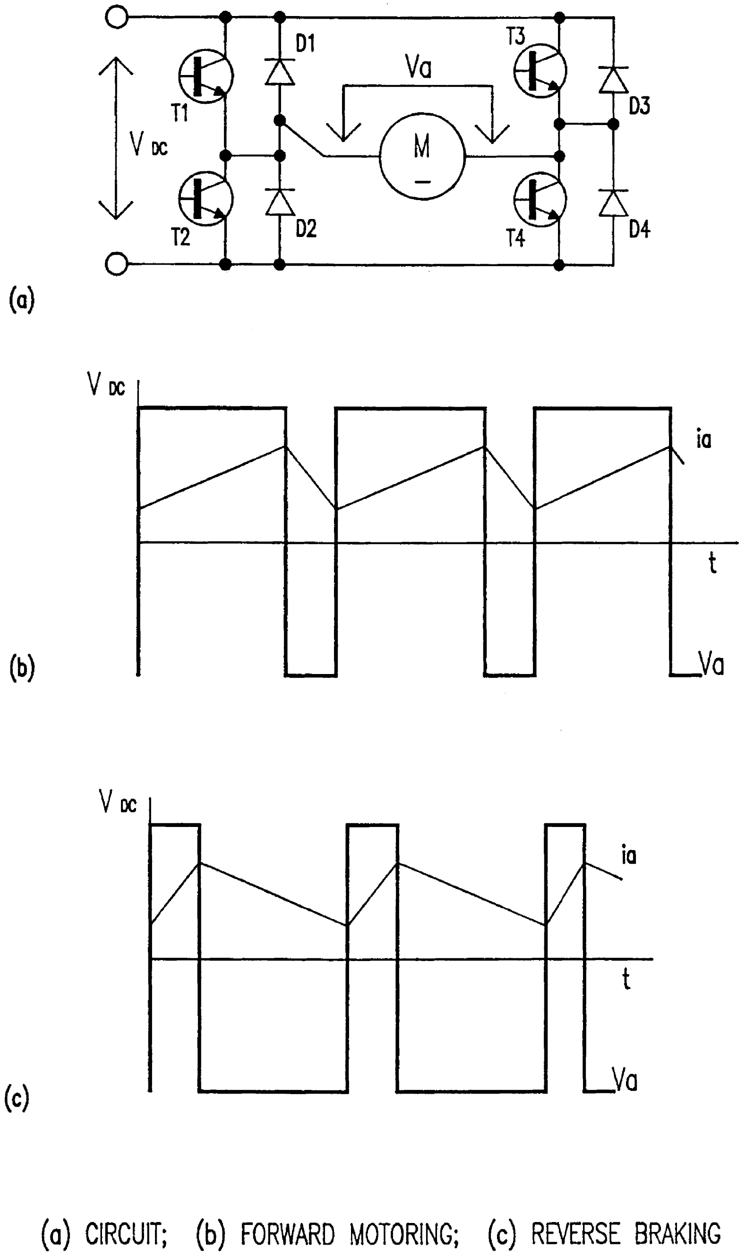


Figure 19.19 Four-quadrant d.c.–d.c. converter

because of its simplicity it is sometimes used in traction applications where reversing is carried out by means of a changeover switch to reverse the armature or field supply.

Four-quadrant d.c.–d.c. converter *Figure 19.19* shows a basic four-quadrant converter capable of supplying reversible voltage and current i.e. of reversing and regeneration.

During motoring, positive output transistors T1 and T4 are switched on during the on-period, whilst diodes D2 and D4 conduct during the off-period. When D2 and D4 conduct, the motor supply is reversed and consequently the voltage is reduced to zero at 50% duty cycle. Any reduction of duty cycle below 50% will cause the output voltage to reverse but with current in the same direction; hence the speed is reversed and the drive is regenerating. With transistors T2 and T3 conducting, the current is reversed and hence the full four-quadrant operation is obtained.

One disadvantage of this converter is that the amplitude of the output ripple voltage is twice that of the simple converter—and the current ripple is therefore worse. This problem can be overcome by a technique known as double-edged modulation. With this technique the flywheel current is circulated via a transistor and a diode during the off-period. For example, after T1 and T4 have been

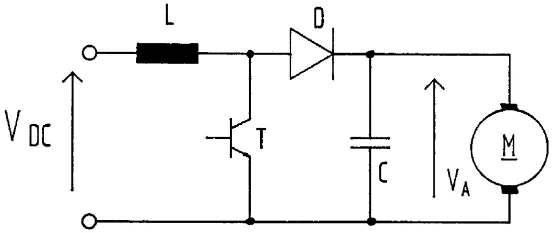


Figure 19.20 Step-up d.c.–d.c. converter

conducting, T4 is turned off and T3 on, so that the flywheel current circulates via T1 and D3. The net effect is a reduction in ripple voltage and a doubling of ripple frequency.

These four-quadrant converters are widely used in high performance d.c. drives such as servos.

19.3.3.3 Step-up d.c.–d.c. converters

As with step down converters many alternative configurations exist for step-up converters. Whilst a full description is not necessary, the principle is of value.

Figure 19.20 shows a much simplified arrangement of a step-up converter. When T is turned on, current builds up in inductor L. When T is turned off, the energy stored in L is transferred to Capacitor C via D. When the capacitor voltage, which is the same as the motor armature voltage reaches the desired level, T is turned on once more. C cannot discharge via T as diode D is reverse biased.

In this way a stabilised voltage typically twice the input d.c. voltage can be obtained. This circuit is particularly useful when operating on low voltage supplies and can lead to very cost effective converter designs.

19.3.4 A.c. to a.c. power converters with intermediate d.c. link

This catagory of a.c. drive commonly termed 'Variable Frequency Inverters' is by far the most important in respect of the majority of industrial applications. It is being considered here as a complete converter, however the input stages have been considered earlier in isolation, and their individual characteristics so described are, of course, applicable. Alternative input stages to some of the drives are applicable.

Also some converters may be used with a variety of machine types. Only practically important combinations are described.

The concept of these 'inverter' drives is well understood—'rectification of fixed frequency, smoothing and then inverting to give variable frequency/variable voltage to feed an a.c. machine'. Within this broad concept two major catagories of drive exist: Firstly, voltage source in which the converter impresses a voltage on the machine, and the machine impedances determine the current. Secondly, current source in which the converter impresses a current on the machine, and the machine impedances determine the voltage.

19.3.4.1 Voltage source inverters

General characteristics The fixed frequency mains supply is a voltage source behind an impedance. Voltage source inverters can be similarly considered, and consequently are very flexible in their application. Major inherent features included:

- Multi-motor loads can be applied—this can be very economical in applications such as roller table drives, spinning machines etc.
- Inverter operation is not dependent upon the machine—indeed various machines (induction, synchronous or even reluctance) can be used provided the current drawn is within the current rating of the inverter. Care should be taken where a low power factor motor is used (e.g. reluctance) to ensure the inverter can provide the required VARs.
- Inherent open-circuit protection, very useful in applications where the cables between the inverter and motor are in some way insecure (e.g. fed via slip-rings, subject to damage etc.).

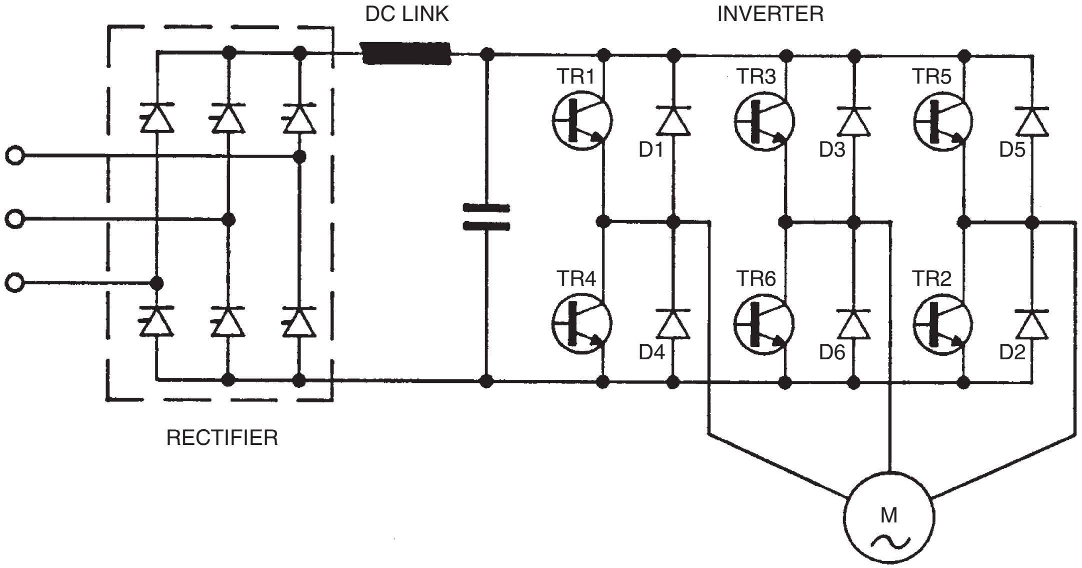


Figure 19.21 Square wave voltage-fed inverter

- Facility to 'ride-through' mains dips can easily be provided by buffering the d.c. voltage link with capacitance or, where necessary, a battery.
- Motoring operation only in both directions is possible without the addition of resistive dumps for braking energy or expensive regenerative converters to feed energy back to the supply.

Six step square wave inverter A typical d.c. link square wave voltage-fed inverter drive is shown in *Figure 19.21*. The three-phase a.c. supply is converted to d.c. in the phase-controlled rectifier stage. The rectified d.c. power is then filtered and fed to the inverter. Note that the d.c. link reactor is small compared to that used in current source designs. Indeed in drives up to about 4 kW it is not practically necessary. Some manufacturers omit the reactor in designs to 400 kW and above, however this has a significant effect upon supply harmonics and unduly stresses the rectifier and filter capacitor.

The inverter switching elements shown as transistors TR1 to TR6 are gated at 60° intervals in the sequence in which they are numbered in the diagram, and each transistor conducts for 180°. The feedback diodes D1 to D6 are connected in inverse-parallel with the transistors, and permit the return of energy from reactive or regenerative loads through the inverter to the d.c. link.

For a star-connected motor, synthesis of inverter output voltage waveforms is shown in *Figure 19.22*. The phase-to-neutral voltage of the inverter has six-step waveshape while the corresponding phase-to-phase voltage has 120° conduction angle. The output frequency is controlled by the rate at which the inverter transistors are triggered into conduction by the inverter control circuitry. Reversing the firing sequence of transistors in the inverter changes the direction of rotation of the motor, and no switching of power leads, either on the incoming supply or to the motor itself, is necessary.

The phase-controlled rectifier regulates the d.c. link voltage and this, in turn, determines the magnitude of the output voltage from the inverter. Thus, the output voltage/frequency relationship may be controlled to regulate the motor flux in the desired manner.

The advantages of the square-wave inverter are high efficiency (98%), suitability to standard motors, potential good reliability and high-speed capability. However, it suffers from low-speed torque pulsations and possible low-speed instability.

In a square wave inverter, each harmonic voltage amplitude is inversely proportional to the harmonic order and hence there are no pronounced high-order harmonics. These are filtered by the motor leakage inductances.

Very high speed motor operation is possible by increasing the output frequency. Faster switching devices such as MOS transistors and insulated gate bipolar transistor (IGBT) can be used to achieve this performance.

It is known that the square wave inverter gives objectionable torque pulsation at low frequency operation, below approximately 5 Hz. This pulsating torque is due to the interactions of low order harmonics with the fundamental voltage, causing a stepping or cogging motion to the rotor running at low speed. Hence, the pulsating torque limits the low frequency operation of the square wave inverter. Appropriate feedback control techniques or flux weakening can attenuate the low speed pulsating torque problems.

The existence of a phase-controlled rectifier to control the voltage of the inverter as illustrated in *Figure 19.21* is an inherent weakness of this circuit. The phase-controlled rectifier will present a low power factor to the a.c. supply, at low speeds, and the d.c. link filter capacitor is large and reduces the response time of the system to voltage and hence speed changes. If the drive system is one for which regenerative braking operation is a requirement, the rectifier has to be of inverse-parallel type. The input power-factor and response time of the drive can be improved by replacing the phase-controlled rectifier with a diode rectifier feeding a d.c. chopper which regulates the input voltage to the inverter. For recovering regenerative energy of the load, a two-quadrant chopper will be necessary. The alternative supply converter arrangement of a diode bridge plus chopper also provides a fixed voltage link which is more economically buffered if mains dip ride-through is required.

The voltage-fed square-wave drive is usually used in low power industrial applications where the speed range is limited to 10 to one and dynamic performance is not important. Recently, this type of drive has largely been superseded by PWM type voltage-fed inverters. Nevertheless, the voltage-fed square wave inverter can be easily adapted to multimotor drives where the speed of a number of induction motors can be closely tracked. It is also used in some high frequency (> 1 kHz) and some high power applications.

Pulse width modulated (PWM) inverter In the PWM inverter drive, the d.c. link voltage is uncontrolled and derived from a simple diode bridge. The output voltage can be controlled electronically within the inverter by using PWM techniques. In this method, the transistors are switched on and off many times within a half cycle to generate a variable-voltage output which is normally low in harmonic content.

A PWM waveform is illustrated in *Figure 19.23*.

A large number of PWM techniques exist each having different performance notably in respect to the stability and audible noise of the driven motor.

Using the PWM technique, low-speed torque pulsations are virtually eliminated since negligible low-order harmonics are present. Hence, this is an ideal solution where a drive system is to be used across a wide speed range.

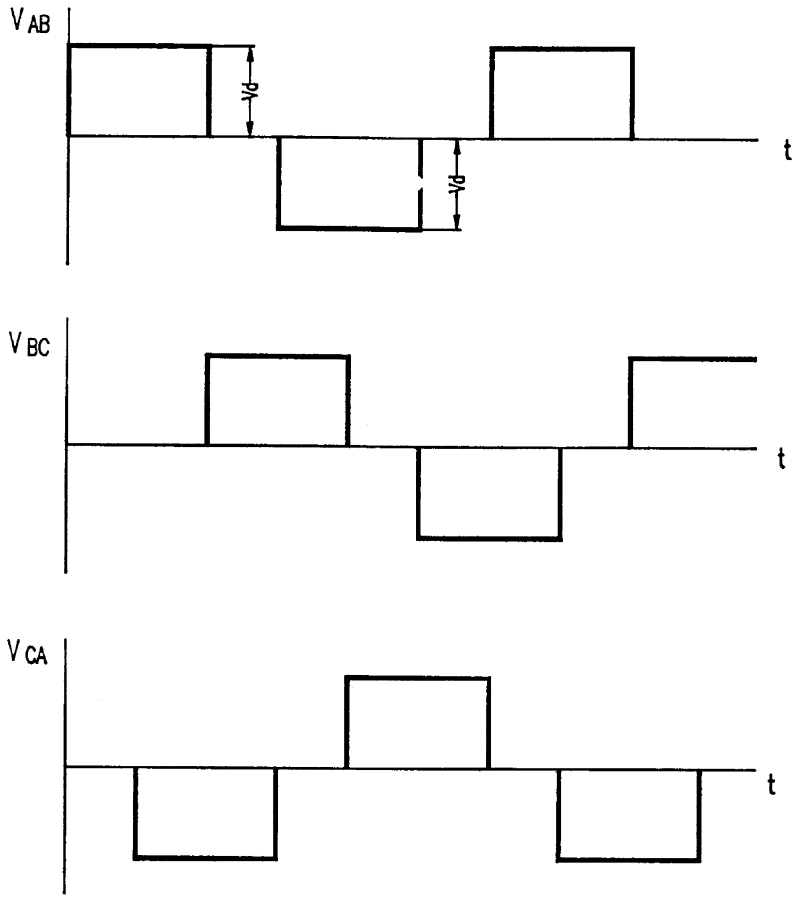
Since voltage and frequency are both controlled with the PWM quick response to changes in demand voltage and frequency can be achieved. Furthermore, with a diode rectifier as the input circuit a high power-factor, approaching unity, is offered to the incoming a.c. supply over the entire speed and load range.

PWM inverter drive efficiency typically approaches 98% but this figure is heavily affected by the choice of switching frequency—the higher the switching frequency the higher the losses in the drive. In practice the maximum fundamental output frequency is usually restricted to 100 Hz in the case of gate turn-off thyristors (GTO) or about 1 kHz for a transistor based system. The upper frequency limit may be improved by making a transition to a less sophisticated PWM waveform with a lower switching frequency and ultimately to a square wave if the application demands it. However, with the introduction of faster-switching power semiconductors, these restrictions to switching frequency and minimum pulse-width have been eased.

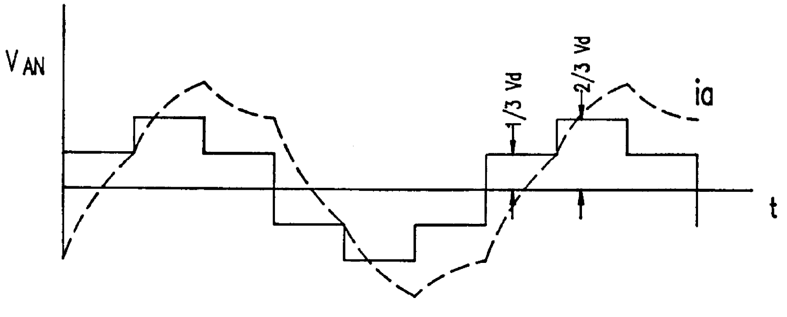
In general, a motor with a large leakage reactance is desirable to limit the flow of harmonic currents and thereby minimise losses.

19.3.4.2 Current source inverters

Whereas each voltage feeding inverter can be used with most forms of a.c. machine, a different design is usually



PHASE-TO-PHASE VOLTAGES



PHASE-TO-NEUTRAL VOLTAGE

Figure 19.22 Square wave voltage-fed inverter—output voltage and current

adopted for synchronous and induction motors. Current source drives are usually, but not always, single-motor systems, and since current is controlled, have simple short-circuit protection.

In contrast to voltage source inverters full four quadrant operation is inherently possible.

Converter-fed synchronous machine Once rotating a synchronous machine generates a.c. voltages which can be used for the natural commutation of a converter connected to its terminals. Indeed the connected synchronous machine behaves as the mains in respect of the a.c to d.c converters described earlier.

Figure 19.24 shows the basic components of the drive system. A low-impedance or ‘stiff’ d.c. current source is required and is obtained from a controlled rectifier and a series reactor. With a stiff current source, the output current wave is not greatly affected by the magnitude of the load.

The synchronous machine can be approximately represented by a counter-emf in series with an equivalent leakage

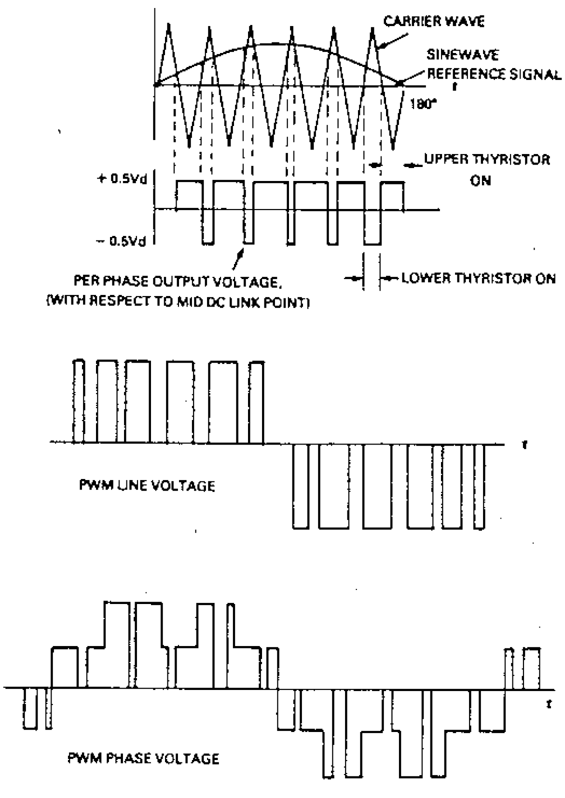


Figure 19.23 Sinusoidal PWM line and phase voltages

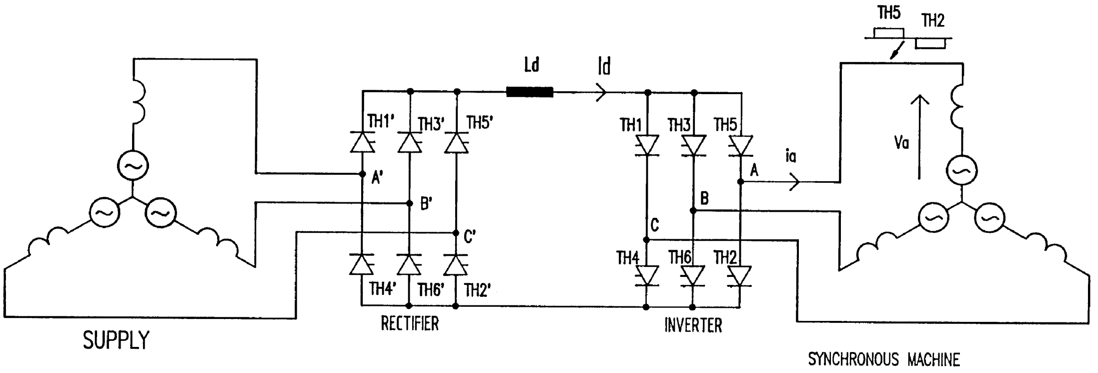


Figure 19.24 Converter-fed synchronous machine

inductance. The d.c. current is switched through the inverter thyristors so as to establish three-phase, six-stepped symmetrical line current waves. Each thyristor conducts for 120° and at any instant one upper thyristor and one lower thyristor remain in conduction.

It is necessary to maintain an approximately constant angular relationship between the rotor and stator m.m.f.s and hence automatically maintain the correct inverter frequency. This is an important point. The inverter does not impose a frequency upon the machine, rather the machine itself determines the frequency. The motor cannot therefore pole-slip. The drive is accelerated by increasing the current fed to the drive which then accelerates and thereby increases the frequency.

As in the d.c. drives, the a.c. supply power factor is poor at low speeds. Full four-quadrant operation is possible without additional components.

Special procedures are necessary for starting these drives since at standstill the machine voltage is not available to commutate the current. In essence this is usually achieved by momentarily switching off the d.c. link current every sixth of a cycle. This allows the thyristors in the inverter to turn off so that the next pair can be fired. Above approximately 5% of rated speed the machine generates sufficient voltage for natural commutation and control is undertaken in a similar manner to that of a d.c. drive.

Applications for this type of drive fall into two main catagories. Firstly, starting converters for large synchronous machines, the converter being rated only for a fraction of the machine rating. Secondly, as large high power (and sometimes high speed) variable speed drives for a variety of applications. Power ratings, typically from 1.5 to 30 MW at speeds up to 8000 rpm are available. Also of import is the fact that high voltage drives are offered with supply voltages up to 5 kV typical, but systems up to 25 kV are in service where the high voltage converter technology is similar to that used for HVDC power converters.

Converter-fed induction motor drive Unlike the Synchronous machine, the induction motor is unable to provide the VARs, or terminal voltage to commutate a converter connected to its terminals. Commercial schemes are available however which are closely based upon the converter fed synchronous machine drive having additional components to provide VAR compensation.

Figure 19.25 shows a basic power circuit. The diagram somewhat belies the potential complexity of the VAR compensator. In its simplest form this could comprise capacitors plus appropriate switches. Control of such a system is somewhat involved. It is often better to use a cycloconverter or even an auxiliary synchronous machine to provide the commutation, and motor VARs.

This system is only appropriate for high power drives, generally above 4 MW where an induction motor is preferred.

Forced commutated induction motor drive This is the most widely used current source inverter at power levels in the range 50–3500 kW at voltages up to normally 690 V.

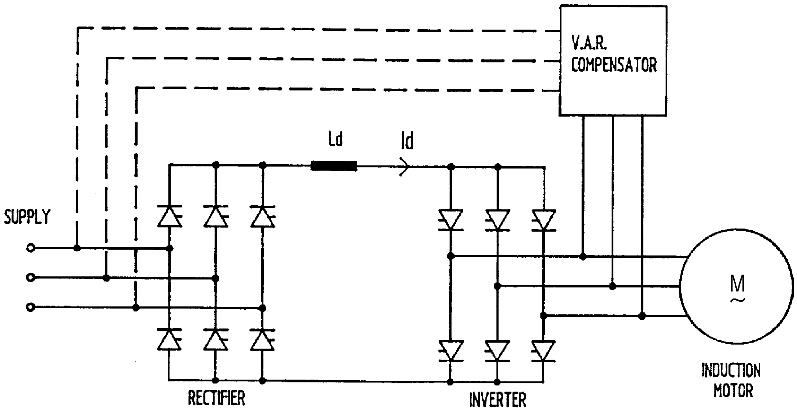


Figure 19.25 Converter-fed induction motor

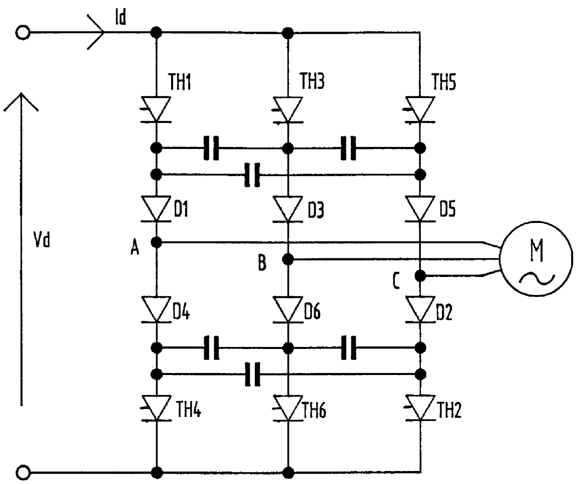


Figure 19.26 Forced commutated induction motor

(High voltage versions 3.3 kV/6.6 kV have been developed however they have not proved to be economically attractive.)

Figure 19.26 shows the inverter and motor of the drive. The d.c. link current I_d, taken from a 'stiff' current source is sequentially switched at the required frequency into the stator windings of the induction motor. The motor voltage waveform is approximately sinusoidal apart from the superposition of voltage spikes caused by the rise and fall of machine current at each commutation. Further distortion is caused by the effects of slot ripple and d.c. current ripple.

The operating frequency range is typically 5–60 Hz, the upper limit being set by the relatively slow commutation process. Special motors with low leakage inductance do offer advantage with this converter and allow reduced capacitance in the inverter and/or higher operating frequency. Below 5 Hz, torque pulsations can be problematic but PWM of the *current* can be used at low frequencies to ease the problem.

This system is most commonly used for single motor applications such as fans, pumps, extruders, compressors etc. where very good dynamic performance is not necessary and a supply power factor which decreases with speed is acceptable.

Static Kramer drive The static Kramer drive is shown in *Figure 19.27*. The drive comprises a slip-ring (wound rotor) induction motor together with an uncontrolled converter, d.c. smoothing reactor and a fully-controlled converter in the rotor circuit.

The diode bridge gives an output voltage V_d which is proportional to the slip of the motor. V_d is opposed by the d.c. voltage of the fully-controlled bridge, a small potential difference being sufficient to circulate current corresponding to the required load torque. Ideally, neglecting losses, the fully-controlled bridge d.c. voltage sets the speed to which the motor will accelerate. Control is therefore very similar to a d.c. drive.

Power can flow in only one direction via the diode bridge which means that motoring torque can be developed only at subsynchronous speeds. For reverse running it is necessary to reverse the phase sequence of the stator supply.

This drive can be very economic when designed for operation over a limited speed range below synchronous

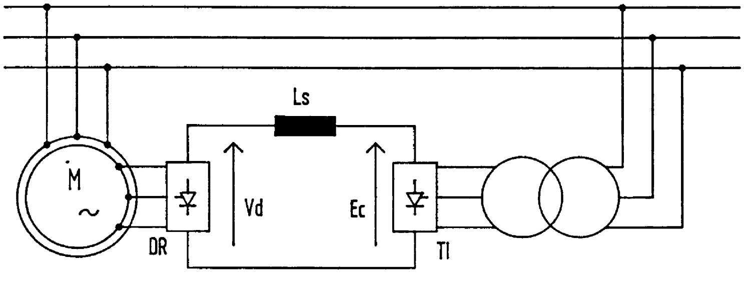


Figure 19.27 Static Kramer drive

speed—this is the useful operating region for fans, pumps etc. The converter bridges required for such limited speed range operation need only be rated at a fraction of that of the machine it is controlling. It is necessary in such designs to provide a starter, usually a resistance to run the motor up to the lowest controllable speed. This means that should there be a fault with the converter equipment, the system can be easily designed to run at full speed without the controller.

Note that the supply harmonic currents and VARS associated with the converter part of the drive may be substantially reduced by adopting a limited speed range solution.

The static Kramer drive finds application mainly at ratings between 1 and 20 MW, with induction motors with 4 or more poles. (Stability problems exist with 2-pole motors which can only be resolved with care.) Speed ranges of 30% are typical (i.e. 70–100% rated speed). The induction motor stator can be wound for any conventional voltage e.g. 6.6 kV, 11 kV.

19.3.5 Direct a.c. to a.c. power converters

This final catagory of power converters converts the fixed frequency, fixed voltage a.c. supply to a variable frequency and/or variable voltage without an intermediate d.c. link.

19.3.5.1 Soft starter/voltage regulator

Figure 19.28, shows a typical soft start comprising inverse parallel connected thyristors in each supply line to an induction motor. Alternative connections are available but the principles are similar. The converter is used to control the voltage applied to the motor and in this way 'soften' the effects of switching an induction motor direct-on-line (DOL). Whilst the converter will control the current drawn from the supply, its most usual application is in controlling torque to provide smooth 'jolt' free acceleration.

Because the stator frequency is unchanged, a reduced running voltage, and hence flux, equates to a large slip which results in additional rotor losses—care must therefore be taken in its application.

In a number of specialised cases, purpose designed high resistance rotors (or slip-ring motors with external rotor resistors) are used to form a variable speed drive—the rational for such a system is based more on history than technology.

More recently such converters have been employed as combined soft starters/power-factor controllers/energy saving devices. The case for significant energy saving by this form of converter is often hard to prove.

A crude form of frequency control is possible by modulating (varying cyclically) the thyristor firing angles at the required output frequency. Whilst commercial systems are available they are of limited value since supply current and motor torque are of poor quality.

19.3.5.2 Cycloconverter

A typical scheme for a cycloconverter drive is shown in *Figure 19.29*. Each motor phase is supplied, in effect, from a dual a.c. to d.c. converter which was described earlier. It is usual to employ circulating current-free converters. To avoid line to line short-circuits isolating transformers are used on the supply side. By modulating the firing angles of the dual bridge converters, a controllable three-phase set of voltages can be produced suitable for feeding polyphase a.c. machines.

The drive is inherently 4-quadrant. The maximum output frequency is limited to approximately half the supply frequency by considerations related to harmonics in the motor currents and torque, stability and dimensions of the drive components. The cycloconverter therefore finds application in low-speed drives. The complexity of the drive also means that only high power systems (>1 MW), or specialised applications (e.g. conveyor drives for use in hazardous environments) are economic.

They are used on large ball mills, minewinders etc. They are also used to feed multimotor loads such as roller tables.

Due to the modulation of the converter firing angles, the harmonic content of the a.c. supply is complex and designs for appropriate harmonic filters somewhat involved.

The cycloconverter is suitable for feeding both induction and synchronous machines. In specialised applications such as wind generators, cycloconverters have been placed in the rotor circuit of a slip-ring induction motor. Such a

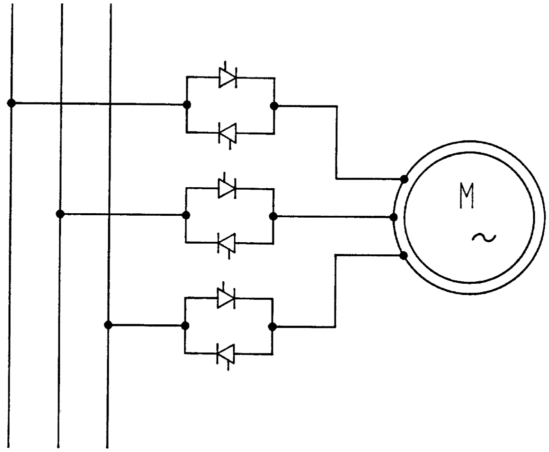


Figure 19.28 Typical soft start

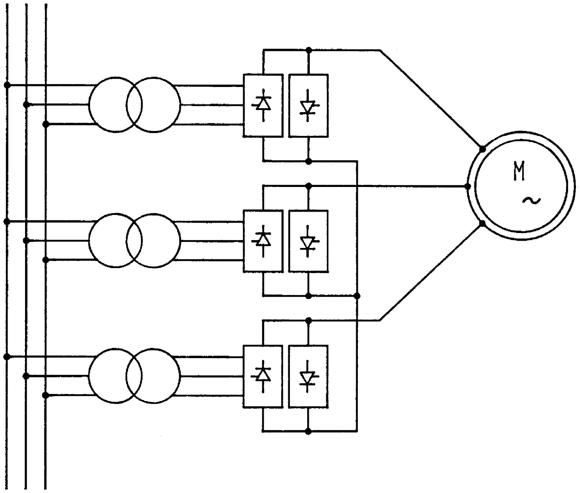


Figure 19.29 Cycloconverter

system, known as a static Scherbius drive which is detailed below.

19.3.5.3 Static Scherbius drive

The static Scherbius drive is closely related to the static Kramer drive, with the single quadrant diode bridge in the rotor circuit replaced by a cycloconverter.

The cycloconverter is used as the voltage and frequency changer between the rotor and the supply. The cycloconverter is inherently regenerative, and the output can be controlled up to half the supply frequency in both phase sequences. It is thus possible for the system to operate as a full 4-quadrant drive. For a given converter rating the range of speed control is therefore twice that of a static Kramer drive.

The relative complexity of the drive limits its application to somewhat specialised high power applications where a very limited speed range only is required and perhaps stringent harmonic current limits have been imposed by the supply authority.

19.3.5.4 Matrix converter

Recently attention has been refocused on the matrix converter shown in *Figure 19.30*. Whilst the basic circuit is not new, recent advances in power devices offer the potential to overcome many of the drawbacks inherent in the circuit when the switches comprise inverse parallel thyristors. Limitations in the maximum output voltage (86% input) means that its application in the commercial industrial market could be problematic. There are prospects in regard to integrated motors and some servo systems where machine voltage is not seen as critical.

Commercial systems are available only for very specialised applications at present. It has yet to be proven to be a practical and cost effective industrial drive although some major drives companies are working on this technology.

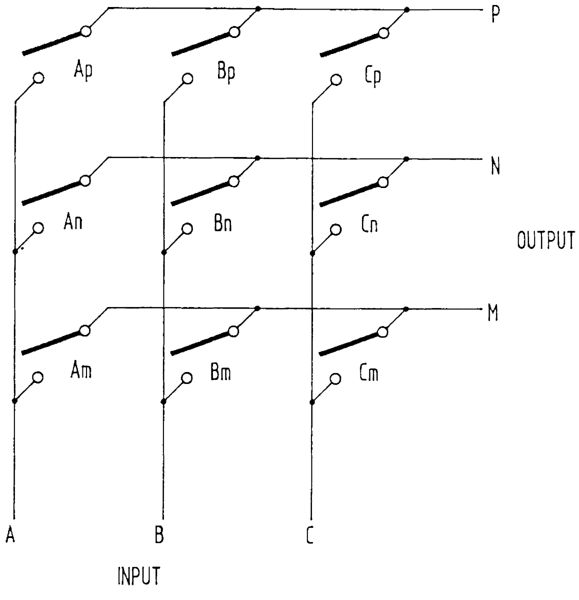


Figure 19.30 Matrix converter

19.4 Drive control

19.4.1 D.c. drive control

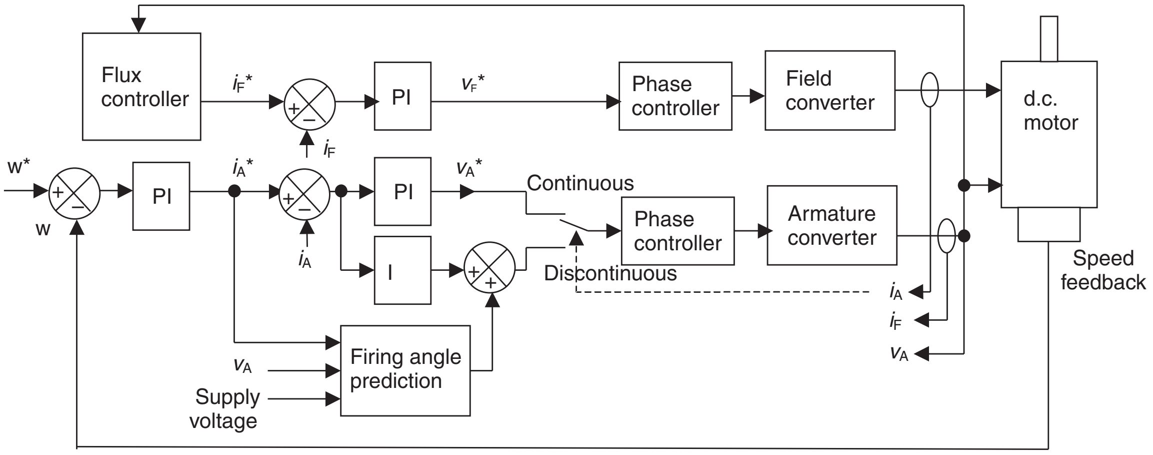
Most medium and large size industrial d.c. motor drives are based on a separately excited motor. The flux is generated by a field winding and the torque by a higher current armature winding fed via a commutator. These two windings are completely independent, and so the flux and torque can be controlled independently as shown below in *Figure 19.31*(*a*). Although the field winding usually has a long time constant, the armature winding time constant is normally very short allowing fast changes of armature current and hence torque.

Flux controller The field converter is either a half or fully controlled thyristor converter. The half controlled converter can only apply positive voltage to the field winding, and so the current in the winding can be increased quickly, but decays slowly. The fully controlled converter can apply positive or negative voltages, and so the performance is the same whether the current is increasing or decreasing. When the motor is rotating at a speed below base speed the field current reference is constant at the rated level for the motor, and so the motor armature voltage increases with speed. The armature voltage reaches its rated level at base speed, and above this speed the flux controller reduces the field current reference, i_F^*, to keep the voltage at the rated level as the speed increases further. The armature voltage feedback can include armature resistance compensation to avoid the effects of the armature resistance drop on the voltage control loop.

Torque controller For four-quadrant operation two thyristor bridges are used as shown in *Figure 19.31*(*b*). Both bridges can apply positive or negative voltage to the motor, but the positive bridge can only supply positive current and the negative bridge negative current. Therefore the positive bridge conducts when positive torque is required and the negative bridge when negative torque is required. The bridges are phase controlled to apply the voltages required by the reference, V_A^*, to the motor. Due to the high voltage ripple in the converter output and the unidirectional nature of thyristors, the current in the armature can be continuous or discontinuous. While the current is continuous the relationship between the voltage reference and the actual applied voltage is a cosine function and the voltage reference can be used directly to control the firing angle of the converter. When the current is discontinuous the relationship is highly non-linear and varies with the voltage level applied to the motor. The drive stores the relationship between the firing angle and motor current for different output voltage levels, and during discontinuous current operation the correct firing angle is selected by the firing angle prediction block for a given current reference. Any errors are trimmed out by an integrator operating on the current error.

When a change in direction of torque is required one bridge must stop conducting and the other bridge must become active. It is clear from the power circuit diagram that only one bridge must conduct at a time during this changeover to avoid a short circuit across the armature supply. It is important that this changeover occurs as quickly as possible to give good dynamic torque control. Modern microprocessor controlled drives enable intelligent methods to be used to keep the bridge changeover delay as short as possible.

(a) Control system


Flux controller
iF*
+
−
PI
vF*
Phase controller
Field converter
d.c. motor
iF
w*
+
−
w
PI
iA*
+
−
PI
vA*
Continuous
Phase controller
Armature converter
iA
I
+
+
Discontinuous
iA
iF
vA
Firing angle prediction
vA
Supply voltage
Speed feedback

(b) Power circuit

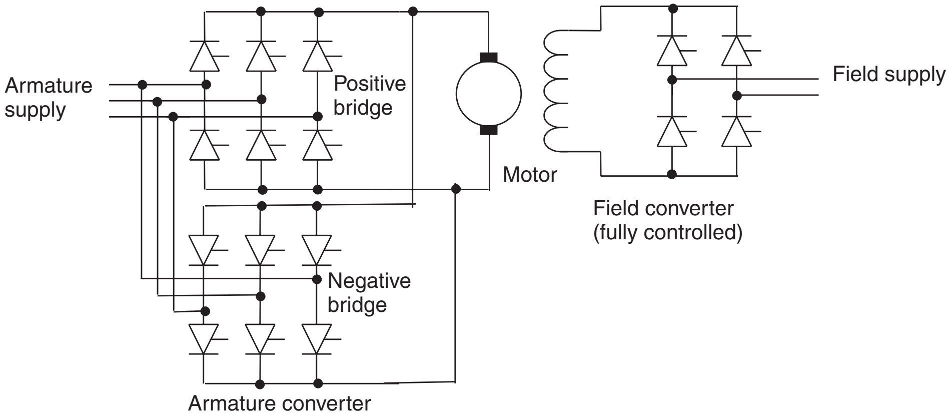

Armature supply
Positive bridge
Motor
Field supply
Field converter (fully controlled)
Negative bridge
Armature converter

Figure 19.31 Separately excited d.c. motor drive

Performance and applications The performance characteristics of a d.c. motor drive can be summarised as:

- The current controller sample rate is limited by the possible commutation rate of the thyristors in the bridges. In general the sample rate and hence the bandwidth of a d.c. drive is ten times lower than that of an a.c. drive.
- The d.c. drive with separately excited d.c. motor is used in similar applications to closed-loop induction motor drives. Because the flux and torque are controlled by separate winding, the decoupling of flux and torque control is not dependent on knowledge of the motor parameters. Therefore accurate control of torque is more easily achieved.
- The thyristor power circuits used in a d.c. drive cost less than an IGBT inverter for an a.c. drive of equivalent power rating. However, a d.c. motor is more expensive than an a.c. motor of equivalent power rating, at power ratings below approximately 200 kW. In general d.c. motors require more maintenance than a.c. motors.
- A full four-quadrant d.c. drive can be constructed with two thyristor bridges, however, the input power factor is poor. A full four-quadrant a.c. drive using an input converter as described above takes currents with significantly less harmonics and with almost unity displacement factor.
- The drive can be used in torque control, without the speed controller.

Although it has been predicted for many years that a.c. drives would replace d.c. drives, this has only happened slowly and many d.c. drives are still manufactured. These are used in many industrial applications especially in larger sizes. The following are some examples of where d.c. drives are used.

- Cranes and hoists
- Lifts
- Material winding

19.4.2 A.c. drive control

There are many types of variable speed drive each suited to different applications or for operation with different types of motor. In this section descriptions are given of typical control systems for a range of different types of variable speed drive operating with a.c. motors

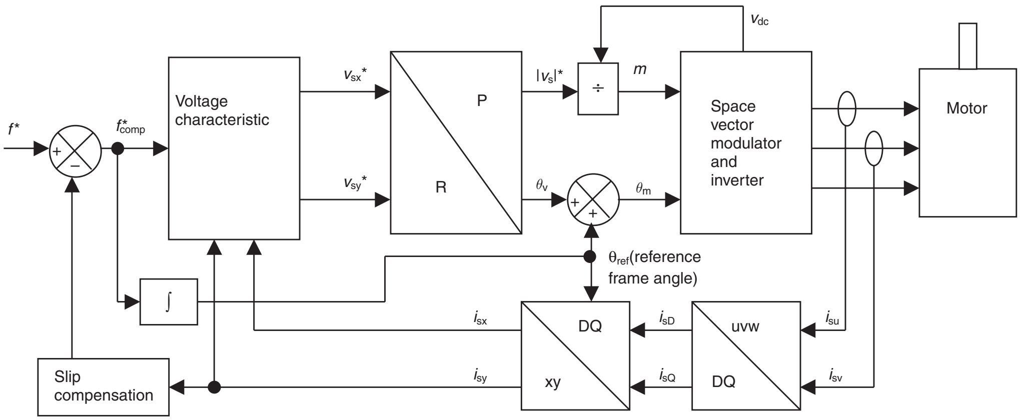


Figure 19.32 Open-loop drive control system

19.4.2.1 General purpose open-loop a.c. drive

The open-loop a.c. drive is a high power variable frequency voltage source. In its simplest form, the output frequency is defined by the user's reference and a suitable fixed frequency to voltage characteristic used to define the output voltage. Although this type of drive is normally designed to supply one or more induction motors connected in parallel, it can also supply other types of a.c. motor or it can be used as a variable frequency/variable voltage power supply. The following description relates to the operation of an open-loop a.c. drive with an induction motor.

Figure 19.32 shows a block diagram of a typical general purpose open-loop drive. For this type of control system feedback is required from the output current of at least two phases of the inverter and the voltage from the d.c. side of the inverter. These feedback signals can be derived from within the drive itself, and so no external feedback, such as the motor speed or position, is required. The control system allows a simple 'open-loop vector strategy' to be implemented. In common with all other a.c. drive control strategies described in this chapter the control system is based on a 'reference frame'.

Performance and applications The performance characteristics of the open-loop drive control system applied to an induction motor can be summarised as:

- Moderate transient performance.
- Full torque production down to approximately 3% of rated motor speed.
- Although a good estimate of stator resistance (Rs) improves torque production at low speeds, the control system will work with an inaccurate estimate, albeit with reduced torque. The stator resistance can be measured by the drive with a simple test.
- Although a good estimate of motor slip improves the ability of the drive to hold the reference speed, the control system will work with an inaccurate estimate, albeit with poorer speed holding. The motor slip depends on the rotor time constant of the motor (Tr) and this cannot be measured easily.
- No position or speed feedback is required from the motor shaft.

This type of drive is used in many applications where moderate performance is required and where providing position feedback would be unacceptable because of the environment or cost, or is simply not necessary. The following are some examples of applications where open-loop drives are used:

- Fans and pumps
- Conveyors
- Centrifuges

19.4.2.2 Closed-loop induction motor drive

The closed-loop induction motor drive, often referred to as a closed-loop vector drive, is used in many applications requiring better performance than an open-loop drive with an induction motor. To obtain the best performance with this type of drive position feedback is required from the rotor, but unlike the permanent magnet servo drive only the change of position and not the absolute position is required. The control system is similar to that used with a permanent magnet servo motor as shown below.

Performance and applications The performance characteristics of the closed-loop induction motor drive can be summarised as:

- Good dynamic performance at all speeds.
- Full torque operation down to standstill.
- A position feedback device that gives the incremental position of the rotor is required.
- Wide power range of motors available so that this type of drive can be used for applications requiring less than 1 kW up to more than 1 MW.
- Suitable for field weakening applications where motors can be operated up to many times base speed.
- The drive can be used in torque control, without the speed controller. The estimate of the flux position to align the reference frame is important as this affects the absolute level of torque produced for a given torque reference. The flux position calculation is dependent on an estimate of the rotor time constant which varies significantly with rotor temperature. However it is possible to include a rotor time constant estimator in the drive control system, so that the drive gives consistent torque control.

Closed-loop induction motor drives are used in many applications where good dynamic performance is required and especially where an induction motor drive is required

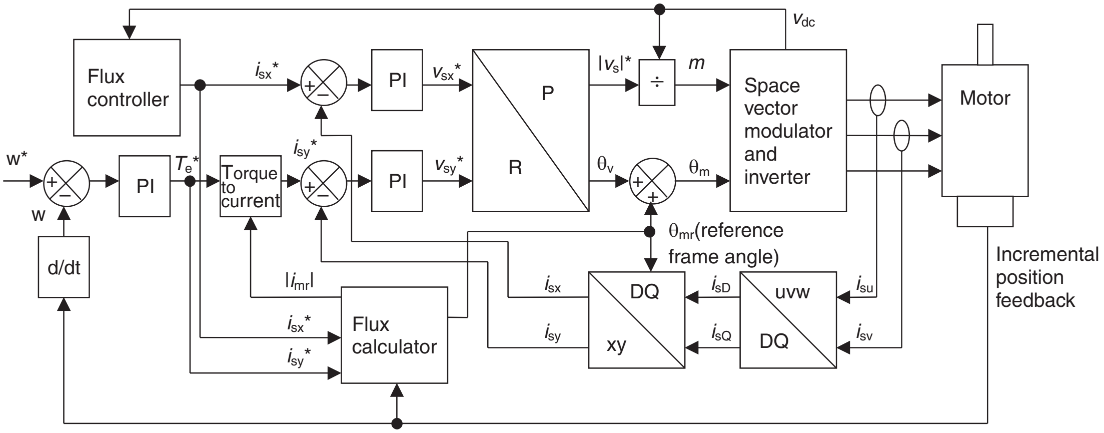

Flux controller
i_{sx}^*
PI
v_{sx}^*
P
R
$|v_s|^*$
÷
m
v_{dc}
Space vector modulator and inverter
Motor
w*
PI
T_e^*
Torque to current
i_{sy}^*
PI
v_{sy}^*
θ_v
θ_m
w
d/dt
θ_{mr}(reference frame angle)
$|i_{mr}|$
i_{sx}
DQ
i_{sD}
uvw
i_{su}
Incremental position feedback
i_{sx}^*
Flux calculator
i_{sy}
xy
i_{sQ}
DQ
i_{sv}
i_{sy}^*

Figure 19.33 Closed-loop induction motor drive control system

to give full torque at standstill. The following are some examples of applications where this type of drive is used:

- Cranes and hoists
- Lifts
- High speed spindle applications
- Material winding

Operation without position feedback The control system described above requires incremental rotor position feedback, but it is possible to implement the scheme without any physical feedback device. This can be done by estimating the rotor position from information available to the drive through the motor voltages and currents. One class of methods used to determine the rotor position, referred to as model based methods, use a model of the motor to calculate the rotor speed and hence the incremental rotor position. When a physical position feedback device is used the drive gives good dynamic performance and operates with full torque even at standstill. When a position estimator is used the dynamic performance is reduced and the minimum speed for full torque operation is similar to that of the open-loop drive. However a closed-loop induction motor drive without position feedback does have the following advantages and disadvantages when compared to an open-loop induction motor drive.

Advantages:

- Light load instability problems that can occur with an open-loop drive are eliminated.
- Torque control operation is improved.
- Starting with a spinning motor is faster.
- Fast closed-loop current control reduces trips under transient conditions.

Disadvantages:

- The motor model is normally dependent on the stator resistance and the rotor time constant. Incorrect estimates of these parameters can cause a significant reduction in performance and low speeds. Without real position feedback information it is difficult for the drive to compensate for variations in these parameters.

19.4.2.3 Permanent magnet servo drive

The permanent magnet servo drive is generally used for applications requiring high performance where motor shaft position feedback can be used. Because the rotor is not symmetrical this feedback must give absolute position within each electrical revolution of the motor. *Figure 19.34* shows the control system of a permanent magnet servo drive. The inverter control and reference frame transformation is the same as for the open-loop drive.

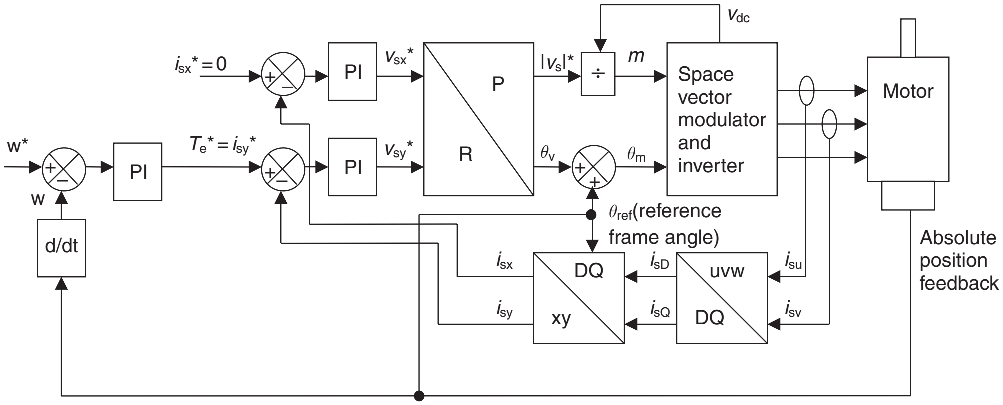

v_{dc}
$i_{sx}^* = 0$
PI
v_{sx}^*
P
R
$|v_s|^*$
÷
m
Space vector modulator and inverter
Motor
w*
PI
$T_e^* = i_{sy}^*$
PI
v_{sy}^*
θ_v
θ_m
w
d/dt
θ_{ref}(reference frame angle)
i_{sx}
DQ
i_{sD}
uvw
i_{su}
Absolute position feedback
i_{sy}
xy
i_{sQ}
DQ
i_{sv}

Figure 19.34 Permanent magnet servo motor drive control system

Performance and applications The performance characteristics of the permanent magnet servo drive can be summarised as:

- Good dynamic performance at all speeds.
- Full torque operation down to standstill.
- Permanent magnet servo motors usually have low inertia. Combined with a fast sample rate for the speed controller and fast torque control, this gives a speed controller with a very high bandwidth.
- Position feedback is required that gives the absolute electrical position of the motor.
- Permanent magnet motors exhibit an effect called cogging related to the geometry of the motor, which results in ripple in the motor torque. This effect can be minimised by good motor design, but can still be a problem.
- The control system described has a zero current reference in the x axis, and so the rotor flux cannot be altered by the drive. This limits the maximum speed of operation as the motor voltage increases with speed. Field weakening is possible, but due to the large effective air gap of a permanent magnet motor, a large amount of current is required to reduce the flux making the drive inefficient in the field weakening region. It is also necessary to limit the maximum speed with field weakening, because although the motor terminal voltages are reduced by reducing the flux whilst the drive is enabled, when the drive is disabled the voltages return to the level that would be produced without field weakening and may damage the power electronics in the drive.
- The drive can be used in simple torque control, without the speed controller shown in *Figure 19.33*.

The following are some examples of applications where permanent magnet servo drives are used:

- Machine tools where precise and dynamic performance is required.
- Pick and place applications where the requirements are less precise, but where rapid movements are required.
- More recently permanent magnet motors with high numbers of poles (e.g. 32 poles) have been in low speed applications such as direct (gearless) drives for lifts.

19.4.2.4 Four-quadrant operation

A.c. voltage source inverter drives are normally based on the power circuit shown below in *Figure 19.35*(*a*). This power circuit has the following drawbacks:

- Although the inverter allows power flow in either direction, the diode rectifier only allows power to flow from the supply into the drive. Therefore the chopper circuit must be used to dissipate unwanted energy in the resistor when braking power is fed from the motor to the drive.
- The currents taken from the supply generally contain significant harmonics.
- The voltage at the inverter input is theoretically limited to the peak line voltage of the supply. In practice this is reduced further by voltage drops in the converter. This limits the maximum output voltage without over-modulation to less than the supply voltage.

If the diode rectifier is replaced with an inverter as shown in *Figure 19.35*(*b*), the input currents can then be controlled to give almost sinusoidal waveforms with unity displacement factor. It is also possible for power to flow in either direction through the input inverter so that the drive gives full four-quadrant operation.

The input inverter can be controlled in a similar way to an inverter supplying a motor, using the reference frame based control system described in previous sections. A control system for the input inverter is shown in *Figure 19.36*.

Performance and applications The performance characteristics of the four-quadrant drive described here can be summarised as:

- Full four-quadrant operation.
- Approximately unity displacement factor and minimal input current harmonics.

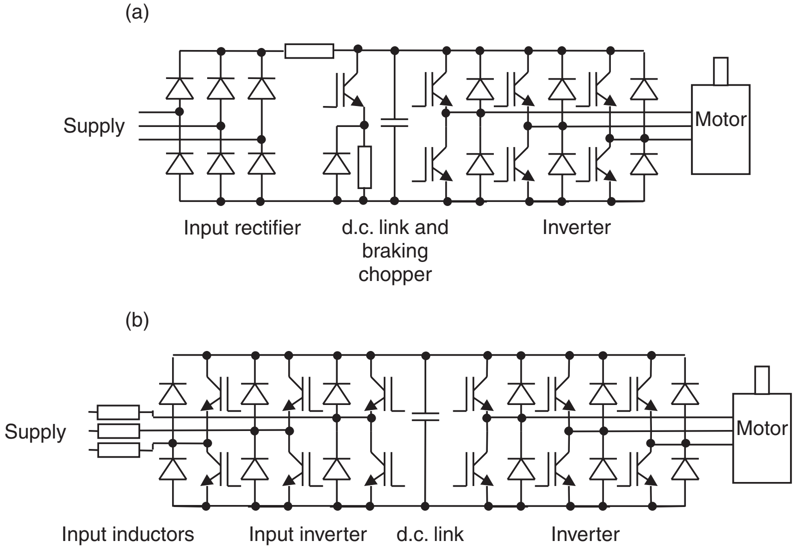


Figure 19.35 Alternative a.c. drive power circuits: (a) standard a.c. drive power circuit; (b) four-quadrant drive power circuit

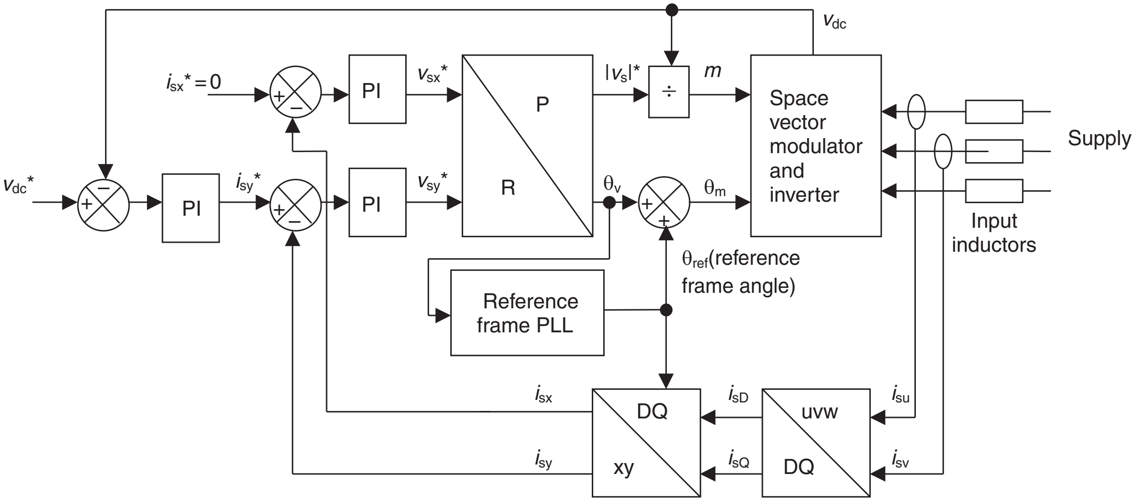

v_{dc}
$i_{sx}^*=0$
PI
v_{sx}^*
P
R
$|v_s|^*$
÷
m
Space vector modulator and inverter
Supply
v_{dc}^*
PI
i_{sy}^*
PI
v_{sy}^*
θ_v
θ_m
Input inductors
θ_{ref}(reference frame angle)
Reference frame PLL
i_{sx}
DQ
xy
i_{sD}
uvw
DQ
i_{su}
i_{sy}
i_{sQ}
i_{sv}

Figure 19.36 Input inverter control system

- Fast transient control of the d.c. voltage at the input to the motor inverter even during motor transient operation.
- The d.c. link voltage of a drive with an input diode rectifier is limited to less than the peak line supply voltage. With an input inverter the d.c. link voltage can be boosted to levels above this.
- No supply voltage feedback required.

Four-quadrant drives are used in applications where good quality input waveforms are required and/or significant braking energy can be returned to the supply. The following are some examples of applications where this type of drive is used:

- Engine test loading systems.
- Cranes and lifts.
- Cable laying winders.

19.4.2.5 Direct torque control

The drive control systems described so far all use space vector modulation to produce the inverter control signals. An alternative method of control that can be used when a motor is connected to the drive is direct torque control.

An example of a control system based on direct torque control is shown in *Figure 19.37*. The torque reference, T_e^*, and the flux magnitude reference, $|\varphi_s|^*$, are derived in the same way as for a drive using a space vector modulator. The hysteresis comparators then produce the required change in flux and torque, and the switching table selects the required inverter state to give a voltage vector that will change the flux and torque as required. The angle of the stator flux, $\alpha_{\varphi S}$, is used to determine which 60° sector the stator flux is in, as different areas of the switching table are used for different sectors.

The direct torque control system requires an estimate of the stator flux and the torque for the hysteresis comparators, and an estimate of rotor speed for the speed controller. These are derived from a model based estimator using the motor currents and an estimate of the motor voltages from the switching table state and the inverter d.c. input voltage, V_{dc}. It is important to note that although the principle of direct torque control appears simple and does not depend on estimates of motor parameters, the motor model used to derive

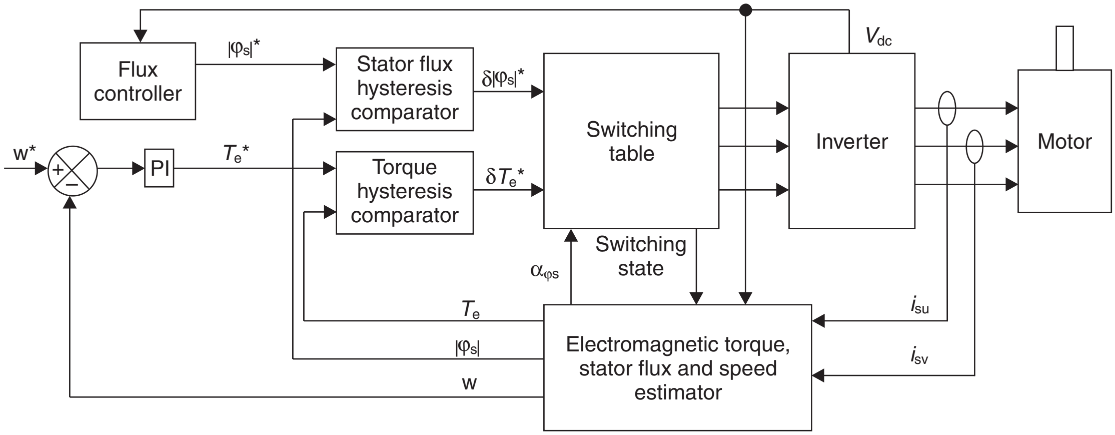

Flux controller
$|\varphi_s|^*$
Stator flux hysteresis comparator
$\delta|\varphi_s|^*$
Switching table
V_{dc}
Inverter
Motor
w*
PI
T_e^*
Torque hysteresis comparator
δT_e^*
$\alpha_{\varphi s}$
Switching state
T_e
$|\varphi_s|$
w
Electromagnetic torque, stator flux and speed estimator
i_{su}
i_{sv}

Figure 19.37 Direct torque control system

the estimates of torque, flux and speed is complex and is heavily dependent on the motor parameters. The following list gives a comparison between direct torque control and a control system based on a space vector modulator:

- The calculations for the current controllers, reference frame translation and space vector modulator are more complex than the direct torque control hysteresis comparators and switching table. However the sample rate required for direct torque control (typically 40 kHz) is much higher than that for a space vector modulator (6–12 kHz), because direct torque control uses a hysteresis method.
- As well as supplying a single induction motor, most induction motor drives can also supply more than one motor in parallel where the motors are different sizes, or the drive can be used as a general purpose variable frequency/variable voltage power supply. Direct torque control cannot be used in these applications, and so a direct torque control drive must also be able to operate with space vector modulation for these applications.
- Because direct torque control is based on hysteresis controllers the inverter has a continuously variable switching frequency.This is considered to be an advantage in spreading the spectrum of the audible noise from the motor, but the range must be controlled so that it does not exceed the maximum allowed by the power electronics of the inverter. Care must also be taken with direct torque control to ensure that changes from one voltage vector to another more than 60° away do not occur repetitively as this can increase the stress on motor insulation.
- A direct torque control drive inherently delivers a change in torque in the shortest possible time within the limits of the sample rate. Because of the sampling and calculation delays usually associated with space vector modulator based systems a change in torque can take several samples. However, deadbeat type algorithms can be used with a space vector modulator systems giving performance which compares well to that of direct torque control.

19.5 Applications and drive selection

19.5.1 General

In order to successfully select and apply the optimum drive system, it is necessary to understand the essential features of both the alternative drive technologies and the load to be driven. The following listing of common loads could prove useful when selecting a drive:

19.5.1.1 Metals industries

Drive duty	*Rating range*	*Comments/ drive type*
Rolling Mill	Up to 1000s of kW	High impact torque loading, constant kW speed range, steel works specification and difficult environment. Closed loop induction motor drives and d.c. drives.
Strip Mill	Up to 100s of kW	Normal torque loading (150% max) constant kW speed range, steel works specification and difficult environment. Closed loop induction motor drives and d.c. drives.
Slitters and perforators	50 to 150 kW	In metal finishers plant, environment and specification easier—normally IP23 enclosure, forced ventilation with filter is acceptable. Drive often integrated with wind/unwind stand drives. Closed loop induction motor drives and d.c. drives.
Wind/unwind	Up to 200 kW	Constant kW rating over build-up range. Regenerative braking with four-quadrant operation. Steel works specification and difficult environment. In metal finishers plant easier conditions apply as above. Closed loop induction motor drives and d.c. drives.
Tube mill	Up to 300 kW	Constant kW rating over pipe diameter range. Usually 4-quadrant with regenerative braking. Environment can be difficult with oil spray present. Closed loop induction motor drives and d.c. drives.
Cast tube spinner	20 to 50 kW	High values of acceleration and deceleration torque required. 4-quadrant regenerative. Typical 4–6 speeds required: clean, spray, fill, spin 1 and spin 2. Difficult environment. Single pipe vent or box enclosed motor with filtered air supply. Closed loop induction motor drives and d.c. drives.
Machine tool spindle	Up to 150 kW 5 to 30 kW typical	Mostly flange mounting and timer belt drive to 30 kW gearbox coupled above. Always field range control, reversible, 4-quadrant drive. Often with encoder for spindle orientation. Forced-vent with filter; often coaxial fan unit to 60 kW. Closed loop induction motor drives. At low powers permanent magnet servo drives and at very high powers d.c. drives are used.

Drive duty	*Rating range*	*Comments*
Machine tool table	Up to 200 kW 20 to 75 kW typical	Foot or flange mounting gearbox otherwise as above. Permanent magnet servo drives. At higher powers closed loop induction motor drives and d.c. drives are used.
Wire drawing	5 to 75 kW	Constant kW speed range with individual motor field control on multi block drives with single controller. Progressive speed increase between heads as wire diameter reduces and speed increases. Dancer arm tension control between heads and tension controlled winder take off. Forced-vent motors require filter against wire end entry. 4-quad. Acceleration/deceleration. Closed loop induction motor drives or at higher powers d.c. drives.

19.5.1.2 Plastics

Drive duty	*Rating range*	*Comments*
Extruder	5 to 400 kW	Constant torque drive with high torque required to start a stiff extruder screw. Environment can be difficult with plastic particles and fume risk. Single pipe vent of motor is advisable. Closed loop induction motor drives and d.c. drives. Open loop induction motor drives may be used in some applications.
Sheet line Reeler	1.5 to 15 kW	Constant kW drive over reel build-up ratio but often sized as constant torque drive. Braking usually mechanical. Environment can be difficult. TEFC IP55 used for low kW ratings. Closed loop induction motor drives and d.c. drives.

19.5.1.3 Rubber

Drive duty	*Rating range*	*Comments*
Banbury Mixer	Up to 1000 kW	Very heavy peak duty, duty cycle rated typically: 250% full load torque for 10 s, 150% for 20 s, 100% for 120 s, 10% for 30 s repeating continuously. Difficult environment with particle rubber and carbon black, single and double pipe vent is usual with some CACA and CACW motors used. Check through Banbury manufacturer's drive specification—safety environment. D.c. drives predominate but closed loop induction motor drives growing in use.
Callender	Up to 500 kW	Environment as above, with easy duty, constant torque, but 200% dynamic braking. Check through manufacturers' drive specification—safety involvement. Closed loop induction motor drives and d.c. drives.

19.5.1.4 Chemical

Drive duty	*Rating range*	*Comments*
Mixer	Up to 150 kW	Generally constant torque but could be rising torque requirement with increasing mix stiffness. Often explosion proof or hazardous location enclosure requirement; environment can be difficult. Closed loop induction motor drives and d.c. drives.
Extruder	Up to 100 kW	Usually constant torque often explosion proof or hazardous location enclosure requirement. Environment can be difficult. Closed loop induction motor drives and d.c. drives.
Stirrers and agitators	Up to 400 kW	High reliability requirement to avoid loss of mix through drive shutdown. Often rising torque with mix stiffness. Energy saving of importance as process can occupy many days. Drive often outdoor mounting with CACA weatherproof enclosed motor. In particularly exposed positions motor and gearbox should have additional protection. Closed loop induction motor drives and d.c. drives.

19.5.1.5 Materials handling

Drive duty	*Rating range*	*Comments*
Conveyor	0.5 to 20 kW	Cascading of multiple drives can be a requirement with progressive speed increase or synchronised drives. Constant torque application with dynamic or regenerative braking. Open loop induction motor drives predominate unless sychronisation/close coordination required where closed loop induction motor drives are used.
Automated warehousing	As above	Usually 3 axis systems, constant torque 4-quadrant 150% full load torque at starting to duty cycle rating. Closed loop induction motor drives predominate. D.c. drives can be used and in less demanding applications open loop a.c. is used.

19.5.1.6 Hoist & Crane

Drive duty	*Rating range*	*Comments*
Lift and hoist	30 to 100 kW	Four-quadrant with 200% full load torque at starting. Starting duty 90–200 starts/hour. Smooth, quiet response very important. Closed loop induction motor drives are widely used. D.c. drives are used. Open loop drives are applied to non safety critical applications and installations where smooth ride quality is not critical.
Crane	3 to 75 kW	Four-quadrant high torque requirement. Sometimes weatherproof enclosure. Operating duty cycle requires evaluation. As for lift and hoist.

19.5.1.7 Concrete pipe manufacture

Drive duty	*Rating range*	*Comments*
Pipe spinner	10 to 100 kW	D.c. motor needs particular protection against water, cement and vibration. Four-quadrant multi-speed drive requirement with regenerative braking. Duty cycle requires evaluation. Closed loop induction motor drives can also be used.

19.5.1.8 Fans and blowers

Drive duty	*Rating range*	*Comments*
Axial flow fan	0. 5 to 40 kW	As cage motors specially adapted for air stream use with impeller on motor shaft. Existing motors will be retained under inverter control and may require slight de-rating or slightly reduced max speed. Inverse cube law relationship between fan kW load and speed. Noise falls as the fifth power of fan speed. Open loop induction motor drives are most widely used.
Centrifugal fan	0.5 to 500 kW	Cube law kW/speed relationship extend acceleration time on large fans to moderate starting current requirement. Power saving important on large fans as is top speed fan noise. Open loop induction motor drives are most widely used.
Rootes type blowers	3 to 200 kW	Positive displacement blowers are constant torque and kW loading is linear with speed into a fixed system resistance. Load pulsates heavily. Rootes blowers are noisy but easily started. Power saving can be important. D.c. drives and closed loop induction motor drives predominate. Open loop induction motor drives can be used with care.

19.5.1.9 Pumps

Drive duty	*Rating range*	*Comments*
Centrifugal pumps	0.5 to > 500 kW	Early cage motors with ‘A’ and ‘E’ class insulation require care over winding temperature under inverter control. Power saving on large drives important. Open loop induction motor drives predominate.

19.5.1.10 Paper and tissue

Drive duty	*Rating range*	*Comments*
Paper machine and pumps	Up to 500 kW	Environment difficult with water, steam and paper pulp present. Pipe vent motors common. Often non standard a.c. and d.c. motors. Usually closely co-ordinated drives in a paper line. Closed loop induction motor drives and d.c. drives.
Winders and reelers	5 to 100 kW	Constant kW range over build-up range. Four quadrant operation with regenerative braking. IP23 motor enclosure with filter is common. Closed loop induction motor drives and d.c. drives.

19.5.1.11 Printing

Drive duty	*Rating range*	*Comments*
Printing press	Up to 200 kW	Some special co-axial motor designs for series connection on line. Field weakening for wide speed range. 4-quadrant with slow ramp acceleration and inch/crawl control plus emergency stop. Pipe vent where ink fumes may be a hazard. Closed loop induction motor drives and d.c. drives.
Folders, unwind & rewind stands	Up to 100 kW	Often integrated in printing line drive with press drive and unwind stand drive under ‘master’ control. Otherwise as above. Closed loop induction motor drives and d.c. drives.

19.5.1.12 Packaging

Drive duty	*Rating range*	*Comments*
Boxing, stamping, folding, wrapping	Up to 75 kW	Mostly 4-quadrant with slow ramp acceleration with inch/crawl and E/stop. Often integrated line control. PM Servo drives are widely used in precision packaging machines. Closed loop induction motor drives and some d.c. drives are also used.

19.5.1.13 Engineering industries

Drive duty	*Rating range*	*Comments*
Test rigs of many types	Up to >15 MW	Test rig drives require careful engineering. Often high speed with fast response, accurate speed and torque measurement, usually 4-quadrant with field weakening control. Engine test rigs require special knowledge of throttle control drive/absorb changeover and power measurement. Drive control/monitoring particularly important. Closed loop induction motor drives and d.c. drives. PM servo drives are also used for precision applications.

19.5.1.14 Wire and cable

Drive duty	*Rating range*	*Comments*
Bunchers and stranders	10 to 150 kW	Generally multiple drives with cage or bow, capstan plus take up drives under integrated control. Constant torque except take-up with 4-quadrant acceleration/deceleration with inch/crawl/E-stop controls. Motors require filter protection against metal dust entry. Closed loop induction motor drives and d.c. drives.
Capstan	5 to 100 kW	As above.
Take-up and unwind stands	5 to 50 kW	As above but constant kW over build-up ratio.
Extruders	5 to 150 kW	See extruders under plastics industry but control often be integrated in cable line drives.
Armourers	10 to 150 kW	As Buncher/Strander drive above.
Caterpillars	1.5 to 30 kW	Constant torque duty and low kW rating in view of low haul off cable speeds. Often integrated in cable line drives. Motor protection generally no problem. Closed loop induction motor drives and d.c. drives.

19.5.1.15 Hydraulics

Drive duty	*Rating range*	*Comments*
Pump and motor test rigs	Up to 250 kW	Hydraulic fluid is a contamination risk. Pipe vent often used. Generally constant torque to medium/high speeds with 4-quadrant drive. Speed torque and power measurement often required with full drive monitoring on endurance rigs. Closed loop induction motor drives and d.c. drives.

19.5.1.16 Electric motors and alternators

Drive duty	*Rating range*	*Comments*
A.c. and d.c. motors/ generators/alternators test bed rigs	Up to >15 MW	All rotating electrical machine manufacturers have elaborate test bed rigs, supplying their own rotating machines and obtaining control systems from the drives industry to their own requirements. Closed loop induction motor drives and d.c. drives.

19.5.1.17 Textiles

Drive duty	*Rating range*	*Comments*
Ring frame machines, carding machines, looms	Up to 150 kW	Difficult environment in which IP55 enclosure has become a standard; ring frame Schrage and d.c. thyristor drive. Use pipe ventilation. All drives constant torque 4-quadrant for speed modulation (ring frame) or best speed holding accuracy with slow ramp acceleration/deceleration on carding drives. Special a.c. cage loom motors are high torque, high slip designs. Today a.c. inverter drives predominate since their characteristics are particularly suitable.

19.5.1.18 Foods, biscuit and confection

Drive duty	*Rating range*	*Comments*
Extruder	5 to 400 kW	Hose proof motors for plant cleaning. Continuous production requiring high levels of reliability, control and monitoring. Otherwise as plastics industry extruders.
Mixer	5 to 150 kW	As above and see chemical industry mixer drive.
Conveyors	0.5 to 120 kW	As above and see material handling industry conveyor drive.

19.6 Electromagnetic compatibility

Electromagnetic compatibility of electrical drive system is a complex but vitally important subject. By their very nature by rapidly switching voltages and/or currents, drives can form very effective noise generators. The design of the drive must therefore be very carefully considered from the early stages firstly to ensure that the drives own controls are not interfered with and secondly that the drive can be applied in a system without adverse effects on other equipment.

Various national and international standards exist for EMC—few have been written with power electronic drives in mind. Drive manufacturers are now making great efforts to design drives compliant with the existing and future standards—they should be consulted for advice on appropriate standards and compliance thereto.

Bibliography and further reading

The documents listed here have been selected to provide the reader with useful sources of information and further reading relating to Electrical Variable Speed drives and their application.

1 MOLTGEN, G., '*Converter Engineering*', John Wiley, ISBN 0-471-90561-5 (1984) (A reference for fundamental power converter operations and relationships)
2 CHALMERS, B. J., '*Electric Motor Handbook*', Butterworths, ISBN 0-408-00707-9 (1988) (A practical reference book covering many aspects of characteristics, specification, design, selection, commissioning and maintenance)
3 Vas, P., *Sensorless Vector and Direct Torque Control*, Oxford University Press, ISBN 0-198-56465-1 (1998) (General background to the theory of vector control of motors)

20

Motors and Actuators

M G Say PhD, MSc, CEng, FRSE, FIERE, AGCI, DIC
Formerly of Heriot-Watt University

J F Eastham

Contents

20.1 Energy conversion 20/3

20.2 Electromagnetic devices 20/3
  20.2.1 Electromagnets 20/3
  20.2.2 Tractive electromagnets 20/3
  20.2.3 Actuators 20/4
  20.2.4 Lifting magnets 20/5
  20.2.5 Crack detectors 20/6
  20.2.6 Separators 20/7
  20.2.7 Clutches 20/8
  20.2.8 Couplings 20/8
  20.2.9 Brakes 20/9
  20.2.10 Magnetic chucks 20/9
  20.2.11 Vibrators 20/10
  20.2.12 Relays and contactors 20/10
  20.2.13 Miniature circuit-breakers 20/12
  20.2.14 Particle accelerators 20/13

20.3 Industrial rotary and linear motors 20/15
  20.3.1 Prototype machines 20/15
  20.3.2 D.c. motors 20/17
  20.3.3 Three-phase induction motors 20/24
  20.3.4 Three-phase commutator motors 20/29
  20.3.5 Synchronous motors 20/31
  20.3.6 Reluctance motors 20/31
  20.3.7 Single-phase motors 20/32
  20.3.8 Motor ratings and dimensions 20/35
  20.3.9 Testing 20/35
  20.3.10 Linear motors 20/40

20.1 Energy conversion

Electromagnetic machines convert electrical into mechanical energy in devices with a limited stroke (actuator, brake, relay etc.) or continuous angular rotation (motor), or linear motion (linear motor).

Mechanical energy involves a force f_m acting over a distance x or a torque M_m acting over an angular displacement θ. Electrical energy involves the displacement of a charge q (a current i for a time t) through a potential difference (p.d.) v. The energies W and corresponding powers $P = \mathrm{d}W/\mathrm{d}t$ are

Mechanical:

$$W_m = f_m x \quad P_m = f_m(\mathrm{d}x/\mathrm{d}t) = f_m u \qquad \text{(linear)}$$

$$W_m = M_m\theta\varsigma \quad P_m = M_m(\mathrm{d}\theta/\mathrm{d}t) = M_m\omega_r \quad \text{(rotary)}$$

Electrical:

$$W_e = vq \qquad P_e = v(\mathrm{d}q/\mathrm{d}t) = vi$$

where $u = \mathrm{d}x/\mathrm{d}t$ is the translational speed and $\omega_r = \mathrm{d}\theta/\mathrm{d}t$ is the rotational speed. In an electromagnetic machine the basic physical conversion mechanism between the two forms of energy is the magnetic field, a characteristic property of electric current. The elements of electromagnetic/mechanical conversion are set out in Sections 2.4.1 to 2.4.3.

20.2 Electromagnetic devices

20.2.1 Electromagnets

Electromagnets for stroke-limited devices (e.g. actuators) are such that estimation of the flux distribution in the air gap (the working region) is difficult. The total magnetomotive force (m.m.f.) produced by a current i in an N-turn coil is $F = Ni$.

20.2.1.1 Coil windings

Most coils for magnetic-circuit excitation are wound by one of the following (usually automated) methods: (1) on a former with end-cheeks; (2) on a bobbin that forms an integral part of the coil and comprises a moulded or fabricated construction of a suitable insultant; or (3) by a winding machine that feeds insulated wire into a self-supporting form, with an epoxy-resin binder.

The coil design is based on the provision of the required m.m.f. for a specified voltage (or current), with an acceptable coil temperature rise on a specified duty cycle.

D.c. excitation For direct current (d.c.), the current i at a coil terminal voltage v is determined by the coil resistance $R = V/i = \rho L_{mt}N/a$, where the N turns have a mean turn length L_{mt} and the conductor, of resistivity, ρ, has a cross-sectional area a. Then

$$a = \rho L_{mt}(Ni)/V = \rho L_{mt}F/V$$

for a total m.m.f. F. The current cannot be determined until the cooling conditions are established. Let the conductor current density be J, so that $i = Ja$; then $N = F/Ja$. The total conducting cross-section of the coil is Na and the gross cross-sectional area of the wound coil is Na/k, where k is the *space factor*.

The power taken by the coil is $P = Vi$, and the consequent temperature rise on continuous operation is $\theta_m = P/cS$. Here c is a cooling coefficient representing the power dissipation per unit of the coil surface area S per degree Celsius rise of surface temperature above ambient. The value of θ_m for continuously rated coils is usually specified. On intermittent or short-time rating the rise is a function of the thermal capacity of the coil.

A.c. excitation For alternating current (a.c.) the current i at voltage V is determined by the coil impedance $Z = R + j\omega L$ at angular frequency ω. An a.c. coil therefore tends to have fewer turns than one for d.c. Further, the coil inductance L varies widely, depending on the saturation of the ferromagnetic parts and in particular on the length of the air gap. A wide gap increases the magnetic reluctance and reduces L, but as the gap length reduces (e.g. by movement of the working parts) the inductance rises. If $\omega L \gg R$, as is usual, the root-mean-square (r.m.s.) value of the m.m.f. approximates to $F = VN/\omega L$ with L estimated for the range of air gap lengths.

In practice, performance is based on data obtained on test. A particular feature is the double-frequency fluctuation of the mechanical force, which produces a characteristic 'chatter' in the closed position of the device; this may have to be mitigated by means of a shading ring.

20.2.1.2 Coil design

Space factor A simple coil wound from a circular-section wire of diameter d, and insulated to a diameter d_i, will pack down in a manner that is affected by the method of winding, one layer partly occupying the troughs in the layer beneath it; the space factor may then approximate to $k = 0.85(d/d_i)^2$. Conductors of small diameter bed less effectively, and the space factor is reduced.

Cooling coefficient A typical value of the cooling coefficient c is 0.075 W/m^2 per °C above ambient. However, cooling conditions vary widely with the efficacy of ventilation.

20.2.1.3 Operating conditions

Whether d.c. or a.c. excited, the current in an operating coil is affected by that movement of the working parts that closes or opens the air gap. Let a quiescent spring-loaded relay (*Figure 20.1*(*a*)) in the open position be connected to a direct source voltage V. The coil current begins to rise exponentially, but the armature does not move until the magnetic force exceeds the spring restraint. Thereafter, the shortening gap increases the coil inductance, setting up a counter electromotive force (e.m.f.) and checking the current rise and the attracting force. Finally, the armature reaches the closed position at the end-stop, dissipating kinetic energy in noise, bounce and mechanical deformation. The sequence of events, in terms of the gap length x, armature speed u, coil current i and time t, is shown in *Figure 20.1*(*b*).

If the coil is energised from an a.c. source there are two further effects: the closing time depends on the instant in the cycle at which the voltage is applied and (more importantly) the operating force fluctuates. Ferromagnetic parts must be laminated to prevent excessive core loss and the counter-effects of eddy currents. Suitable sheet steel for the purpose has a core loss of less than 5 W/kg.

The force fluctuation can be reduced (but not eliminated) by a *shading* ring (*Figure 20.2*) embedded in one of the pole faces flanking the gap. Currents induced in the ring delay part of the pole flux. Thus the combination of shaded and unshaded flux gives a resultant that still fluctuates but does not at any instant fall to zero.

20.2.2 Tractive electromagnets

Two forms of tractive electromagnet are shown in *Figure 20.3*. Type (*a*) usually has cylindrical poles, sometimes with shouldered ends to retain the coils, a rectangular yoke to which the

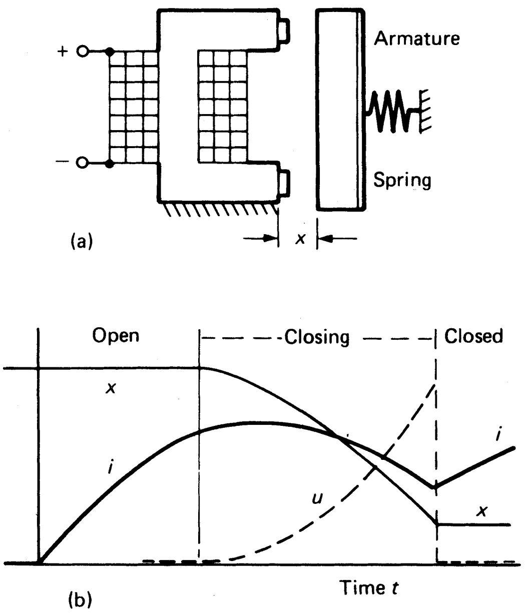


Figure 20.1 Operation of a d.c. relay

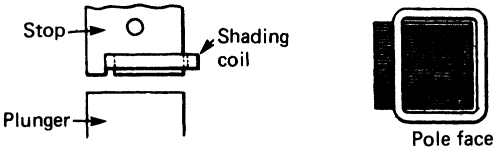


Figure 20.2 Shading coil or ring

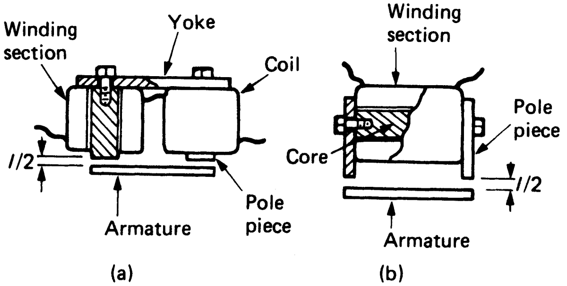


Figure 20.3 Tractive electromagnets

poles are screwed or bolted, and a rectangular armature. Two exciting coils are used; they are connected to give opposite polarities at the respective pole ends. Type (*b*) has a single coil mounted on a cylindrical core to which the rectangular pole-pieces are attached. In both cases the total air gap length is the sum of the gaps at the respective poles; in some designs, however, the armature is hinged to the pole-piece at one end. In this case the free end forms the major gap.

In (*a*) let each polar surface have an area of 250 mm^2 and be required to exert a total force of 1.0 N on the armature when both gaps are 3.0 mm long. Then with *d.c. excitation*, $f = 400\,000B^2 A$ giving $B = 70$ mT, for which $H = 57\,000$ A-t/m. For a total gap of 6 mm the gap m.m.f. required is 340 A-t. Adding 10% for the iron circuit and 25% for leakage, the total excitation required is about 450 A-t, from which the coil design follows.

With *a.c. excitation* it is necessary to estimate the inductance in the open and closed positions, and to adjust the number of turns for a given operating voltage so that adequate force is available. The change of magnetic flux between the two extreme armature positions is very much less than with d.c. operation, so that for the same (average) force in the open position, that in the closed position is only a little greater.

20.2.3 Actuators

20.2.3.1 D.c. actuators

Three typical arrangements for d.c. actuators are shown in *Figure 20.4*. Form (*a*) is convenient for small devices as the frame can be bent from strip; it is common for overcurrent and undervoltage relays. Form (*b*) may have a cast frame, and provides parallel flux paths through the iron. In (*c*) the cylindrical iron circuit presents a low reluctance, the circuit being completed by a lid attached by studs or screwed into the cylindrical body.

The iron *end-stop* should project well into the coil to improve flux concentration. It may be integral with the frame or screwed into it (in which case it can be used to locate and secure the operating coil). The plunger passes through the frame at the *throat*, the reluctance of which can be reduced by minimising the annular gap and extending the effective axial length, as shown at (*b*) and (*c*).

A typical iron clad actuator in part section is shown in *Figure 20.5*. With the dimensions $a = 220$ mm, $d = 65$ mm, $x = 63$ mm and $y = 450$ mm, the coil may develop about 15 kA-t to give a pull of 400 N across a 25 mm gap in the open position. The brass pin forms a stop, and cushions the plunger by expelling air through the vent.

With a flat-ended plunger the stroke is equal in length to the magnetic air gap. Maximum work (force × displacement) occurs with a short stroke. By using a coned plunger (see *Figure 20.5*), maximum work is obtained with a longer stroke. If the cone angle is 60°, the comparable stroke is twice that for a flat-ended plunger for about the same magnetic pull. It is possible to obtain a wide variety of characteristics by modifying the shapes of the stop and plunger ends.

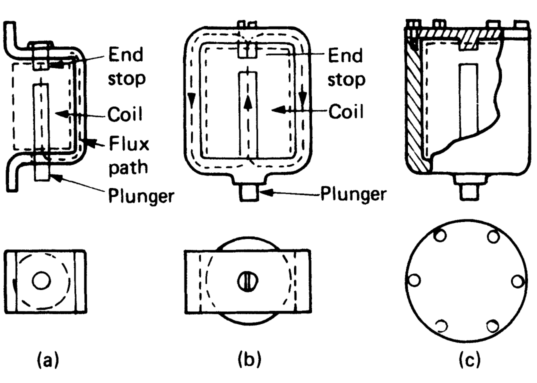


Figure 20.4 D.c. actuators

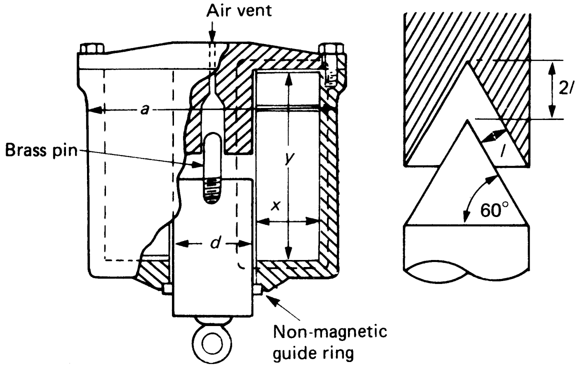


Figure 20.5 Iron clad 'pot' actuator

20.2.3.2 A.c. actuators

A common arrangement for a single-phase device is that shown in *Figure 20.6*. The E-type laminations are clamped. In the plunger, rivets should lie in a line in the flux direction to minimise eddy currents. To keep down the 'holding' current the plunger and stop ends should be flat.

Because of the many variables concerned, the design is complicated. An empirical rule is to allow 1.5 mm^2 of plunger cross-section for every 1 N of force; this corresponds to a peak flux density of 0.8 T in the laminations. The size of the coil (and therefore the main dimensions) may be taken as having a length 2.5–3 times the stroke and a depth equal to the stroke. The number of turns N is estimated from

$$N = \mathcal{V}/4.4 f B_{\mathrm{m}} A$$

where $B_{\mathrm{m}}A$ is the peak flux and f is the frequency. Final adjustment of N is made on test; it is reduced if the force is too low.

20.2.3.3 Polyphase actuators

A three-phase actuator has three limbs. Because of the phasing, the net force on the laminated bar armature assembly is never zero, and shading is not necessary.

The typical unit (*Figure 20.7*) is for operating a brake. It has three limb-coils E connected in star. The armature A is shown in the lifted (energised) condition. The plunger rod, fitted with a piston in the dashpot D, cushions the end of the stroke. A valve in the piston allows unretarded drop-out for quick brake application.

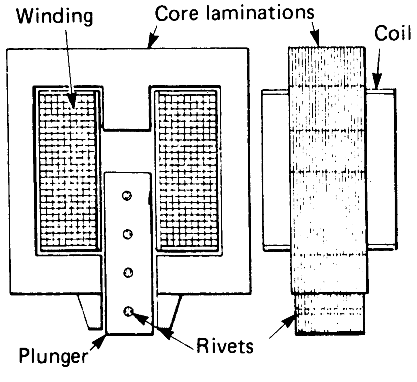


Figure 20.6 Single-phase actuator

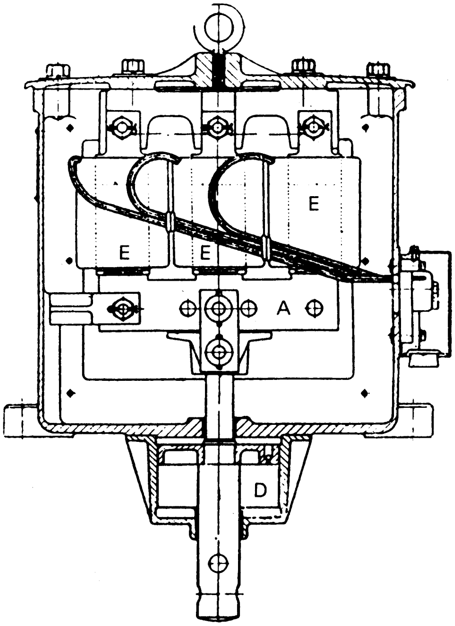


Figure 20.7 Three-phase actuator

20.2.4 Lifting magnets

Lifting magnets are of use in the handling of iron and steel, as they dispense with hooks and slings. The maximum load of a magnet varies with the material to be lifted. A magnet capable of lifting 1 t of scrap may raise a 20 t load in the form of a thick solid piece with a flat upper surface. As the excitation is limited by temperature rise of the coil, the lifting capacity is also dependent on the duty cycle. For the comparatively arduous conditions normally ruling in industrial use, a robust and weatherproof construction is essential.

20.2.4.1 Circular magnets

The essential features of a circular magnet are shown in *Figure 20.8*. As the magnetic properties of the material lifted and the air gaps between the magnet poles and the material are both arbitrary and subject to wide variation, the design of a lifting magnet is generally based on thermal considerations. A given carcass and winding are assigned an empirically derived power rating such that the temperature rise of the coil is not excessive. The designer's aim is then to secure the maximum effective ampere-turn excitation and working flux density by adjustments of iron and conductor materials and heat dissipation. Allowances in design must be made for the development of adequate pull under conditions of low line voltage (e.g. 80% or less of nominal), and high conductor resistivity when hot.

The majority of lifting magnets, except those of small size, have a winding of flat strip, which is more adaptable than wires of circular section to the attainment of a good space factor with the large conductor areas generally necessary. Aluminium is sometimes employed in preference to copper for the advantage of weight economy: the weight of a magnet is

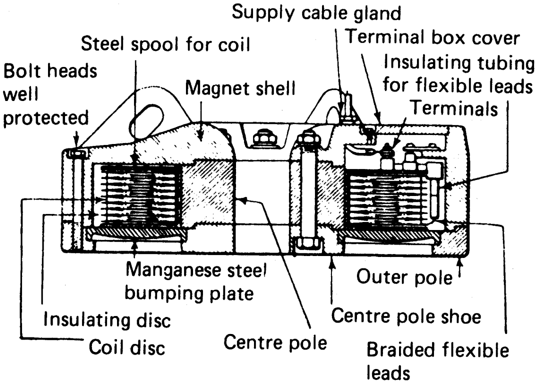


Figure 20.8 Circular lifting magnet

important as it represents a useless additional load to be moved every time its crane is operated. The winding in *Figure 20.8* is shown diagrammatically: it comprises a number of flat spirals with heat-resistant insulation.

The general dimensions are such that the diameter of the inner pole-face is about one-third of the overall diameter d. The load lifted is proportional to d^2 and the power rating (in kilowatts) is of the order of $4d^2$ (with d in metres). As regards the load, only in exceptional cases does close and uniform contact occur between magnet and load surfaces. The tendency for flux concentration over small contact regions develops local saturation and increases the effective gap.

20.2.4.2 Rectangular magnets

Materials of regular shape, such as sheets, bars, pipes, etc., are well suited to lifting by rectangular magnets. The general construction is similar to that of the circular type, the body being formed of a box-shaped steel casting with a central projection to give the inner polar surface.

The approximate lifting capacities of circular and rectangular magnets are given in *Table 20.1.*

20.2.4.3 Control

Simple on/off switching is not practicable because of the high level of stored magnetic energy. The general control features needed are: (1) discharge resistors connected across the winding just prior to disconnection to reduce contact arcing and limit inductive e.m.f; (2) auxiliary resistors introduced into the coil circuit after a predetermined time to limit coil temperature rise; and (3) reversal of polarity at a low current level to overcome remanence and so release small pieces such as turnings or scrap.

20.2.5 Crack detectors

Electromagnetic crack detection depends on the fact that, in magnetic material, the magnetic susceptibility of a fault is markedly inferior to that of the surrounding material. The success of the whole technique of magnetic crack detection depends largely on the care taken to ensure correct strength and direction of magnetisation. The following methods are used:

(1) *Needle method.* The surface to be tested is explored with a small magnetic needle. This needle carries a pointer which moves over a scale, with a right and left motion as the needle turns on its pivot to align with the field distortion passing beneath it in the direction of the arrow. Thus the fault is detected. The sensitivity is increased by using a mirror and light beam.

(2) *Powder method.* The part to be tested, previously cleaned, is laid across the arms of the machine, and the circuit-

Table 20.1 Approximate lifting capacities

Circular magnets: *load lifted* (t)

Material	*Magnet diameter* (m)				
	1.6	*1.4*	*1.2*	*1.0*	*0.6*
Skull-cracker ball	18	15	10	7	3
Slabs	27	23	16	9	3
Pig-iron	1.3	1.0	0.6	0.3	0.1
Broken scrap	0.8	0.5	0.4	0.3	0.1
Cast-iron borings	1.0	0.7	0.5	0.3	0.1
Steel turnings	0.5	0.3	0.2	0.1	0.005

Rectangular magnets: *plate area lifted* (m^2)

Plate stack			*Magnet dimensions* (m)			
Plate thickness (mm)	*Longest plate* (m)	*Maximum No. of plates in stack*	*0.6 × ~~0.~~4*	*1.0 × ~~0.~~4*	*1.4 × ~~0.~~4*	*2.0 × ~~0.~~4*
0.4	1.5	80	0.9	2.3	3.5	4.6
1	2.8	20	1.8	4.3	6.5	8.7
3	4.2	10	2.4	5.5	8.3	11
6	6.7	5	2.8	6.5	9.7	13
12	9.5	3	3.2	7.5	11.3	15
25	13.5	2	3.2	7.5	11.3	15

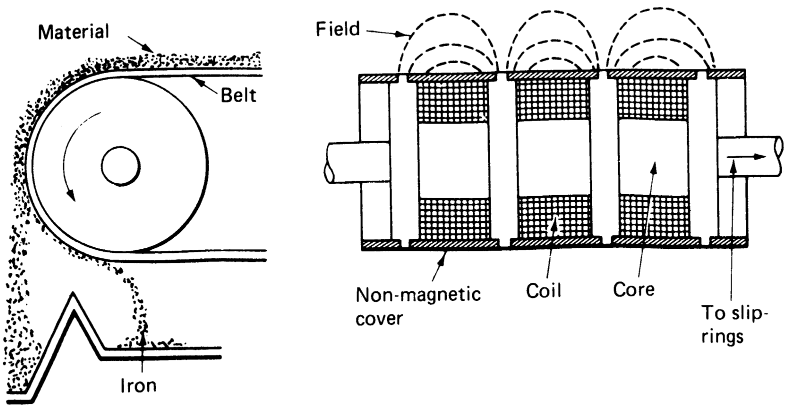


Figure 20.9 Magnetic pulley separator

closing push-button switch depressed and released quickly. The article is then removed and sprinkled with special powder, the excess of which is blown away or shaken off; it will then be found that the defects are clearly indicated by the magnetic patterns.

(3) *Fluid method.* This resembles the powder method but employs a fluid (e.g. paraffin) containing finely divided magnetic material in suspension.

Each of these techniques can be applied to the detection of cracks or other flaws in parts which have been magnetised. There are two methods of attaining this magnetisation in normal commercial use.

In the first method the part to be tested is placed between the poles of an electromagnet, in which case the direction of the field is from pole to pole. The second method utilises the fact that a concentric magnetic field forms round an electric current. A heavy low-voltage current is passed through the part itself, or through a current-bar adjacent to it or threaded through it.

As only those cracks or flaws will be shown up which cut across the magnetic field, it will readily be understood that the first method is most suited to the detection of transverse cracks, the second to the detection of longitudinal ones. However, apparatus designed for testing by means of the second method may be adapted to the detection of transverse cracks by encircling the part with several turns of cable through which the heavy current is passed.

20.2.6 Separators

The bulk handling of material, particularly where the process involves crushing or grinding, may require the use of a magnetic separator for removing unwanted or tramp iron and steel, or for quickly separating ferrous from nonferrous scrap metals. Successful operation depends on uniformity of the feed thickness, and often an installation must include a suitable conveyor/feeder.

20.2.6.1 Types

Magnetic pulley This form of separator (*Figure 20.9*) comprises a number of circular cores and poles, the magnetic axis being that of the shaft. Coils encircle the cores, with d.c. (or rectified a.c.) excitation, and set up a magnetic field pattern. Iron attracted to the pulley surface is removed by aid of the conveyor belt, the material being drawn away from the magnetic field region. When the belt speed or width, or the thickness of the feed, is unsuitable for a single pulley, two may be used, one at each end of a short auxiliary belt that receives its feed from the main conveyor.

Drum This has an advantage over the pulley type in respect of its more effective separation. A drum can operate in conjunction with a belt conveyor if placed below the head pulley and a suitable guide. Feed is readily arranged down a chute or directly on to the feeder tray, if one is provided. A common type of feeder has the tray oscillated by an eccentric motion, or vibrated in a straight-line motion, at about 15 Hz.

Suspension A structure resembling a lifting magnet is suspended over a conveyor belt. It operates successfully on feeds containing awkward shapes of tramp iron at a belt speed up to 2.5 m/s. The magnet will not automatically discharge its load, but the large gap can contain a considerable load. The power rating is large.

Disc Most machines utilise rotating discs with a sharp or serrated periphery, set above the conveyor belt and over the magnet. Separation of iron depends on the change of polarity of a given region of the disc as it rotates, so that ferrous particles can be released.

Induction roll A powerful magnet (*Figure 20.10*) is provided with a return path plate. Rollers set between them are

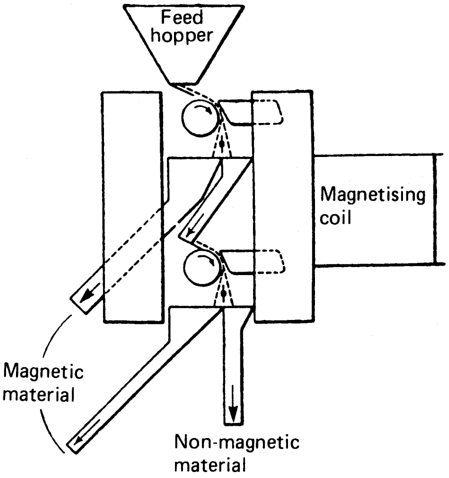


Figure 20.10 Induction roll separator

magnetised by induction. Material is fed into the top. Non-magnetic pieces fall through under gravity, while ferrous material adheres to the roller and is carried round and detached. Up to eight rollers in tandem may be used.

Wetherill The Wetherill separator has a single magnet unit mounted either side of a conveyor belt on which the material to be treated is passed beneath the upper magnet pole (*Figure 20.11*). Another belt is arranged over the upper pole of each magnet to take off the extracted ferrous material. The success of the separator depends on the shape of the magnet poles: the lower is flat and the upper is arranged with a ridge to concentrate the field. As the material passes under the magnets, each ferrous particle jumps towards the upper pole and is intercepted by the take-off belt, which in turn carries it to the side where it is discharged in a continuous operation. In practice, several magnets are used; the number of products that can be separated in a single operation is determined by the number of take-off belts, of which there are two per magnet unit.

20.2.6.2 Ore separation

For dealing with material in large lumps the magnetic field must have a deep penetration. This involves widening out the poles. The flux density is inevitably weakened. Thus, while feed depths of 250 mm are usual with a drum of 1 m diameter for the removal of tramp iron, the depth must be cut to, say, 75 mm when feebly magnetic material is operated on.

An important branch of separation deals with the subdivision and concentration of ores, the constituents of which have permeabilities very much lower than that of iron. Data on this point are given in *Table 20.2*. The process may be performed in several ways. A single product can be removed from the bulk; several constituents may be removed, each separately, in a single operation; or the separation may be carried out by a wet process.

The general design for feebly magnetic materials differs from that for the removal of tramp iron, essentially in the necessary flux density. A material with a permeability of 1% of that of iron may require a gap density exceeding 1.6 T, and the field must be divergent. For this purpose the lower pole over which the material passes is made flat; the upper pole, whether fixed or moving, is provided with a concentrating V-edge so that particles travel to it out of the general bulk of the material treated.

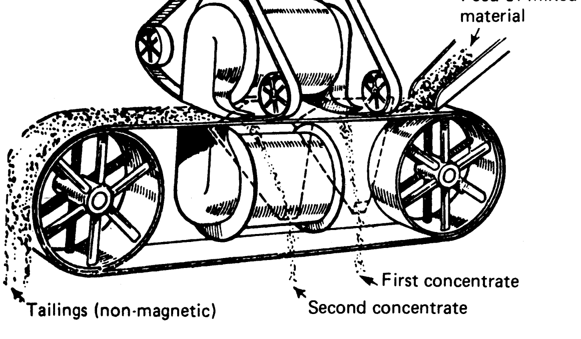


Figure 20.11 Wetherill separator

20.2.7 Clutches

The conventional clutch consists essentially of two members: the field member, which carries the exciting winding, and the armature member, consisting virtually of a steel ring which becomes attracted to the field member when the winding is energised. The engaging surfaces of these members have a friction lining for taking up the load when the clutch engages, and means are provided for spring disengagement of the armature when the winding is de-energised. As the clutch rotates in operation, it is necessary to employ slip-rings and brushes for the current supply.

A special type of clutch with a double friction lining is shown in section in *Figure 20.12*. In this case the field and armature members rotate together on the same shaft. The other shaft carries on a spring plate the lining carrier member. The two friction surfaces on this member engage between the armature and field members when the field coil is energised. General particulars for representative sizes of this type of clutch are given in *Table 20.3*.

20.2.8 Couplings

Eddy-current couplings resemble induction motors in that they develop torque by 'slip', and the throughput efficiency falls with decrease of speed. In selecting a coupling the critical factors are the speed range and the load torque variation therein.

The essential features are shown in *Figure 20.13*. The outer member (the loss drum) is mounted on the shaft extension of the drive motor, and the inner member (the pole system) on the driven shaft. Operation depends on the induction of current in the loss drum by e.m.f.s resulting from the speed difference between the driving and driven shafts. The two types illustrated are:

(1) *Interdigitate*. This is common for drives transferring up to about 100 kW. The loss drum is of plain ferromagnetic material of low resistivity, normally with forced cooling. The 'claw'-shaped pole structure gives a multipolar field by means of a single exciting coil. There is substantial interpolar leakage flux.

(2) *Inductor*. The toothed rotor produces an alternating flux density pattern in the loss drum by the modulation of the airgap permeance. An annular exciting coil, fixed or rotary, causes the two air gaps to have opposite polarity,

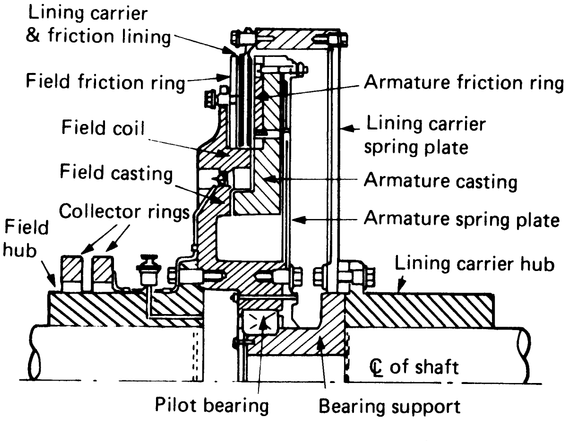


Figure 20.12 Electromagnetic clutch with double friction lining

Table 20.2 Relative attraction force (iron = 100) of various materials

Apatite	0.2	Dolomite	0.2	Lithium	0.5	Quartz	0.4
Argentite	0.3	Fluorite	0.1	Magnesium	0.8	Rutile	0.4
Biotite	3.2	Franklinite	35.4	Magnetite	40.2	Siderite	1.8
Bornite	0.2	Garnet	0.4	Manganese	8.9	Strontium	3.4
Cerium	15.4	Haematite	1.3	Molybdenite	0.3	Titanium	1.2
Chromium	3.1	Ilmenite	24.7	Palladium	5.2	Tungsten	0.3
Corundum	0.8	Limonite	0.8	Pyrrhoite	6.7	Zircon	1.0

Table 20.3 Clutches with double friction linings

Overall diameter (m)	*Max. power per 100 rev/min* (kW)	*Max. torque* (kN-m)	*Max. speed* (rev/min)	*Kinetic energy at 100 rev/min* (kJ)	*Mass* (kg)	*Current at 240 V* (A)
0.6	33	3	1200	0.5	300	0.9
0.8	95	9	900	1.8	520	1.4
1.0	200	20	700	5	960	2.1
1.2	360	36	600	11	1300	2.6
1.5	670	67	480	31	2200	3.2
1.8	1250	125	400	67	3400	4.2
2.1	1600	160	250	125	4700	4.7

the flux between them completing its path through the loss drum.

20.2.9 Brakes

The three basic forms of brake are: (i) solenoid-operated, (ii) tractive, and (iii) a thruster (electrohydraulic). In each case a brake-band or (more commonly) a brake-shoe is pressed against the brake-drum, either by weights or by springs operating through a lever. The use of springs is preferable, especially with large brakes, as the cushioning of the shock due to a falling weight introduces additional problems of design as well as limiting the positions in which the brake may be mounted. The brake is released by the operating force acting against the force due to the resetting spring. The brake is held in the off position for as long as the controlling circuit is energised.

The pressure used on the friction surfaces and the coefficient of friction are of the same order as for clutches. The pressure employed should be such as to give a reasonable rate of wear, and the figure chosen will determine the width of shoe required for a given operating force and wheel diameter. In general practice there are two brake-shoes, each embracing about one-quarter of the wheel circumference.

20.2.9.1 Solenoid brake

The brake is held 'off' by a solenoid/plunger device acting through leverage against spring loading, the latter being adjustable to suit the brake-torque requirements. If the brake is energised only for short periods, with intervening periods of rest (with the brake on), it is usually possible to fit a coil giving more ampere-turns than are obtainable with a continuous rating and thus to use a greater resetting spring pressure, giving increased braking torque.

20.2.9.2 Tractive brake

The example in *Figure 20.14* embodies a tractive electromagnet operating on inner and outer disc armatures A when the magnetising coil B is energised. The mechanical features are the adjusting wedge C, the brake-shoes D, the adjusting nuts E for the outer shoe-lever F, the torque spring G and its adjuster H, the tie-rod J, the terminals K, and the shoe-clamping screws L. Armatures AA rest in slots in the base and tend to remain against the slot abutments as a result of spring pressure and magnetic force. The powerful mainspring forces the armatures AA apart, causing the inner to apply pressure to the right-hand shoe and the outer to the left-hand shoe through the tie-rod J. When coil B is energised, the armatures AA mutually attract, so releasing the brake.

20.2.9.3 Thruster brake

The thruster brake employs a hydraulic thruster cylinder, with a piston acting under the fluid pressure produced by a small motor-driven pump unit. The power consumption is relatively low, but there is a short time-lag in brake response.

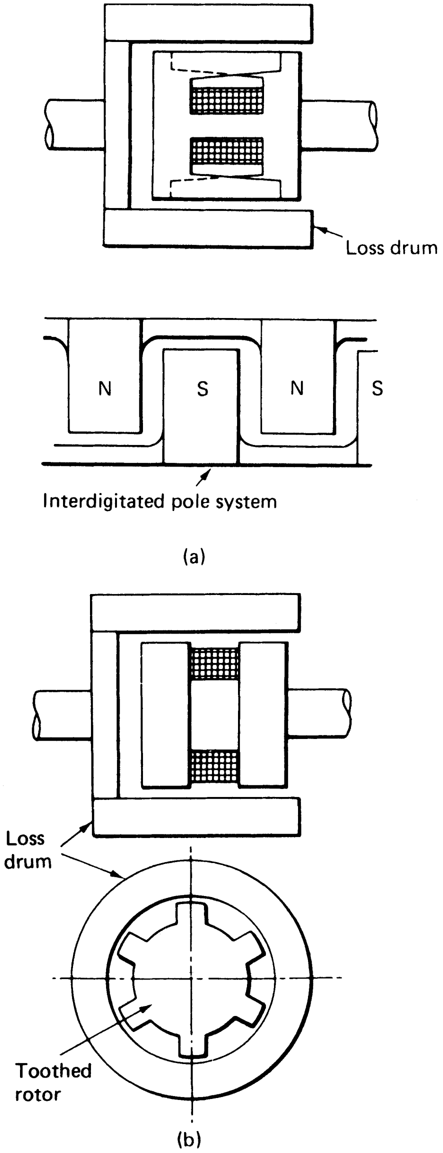
20.2.10 Magnetic chucks

In cases where awkwardly shaped ferrous-metal parts have to be machined in any quantity, the electromagnetic chuck forms a valuable auxiliary to various kinds of machine tools. The chuck contains a number of distributed windings which when energised from a d.c. source produce a concentrated and uniform field at the surface of the chuck, which is ground flat so as to form a suitable base-plate for accurate machining operations. The magnetic pull on ferrous materials in contact with the chuck surface is sufficient to prevent movement under all normal machining stresses.

When the current is switched off, the residual magnetism is in some cases sufficient to prevent easy removal of the part. The usual form of control switch accordingly has a demagnetising position.

The principle can be applied to rotating chucks, in which case slip-rings are necessary to convey exciting current to the windings.

In some cases *permanent-magnet* chucks can be employed. Hold and release of the workpiece are effected by an operating



(a)

(b)

Figure 20.13 Eddy-current couplings

lever which, in the off position, closes the flux paths of the magnets through high-permeability bridges and reduces the flux through the work. With either electro- or permanent-magnet forms, the workpiece may have to be demagnetised after machining.

20.2.11 Vibrators

A vibrator generator develops a vibro-motive force of adjustable magnitude and frequency for the noise, fatigue and vibration testing of small structures and for the assessment of mechanical resonance.

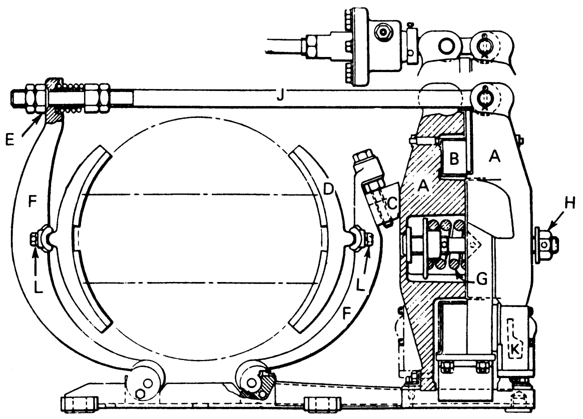


Figure 20.14 Tractive electromagnetic brake

20.2.11.1 Electrodynamic vibrator

Figure 20.15(*a*) shows the essential features of an electrodynamic vibrator, which are those of a powerful loudspeaker mechanism in which a circular coil, carrying an alternating current and lying in a constant radial magnetic field, develops vibratory force and displacement of corresponding frequency. A construction of the form shown can be adapted to develop torsional vibration by pivoting the armature centrally.

20.2.11.2 Magnetostrictive vibrator

The magnetostriction effect can be employed by placing the a.c. exciting coil around a stack of magnetostrictive material (*Figure 20.15*(*b*)). Mechanical amplification of the very small displacement is provided by a truncated drive rod. Vibrators of this kind are generally fixed-frequency devices, but they are suitable for relatively high frequencies only.

Single-frequency low-power vibrators can be constructed with piezo-electric drive. As large crystals are not readily available, these vibrators are usable only in the ultrasonic frequency range.

20.2.12 Relays and contactors

Relays and contactors, a.c. or d.c. excited, are widely employed for low- and high-power switching. The basic features are shown in *Figure 20.16*.

20.2.12.1 Contactors

The term ‘contactor’ applies to power-control devices. For d.c. operation the contactor is made single- or double-pole as required. When the coil is energised, a magnetic field is established across the air gap and the armature is attracted to the pole to close the contacts. The moving contact has a flexible conductor attached to it in order to avoid passing current through the hinge. The destructive effects of d.c. arcs are such as to make necessary an arc shield and magnetic blow-out arrangement. The blow-out winding carries the main current and its connection is so arranged that the arc is expelled from the contact region when the contacts separate.

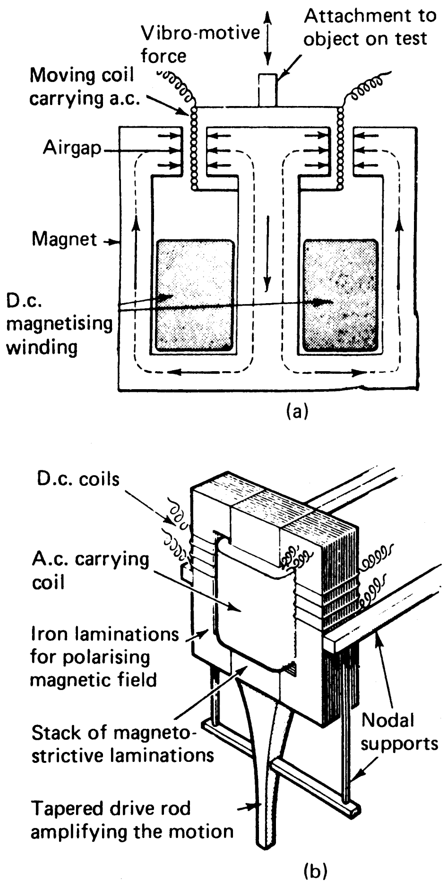

Attachment to object on test
Vibro-motive force
Moving coil carrying a.c.
Airgap
Magnet
D.c. magnetising winding
(a)
D.c. coils
A.c. carrying coil
Iron laminations for polarising magnetic field
Stack of magneto-strictive laminations
Nodal supports
Tapered drive rod amplifying the motion
(b)

Figure 20.15 Vibrators

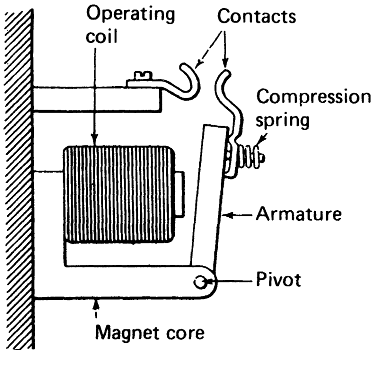


Figure 20.16 Elements of a contactor

For a.c. service the contactor normally has two or three poles. The magnetic circuit is laminated and the pole-face has a shading coil to reduce 'chatter'. Blow-out coils may not be provided because the principle operates less effectively on a.c.; reliance may be placed on extinction at a current zero. A typical a.c. contactor is illustrated in *Figure 20.17*.

Ratings These have been standardised. The severity of operating conditions varies considerably according to the class of service. Although the cleaning action on the contacts due to frequent operation is desirable in removing cumulative high-resistance films which tend to increase heating, this class of service causes greater contact wear and erosion for a given loading than would occur with less frequent operation. Conversely, very infrequent operation which involves the contacts carrying current for long periods is not onerous from the viewpoint of wear and erosion but is conducive to the formation of high-resistance surface films unless a suitably low temperature is maintained so as to limit the formation of the films. The permissible temperature rise for different types of contact is given in *Table 20.4*. Operation must be satisfactory with the shunt windings at final rated temperature and with reduced operating voltage (80% of normal for d.c., 85% for a.c.).

Table 20.4 Temperature limits for contacts

Type of contact	*Temperature rise* (°C)
Solid copper in air	
Standard rating	65
Uninterrupted rating	45
Solid copper in oil	45
Laminated copper in air or in oil	40
Solid silver or silver faced in air	80
Carbon	100

20.2.12.2 Relays

The electromagnetic relay operates one or more sets of contacts by the attraction of a movable armature towards a magnetised core. The representative types shown in *Figure 20.18* are: (a) the 'telephone' type with pivoted armature; (b) the 'commercial' version of (a); (c) a mercury switch with hinged armature; and (d) a spring-suspended armature. An important feature is the operating time.

High-speed operation may be obtained by one or more of the following methods: (i) lamination of the magnetic circuit to minimise eddy-current delay; (ii) reduction of the mass of moving parts; (iii) use of a large coil power; or (iv) reduction of coil inductance.

Low-speed operation, sometimes needed to introduce a time-lag, is obtained by: (i) use of a lag (or *slugging*) coil comprising an additional and separate short-circuited loop or winding; (ii) use of a series inductor or shunt capacitor; or (iii) addition of an external time-delay relay.

Design features Contact sets may be normally open or normally closed, and both types may be fitted on the same relay mechanism. The arrangement is determined by the operating sequence required: i.e. make, break, change-over, make-before-break, break-before-make. The contact size and material must be chosen in accordance with the rating and electrical characteristics of the circuits controlled.

Ideally, the contacts should operate cleanly and with no bounce. They should be of adequate size and of the most suitable material. In extremely low-voltage circuits the contact resistance is usually an important consideration and special precautions may also have to be taken to ensure reliable operation under conditions of vibration or shock.

Similarly, in cases of high-current switching it may be necessary to ensure wide separation of the contacts or even to

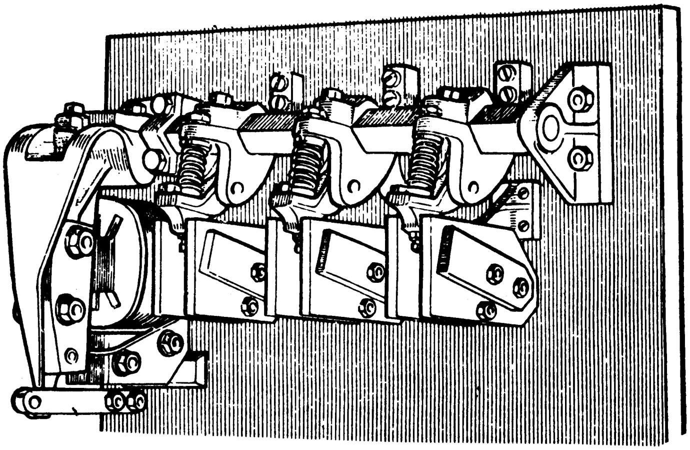


Figure 20.17 Triple-pole a.c. contactor

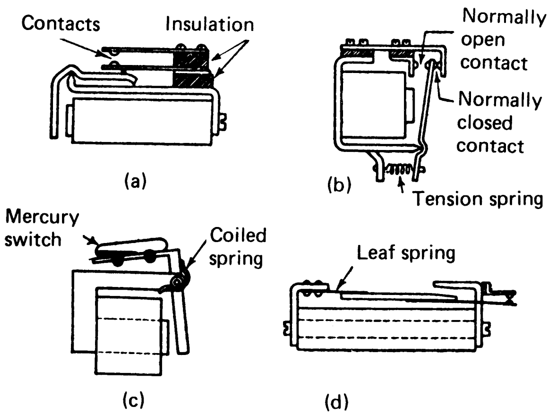


Figure 20.18 Electromagnetic relays

arrange for several gaps to operate in series. In some cases it may be necessary to use arc-suppressing circuits.

The number and type of the contacts and springs determines the switching operation to be performed by the relay; this factor also determines the work to be done by the magnetic circuit. It follows, therefore, that the choice of a suitable coil and iron circuit design is determined by the contact arrangement of any particular relay. Various configurations of magnetic circuits and materials are used in the relays under review, depending upon their particular application. For example, in the high-sensitivity relays, where the air gap has to be kept to a minimum, it is necessary to use materials having a very low residual magnetism and high permeability.

The power required to operate the relay is determined by the spring-set arrangement and the magnetic circuit. The method of construction is important, since it largely determines the safe operating temperature of the winding and this, in turn, governs the coil power and the maximum pull available at the armature. By increasing the area of the flux path while maintaining the ampere-turns and coil power constant, the total air gap flux, and therefore the armature pull, can be increased and the increased coil area will permit cooler operation of the coil. This may, however, lead to a relay that is physically larger than can be tolerated. In practice, therefore, it is more reasonable to build a relay of a given size and to use other means to amplify the controlling power.

The continuous power input to a given relay coil is limited only by the maximum temperature that the coil insulation can withstand without breakdown. This temperature is governed by the environmental conditions as well as by the coil construction and the quality of the insulating material.

Many of the functions performed by the electromagnetic relay have been taken over by solid-state switching.

20.2.13 Miniature circuit-breakers

The miniature circuit-breaker (m.c.b.) is, for the control of small motors and domestic subcircuits, considered primarily as an alternative to the fused switch. The appropriate British Standard is BS 3871, which lays down specific technical requirements. The usual form of the m.c.b. embodies total enclosure in a moulded insulating material. As the operating mechanism must be fitted with an automatic release independent of the closing mechanism, the m.c.b. is such that the user cannot alter the overcurrent setting nor close the breaker under fault conditions. At the same time the m.c.b. must tolerate harmless transient overloads while clearing short circuits. For most practical conditions, a change-over from time-delay switching to 'instantaneous' tripping at currents exceeding 6–10 times full-load rating is suitable.

20.2.13.1 Tripping mechanisms

Methods of achieving the required operating characteristics can be classified as (i) thermomagnetic, (ii) assisted thermal and (iii) magnetohydraulic. In the *thermomagnetic* method the time-delay is provided by a bimetal element, and the fast trip by a separate magnetically operated mechanism based on a trip coil. In the *assisted thermal* method the bimetal is itself subjected to magnetic force. The *magnetohydraulic* mechanism incorporates a sealed dashpot with a fluid and a spring restraint, the dashpot plunger being of iron and subject to the magnetic pull of the trip coil. The essential features are illustrated in *Figure 20.19*.

Thermomagnetic The bimetal element shown in *Figure 20.19*(*a*) may carry the line current or, for low current ratings, be independently heated. Its flexure operates the trip latch through a crank. On overcurrent the magnetic force acts directly on the latch bar, with or without the aid of the bimetal deflection.

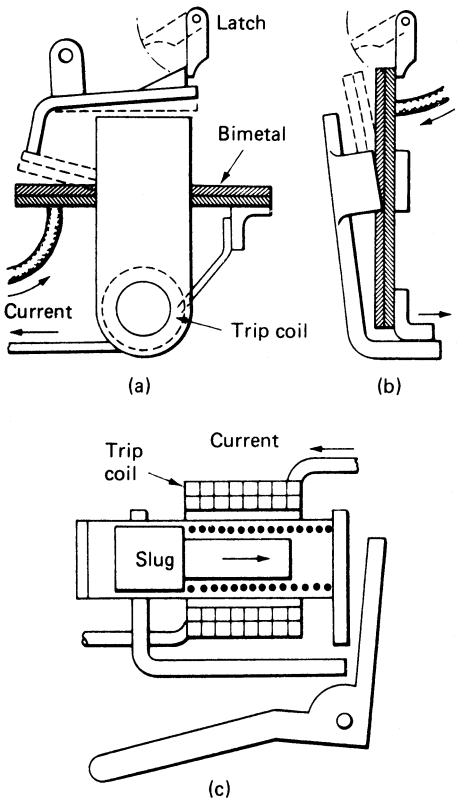


Figure 20.19 M.c.b. trip mechanisms

Assisted thermal The time-delay characteristic is provided by a bimetal element, and instantaneous tripping by magnetic deflection of the bimetal. The operation is shown in *Figure 20.19*(*b*). A bar of magnetic material is placed close to the bimetal element, and the magnetic field set up by the current develops a pull on the bimetal such as to increase its deflection and release the trip latch. The magnetic effect is proportional to the square of the current and so becomes significant on overcurrent. However, as the position of the bimetal element on the occurrence of a short circuit is arbitrary, there is no well-defined change-over point at which instantaneous tripping occurs.

The method is cheap and simple, but is difficult to design for low-current (e.g. 5 A) breakers because the operation tends to be sluggish, particularly at fault-current levels that are less than 500 A.

Magnetohydraulic This method, shown in *Figure 20.19*(*c*), combines in one composite magnetic system a spring-loaded dashpot with magnetic slug in a silicone fluid, and a normal magnetic trip. When the line current flows, the magnetic field produced by the trip coil moves the slug against the spring towards the fixed pole-piece, so reducing the reluctance of the flux path and increasing the magnetic pull on the trip lever. If it reaches the end of the dashpot, the pull is sufficient to operate this lever and trip the circuit-breaker. On sudden overcurrent exceeding 6–10 times full-load value, there is sufficient pull at the fixed pole-piece to attract the armature of the trip lever regardless of the position of the slug in the dashpot. The characteristic is more definite and satisfactory for low-current ratings than that of the assisted thermal mechanism.

20.2.13.2 Operating features

Thermal operation by bimetal elements implies that the effective current rating is a function of the ambient temperature. It is the practice, if complete ambient compensation is not fitted, to rate m.c.b.s in such a way as to allow for the type of enclosure. With magnetohydraulic devices the tripping is independent of the ambient temperature over a specified range, the small variations due to change of viscosity of the damping fluid being minimised by use of a fluid with a nearly flat viscosity–temperature characteristic.

The combination of thermal and magnetic functions is not easily controlled for low current ratings, and for m.c.b.s with such ratings the tolerances on operation must be wider than they are for larger currents.

Normally, m.c.b.s are suitable only for a.c. circuits. As with all a.c. switchgear, the problems of breaking efficacy are associated not only with the actual short-circuit current but also with its asymmetry and power factor.

As m.c.b.s can be linked to give two- and three-pole versions, so arranged that a fault on one pole will produce complete circuit isolation, the risk of single-phasing in motor control is effectively eliminated. In other directions, however, m.c.b.s cannot necessarily replace fuses: they do not possess the high short-circuit breaking capacity of the modern h.r.c. fuse, nor do they have its inherent fault-energy limitation. If, therefore, conditions are such that back-up protection has to be provided for m.c.b.s, the 'take-over' zone should be of the order of 1.0–1.3 kA.

20.2.14 Particle accelerators

Modern accelerators produce high-energy beams of electrons, ions, X-rays, neutrons or mesons for nuclear research, X-ray therapy, electron irradiation and industrial radiography. If a particle of charge e is accelerated between electrodes of p.d. V it acquires a kinetic energy eV electron-volts ($1\,\text{MeV} = 1.6 \times 10^{-13}$ J). Accelerators are classified as *direct*, in which the full accelerating voltage is applied between the two electrodes; *indirect*, in which the particles travel in circular orbits and cyclically traverse a region of electric or magnetic field, gaining energy in each revolution; and *linear*, in which the particles travel along a straight path, arriving in correct phase at gaps in the structure having high-frequency excitation, or move in step with a travelling electromagnetic wave.

20.2.14.1 Direct accelerators

The Cockcroft–Walton multiplier circuit has two banks of series capacitors, alternately connected by rectifiers acting as change-over switches according to the output polarity of the energising transformer. The upper limit of energy, about 2 MeV, is set by insulation. A typical target current is 100 μA.

The Van de Graaff electrostatic generator is capable of generating a direct potential of up to about 8 MV of either polarity. It has an endless insulating belt on to which charge is sprayed from 'spray-set' needle-points at about 50 kV. The charge is carried upwards to the interior of the high-voltage (h.v.) electrode, a metal sphere, to which it is transferred by means of a second spray set. H.v. insulation difficulties are overcome by operating the equipment in a tank filled with a high-pressure gas, e.g. nitrogen–freon mixture at 1500 kN/m^2. In two-stage Van de Graaff generators for higher energies, negative hydrogen ions are accelerated from earth potential to

6 MeV; they are then fired into a thin beryllium foil 'stripper' which removes the electrons from the outer shells of the atom and leaves the remanent ions moving on with little change of energy but with a positive charge. The second stage brings these ions back to earth potential and the total energy gain is 12 MeV. To bring ions on to a small target the accelerating and deflecting fields must be accurately controlled, and scattering limited by evacuating the accelerator tubes to very low pressure. The energies are sufficient for the study of nuclear reactions with the heaviest elements.

20.2.14.2 Indirect (orbital) accelerators

Indirect (orbital) accelerators may have orbits of approximately constant radius with a changing magnetic field (betatrons and synchrotrons) or orbits consisting of a series of arcs of circles of discrete and increasing radii in a constant magnetic field (cyclotrons and microtrons).

Betatron The betatron is unique in that the magnetic field not only directs particles into circular orbits but also accelerates them. The magnet has an alternating field of which only one quarter-period is used. Electrons are accelerated in an evacuated toroidal chamber between the poles of the magnet. They are injected at an energy corresponding to a low magnetic field, which bends them in circular orbits round the toroid. A cross-section of the poles and vacuum chamber is shown in *Figure 20.20*. As the magnetic flux through an electron orbit increases during the cycle of alternation, the electron experiences a tangential force, and its gain in energy per revolution is the voltage that would be induced in a loop of wire in the orbit. As the electron gains energy, the magnetic guide field intensity at the orbit increases at a suitable rate. To keep the electron on a constant radius from injection to peak energy requires the rate of change of intensity at the orbit to be half that of the mean flux per unit area within the orbit. At peak energy (or earlier) the electrons are caused to move away from their equilibrium orbit and to strike a target inside the vacuum chamber, producing X-rays or corresponding energy. The output consists of short pulses of radiation whose repetition rate is the frequency of the magnet excitation. Energy limitations are set by the size and cost of the magnet and the radiation loss when a high-energy electron has circular motion.

Synchrotron The synchrotron uses an annular magnetic guide field which increases as the particles gain energy, as in the betatron, so that they maintain a constant orbit radius. Electrons are initially accelerated by the action of central 'betatron bars' which saturate when the main magnetic field corresponds to an energy of 2–3 MeV when electrons travel at a velocity only 1–2% less than the velocity of light. Further gain of energy is produced by radio-frequency (r.f.) power at the frequency of orbital rotation (or a multiple of it) that is fed to resonators inside the vacuum chamber. The particles become bunched in their orbits so that they pass across the accelerating gap in the resonator at the correct phase of the r.f. field. The limitation on electron acceleration is now mainly set by radiation losses due to circular motion.

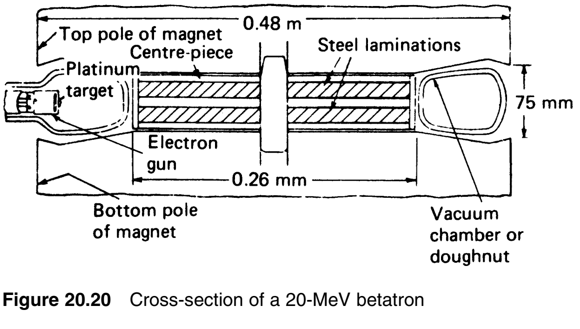


Figure 20.20 Cross-section of a 20-MeV betatron

Protons are injected at about 500 keV, which produces a velocity of only 3% of that of light. Further acceleration changes the frequency of orbital rotation. For a proton synchrotron the magnetic guide field strength and the r.f. power frequency have to be varied accurately over large ranges.

Cyclotron This early form of accelerator consists of a vacuum chamber between the poles of a fixed-field magnet containing two hollow D-shaped electrodes which load the end of a quarter-wave resonant line so that a voltage of frequency 10–20 MHz appears across the accelerating gap between the 'D's. Positive ions or protons are introduced at the centre axis of the magnet and are accelerated twice per rotation as they spiral out from the centre. The relation between particle mass m, charge e, magnetic flux density B and frequency f is $f = Be/2\pi m$. Energy limitation is set by the relativistic increase of mass, which limits the speed of high-energy particles so that their phase retards with respect to the r.f. field.

Synchrocyclotron In this device the energy limitation of the cyclotron can be removed by modulating the oscillator frequency to a lower value as a bunch of particles gains energy.

Microtron In the microtron, or electron cyclotron, electrons are accelerated in a vacuum chamber between the poles of a fixed-field magnet. The orbits consist of a series of discrete circular arcs which have a common tangent at a resonant cavity in which the electrons gain their successive increases of energy from an r.f. electric field. The highest energy achieved with such a machine is 6 MeV, and mean currents are less than 1 μA.

20.2.14.3 Linear accelerators

Indirect accelerators of protons have so far used a resonant cavity in which drift-tube electrodes are introduced that distort the fields and enable particles to be shielded from field reversals. Particles are accelerated between gaps and move between centres of successive gaps in one complete period of oscillation (*Figure 20.21*). Oscillators operating at about 200 MHz and a pulse power of 1–2 MW are used to excite the cavity for some hundreds of microseconds. Injection is by a Cockcroft–Walton or Van de Graaff device.

An important device for electron acceleration is the travelling-wave accelerator, using megawatt pulses of r.f. power at 3000 MHz. The power is propagated along a cylindrical waveguide loaded with a series of irises. A travelling wave is set up with an axial electric-field component, and correct dimensioning of the iris hole radius a and the waveguide

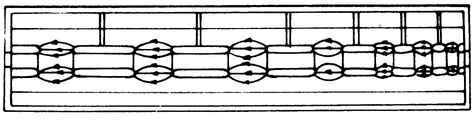


Figure 20.21 Field resonant cavity for a proton accelerator

radius *b* (*Figure 20.22*) enables the propagation velocity and the field-intensity/power-flow relation to be varied. An electron injected along the axis with an energy of the order of 45 keV is accelerated by the axial field, and as its velocity changes it remains in correct phase with the travelling field, the propagation velocity of which is varied to match. A fixed axial field is required to provide for electron focusing.

High-energy machines with low beam-currents have been used in the USA, low-energy machines with high beam-currents in the UK. The 25 MeV Harwell accelerator has a length of 6 m divided into six sections, each fed by a 6 MW klystron amplifier to give a peak beam-current of 1 A and a mean outpot power of 30 kW.

20.2.14.4 Large machines

The Harwell proton synchrotron gives the particles an energy of 7 GeV in an orbit of radius 19 m within a 7000 t magnet. The magnet takes 10 kA to raise the orbit flux density to about 1.4 T in 0.75 s, to hold this value for 0.25 s and to reduce it to zero in 0.75 s, with a repetition frequency of about two cycles per hour. The inductance of the magnet is about 1.1 H, and to produce a rate of change of current of 10/0.75 = 13.3 kA/s the magnet supply voltage must be about 14 kV. The peak stored energy is 40 MJ. The supply is from a pair of 3750 kW/75 MW motor/generators through rectifiers.

The new accelerator for CERN near Geneva is to use the existing 25 GeV machine to inject particles into a 300 GeV proton synchrotron with an orbit of diameter about 2.2 km. The magnet will employ superconducting exciting windings giving a flux density of 4–6 T. The design is such that it can be built initially with only alternate magnet sections, and upgraded in energy later without basic alteration of the main structure, possibly to 800 GeV.

20.3 Industrial rotary and linear motors

The elements discussed in Section 2.4.3.6 indicate that there are two methods of developing a mechanical force in an electromagnetic machine.

(1) *Interaction.* The force f_e on a conductor carrying a current *i* and lying in a magnetic field of density *B* is $f_e = Bi$ per unit length, provided that the directions of *B* and *i* are at right angles; the direction of f_e is then at right angles to both *B* and *i*. This is the most common arrangement.

(2) *Alignment.* Use is made of the force of alignment between two ferromagnetic parts, either or both of which may be magnetically excited. The principle is less often applied, but appears in certain cases, e.g. in salient-pole synchronous machines and in reluctance motors.

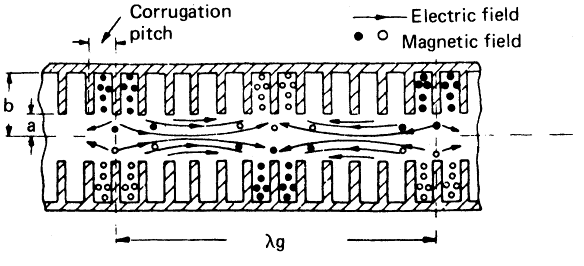


Figure 20.22 Fields in a corrugated waveguide

20.3.1 Prototype machines

Three basic geometries (*Figure 20.23*) satisfy the requirement for the relative orientations of *B, i* and f_e. For a rotary case the *cylindrical* form is the most common, while the *disc* with its short axial length suits particular applications. The *flat* form is employed for linear motion.

Such machines are almost exclusively *heteropolar* (i.e. have alternate north and south poles). To maintain unidirectional interaction force, the direction of the current in a given rotor conductor must reverse as it passes from a north pole to a south pole region.

20.3.1.1 Heteropolar cylindrical machine

A heteropolar cylindrical machine for a two-pole magnetic circuit is shown in *Figure 20.24*. The active region is the air gap between stator and rotor. In *Figure 20.24*(*a*) the stator conductors are arranged (normally in slots) on the surface and are connected so as to develop the current-sheet pattern indicated, giving rise to a distributed m.m.f. of peak value F_1 on the axis of the winding. A corresponding rotor current-sheet pattern sets up an m.m.f. F_2. If the m.m.f. distributions are assumed to be sinusoidal, the torque on the rotor can be shown to be

$$M_e = kF_1F_2 \sin \lambda$$

where λ is the *torque angle* between the stator and rotor winding axes, and *k* is a function of the air gap dimensions. For $\lambda = 0$ the m.m.f.s F_1 and F_2 are in alignment and there is no torque. Displacement of the rotor increases the torque, which reaches a maximum for $\lambda = \pi/2$ rad. For further displacement the torque falls, to become zero again for $\lambda = \pi$ rad.

The machine in *Figure 20.24*(*b*) has a fixed optimum torque angle $\lambda = \pi/2$ rad. Here the rotor must have a closed winding and be provided with a commutator or alternative switching device so that each conductor, as it passes from a north to a south polar region, has its current automatically reversed. Then the direction of F_2 is fixed for all operating conditions. As the direction of F_1 is also fixed, it is usually developed by salient poles.

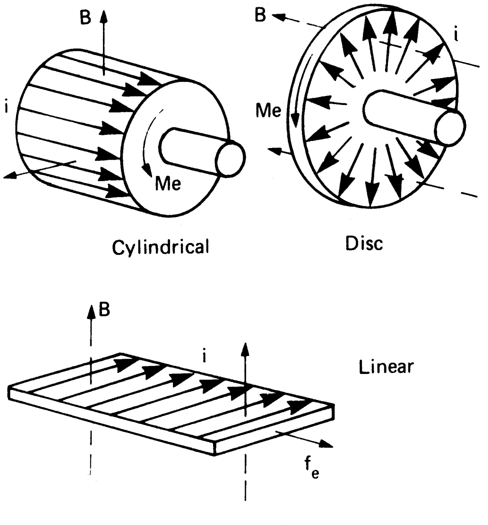


Figure 20.23 Basic geometries for electromagnet machines

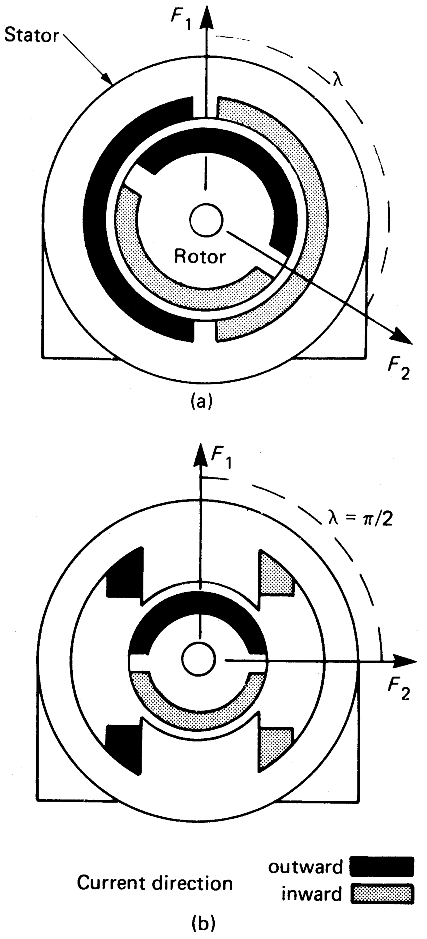


Figure 20.24 Prototype two-pole cylindrical machines

20.3.1.2 Types of machine

The three most common machines—synchronous, induction (asynchronous) and commutator—are all heteropolar and have at least one member cyclindrical. They are distinguished by the nature of the supply (a.c. or d.c.) and that of the air gap flux (travelling wave or fixed axis).

Travelling-wave gap flux The stator current-sheet pattern in *Figure 20.24*(*a*) is set up by a three-phase winding ABC as in *Figure 20.25* for a two-pole machine. With the phase windings excited by balanced symmetrical three-phase currents of frequency f_1, the sequential cyclic reversal of currents in the displaced windings shifts the current-sheet pattern, as shown for peak current (i) in phase A and (ii) in phase B (one-third of a period later). Thus the stator m.m.f. F_1 produces a travelling wave of m.m.f. and air gap flux (often called a 'rotating field') moving at synchronous speed $n_s = f_1$ (rev/s) or angular speed $\omega_1 = 2\pi f_1$ for a three-phase supply of frequency f_1 to a two-pole machine; in general $n_s = f_1/p$ and $\omega = 2\pi f_1/p$ for a machine with p pole-pairs.

Let the rotor, rotating at angular speed ω_r, have a three-phase winding carrying currents of frequency f_2; then it has an m.m.f. F_2 rotating at angular speed $\omega_2 = 2\pi f_2$ with respect to the *rotor* body, and therefore at $\omega_r \pm \omega_2$ with respect to the *stator*. For a steady unidirectional torque to be developed, F_1 and F_2 must rotate in synchronism to preserve unchanging the torque angle λ. Thus $\omega_r \pm \omega_2 = \omega_1$ is the essential running condition.

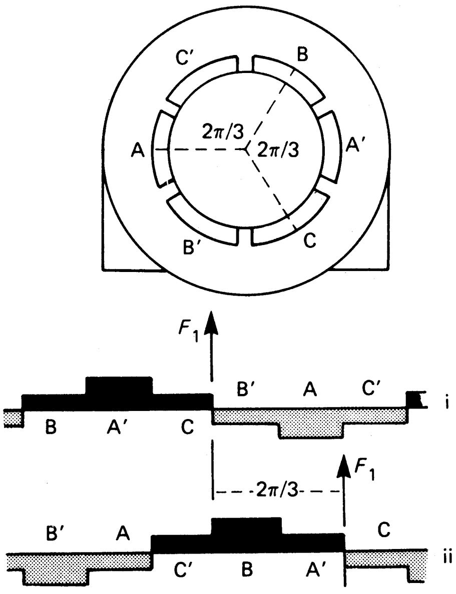


Figure 20.25 Production of a travelling-wave field

Synchronous machine The rotor is d.c. excited, so that F_2 is 'fixed' to the rotor body; then $\omega_2 = 0$ and $\omega_r = \omega_1$. The rotor must therefore rotate synchronously with the stator travelling-wave field. The torque angle accommodates to the torque demand up to a maximum for the torque angle $\lambda = \pi/2$ rad. The machine can operate in both generator and motor modes by simple reversal of the torque angle.

Induction machine The rotor winding, isolated and closed, derives its current inductively from the stator. If the rotor spins at synchronous speed, its conductors move with the stator field and no rotor current can be induced. However, if ω_r is less than ω_1 by a fractional 'slip' $s = (\omega_1 - \omega_2)/\omega_1$, the rotor conductors lie in a field changing at slip frequency $s\omega_1$, and currents of this frequency are induced to provide an m.m.f. F_2 travelling around the rotor at this frequency. This gives $\omega_r = \omega_2 = \omega_1$, the required condition. Torque is developed for any slip s other than zero (synchronous speed), with F_1 and F_2 mutually displaced by the torque angle. By driving the machine above synchronous speed the slip and torque are reversed, and the machine generates.

Fixed-axis gap flux In the usual constructional form (*Figure 20.24*(*b*)), the gap flux is produced by the stator m.m.f. F_1, generally with the poles salient. The rotor m.m.f. F_2 has the optimum torque angle $\lambda = \pi/2$ rad. As the rotor spins, the current of an individual conductor is reversed as it passes from the outward- to the inward-directed region of the current sheet in the process of *commutation*. In consequence the machine can develop torque at standstill and at any practicable speed.

D.c. commutator machine Both stator and rotor windings are d.c. excited. The torque is smooth and continuous, with simple control of speed and both motor and generator operation. In small d.c. motors the stator may be magnetised by permanent magnets, dispensing with the exciting winding.

Single-phase commutator machine As simultaneous reversal of F_1 and F_2 does not affect the direction of the torque, the d.c. motor can be operated on a one-phase supply with the stator and rotor windings connected in series. However, the torque has a double-frequency pulsation about a unidirectional mean.

Other forms There are many variants, especially in small and miniature machines. Single-phase induction motors require special starting techniques ('split-phase', 'shaded-pole'). Some operate on alignment torque ('reluctance', 'hysteresis', 'brushless', 'stepper'). A few large homopolar d.c. machines have been devised.

Disc motors The geometry shown in *Figure 20.23* has been applied with permanent-magnet multipolar field systems to machines that must have a very short axial length, e.g. for driving cooling fans for motor vehicles.

Linear motors These are most usually based on the three-phase induction-motor principle.

20.3.2 D.c. motors

In spite of the fact that a standard d.c. motor costs 1.5–2 times as much as a cage induction motor, and that alternating current is universal for general power distribution, the scope for d.c. motors is still large, particularly for drives requiring speed control or some other special feature. D.c. motors are built in all sizes from fractional-kilowatt up to about 4 MW, the upper limit being imposed by commutation problems.

In addition to the standard types of motor (shunt, series or compound) which are normally fed from a constant-voltage d.c. supply, many modern d.c. motors incorporate thyristors enabling them to operate from a standard a.c. supply. The thyristors rectify the alternating voltage and, by gate control, enable the resulting direct voltage to be varied, thereby giving a wide range of speed control.

20.3.2.1 Characteristics

The connections and basic torque–current and speed–torque characteristics of standard shunt, series and compound motors are given in *Figure 20.26*. Motors with separately excited fields are often used for control purposes.

When connected to a d.c. supply of voltage V a motor takes a current $I=P/V\eta$ when developing a useful output power P. The efficiency η varies with the rating: typical rated currents are given in *Table 20.6* for a range of operating voltages. The rotational e.m.f. in an armature of resistance r over the brushes and carrying a current I_a is $E=V-I_ar_a$. The power-conversion relation is $EI_a=M\omega_r$, where M is the torque and $\omega_r=2\pi n$ is the angular speed. Then

$$E=V-I_ar_a=2(p/a)nN\Phi=Kn\Phi \quad M=K\Phi I_a/2\pi$$

where Φ is the flux. Here $K=2(p/a)N$ involves the number of pole-pairs (p), the total number of turns (N) on the armature and the number of pairs of parallel paths (a) of the winding. For simple lap and wave windings, $a=1$ and $a=p$, respectively. In a shunt-connected machine the total input current I is the sum of the armature and field currents, $I=I_a+I_f$. In a series-connected machine the same current I flows in both field and armature windings.

Shunt excitation The field winding has the constant terminal voltage applied to it so that the flux will be approximately constant and the torque will be proportional to the armature current. Speed is proportional to E and is approximately constant since I_ar_a is normally not more than 3–5% of V at full load. In practice, the flux will, owing to armature reaction, be distorted when the machine is loaded, the flux density under the leading pole tips being increased and that under the trailing pole tips decreased. Owing to saturation of the iron in the teeth under the leading pole tips, the increase in density there is less than the decrease in density under the trailing tips, so that there is a net reduction of flux of 2–3% at full load. The drop in speed from no load to full load is therefore less than

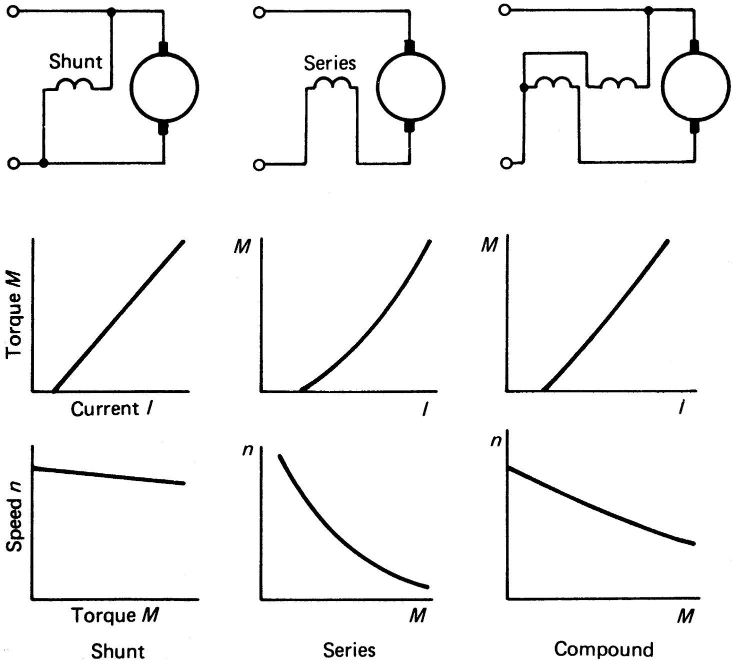


Figure 20.26 D.c. motors: basic characteristics

would be expected from the speed equation—in some cases this action even gives a rising speed characteristic, a disadvantage which can be corrected by the use of a small series field winding. Increase in temperature from cold to hot raises the resistance of the field winding and reduces the current in it, thereby reducing the flux and increasing the speed for a given load.

Motors designed to give a wide range of speed control by variation of the field or to be used in situations where sudden and heavy load fluctuations occur are often fitted with a compensating winding in the pole face to neutralise the effect of armature reaction and prevent flux distortion; such windings are connected in series with the armature so that neutralisation is correct at all loads.

The shunt motor can be used for the drive of machine tools, pumps and compressors, printing machinery and all forms of industrial drive requiring a speed which is approximately constant and independent of the load.

Separate excitation This is applied widely to control motors, particularly where speed variation is required over a considerable range. For a given field voltage, the characteristics resemble those of a shunt motor. Separate excitation (but without the facility of field control) applies to small motors with permanent-magnet field systems.

Series excitation The field m.m.f. is produced by the motor current so that at low currents where the iron is unsaturated Φ is approximately proportional to I, but at high currents (1.5–2 times full-load current) Φ tends to become constant as the iron saturates. The starting torque, when the current is above the full-load value, is thus greater than for a shunt motor with corresponding full-load current and flux. The speed at heavy currents drops to a low value on account of the increase of flux.

The high starting torque and falling speed–torque characteristic make the series motor suitable for driving hoists and cranes, for traction and rope haulage and for driving fans, centrifugal pumps or other apparatus where there is no danger of the motor being run light.

Compound excitation Where a drop in speed between no load and full load greater than that obtainable with a plain shunt motor is required, a series winding may be added to assist the shunt winding, giving a speed–torque characteristic having any desired amount of droop. Instability of a shunt motor due to the weakening of the field by armature reaction can also be cured by the addition of a series winding known as a *series stability* winding. The chief application of the compound motor arises when the motor is used in conjunction with a flywheel—a fairly steep droop to the speed–torque characteristic is then necessary in order to enable the flywheel to give up its stored energy when a sudden load comes on. Compound motors are also used for driving pumps, compressors and other heavy-duty machinery.

If the series winding is arranged to oppose the shunt winding, a motor with a flat or even a rising speed–torque characteristic can be designed, such a motor being known as a differentially compound motor. Such motors are, however, very rarely used.

20.3.2.2 Construction

Motor design aims at economy of materials and the reduction of loss. Further, as most industrial d.c. machines are fed from an a.c. supply through thyristors, an all-laminated magnetic circuit reduces the effect of supply harmonics. There is increasing use of square frames, either of rolled steel or of laminations.

Poles Constructed separately, the pole body and shoe are assembled from laminations, or a solid body is provided with a laminated shoe. The poles are bolted to the yoke and retain the field windings. Commutating poles may be solid or, more usually, laminated. Motors of rating below about 10 kW may have half as many commutating as main poles.

Field windings Main *shunt-field* windings are of circular- or rectangular-section wire, insulated and wound on a former. The whole is then taped, impregnated, slipped on to the pole and held by the pole-shoe. Large machines may have the turns wound on a bobbin of pressboard or of steel lined with micanite. *Series windings* to carry currents exceeding 50 A are more generally of copper strip wound on edge, and a similar construction is used for *compole* windings.

Armature core This is built from core-steel laminations (0.35–0.6 mm), coated on one side with an insulating varnish and bolted or clamped between thick end-plates. For diameters up to about 1 m the stampings may be made in disc form; above this size they are in sectors keyed to the shaft hub. If the core length exceeds about 20 cm, radial ducts are provided, each 5–6 cm. Axial ducts are employed with small machines. Machines of ratings up to 50 kW may have slots skewed to reduce noise.

Commutator and brushgear The commutator is conventionally made by assembling hard-drawn copper sectors inter-leaved with 0.7–2 mm sheet mica, these separators being ‘undercut’ by about 1 mm. The brushes, of a suitable carbon/graphite content, are mounted in boxes with spring loading to hold them against the commutator surface with a medium to strong pressure depending on the application. The circumferential brush width is typically 2–3 sector widths (10–20 mm) and about 30 mm axially. One brush-arm per pole is employed except for certain four-pole wave-wound machines which have two brush-arms in adjacent positions to facilitate maintenance.

Armature winding Almost all motors other than very large machines use a simple two-circuit wave-winding. Conductor wires of section 1 mm^2 or less are circular and enamel insulated, the conductors of a coil being taped half-lap before being placed in the slots. Larger machines have former-wound coils of rectangular-section conductors, insulated by half-lapped tape. The coils are assembled from the conductors and taped before insertion. The slots are lined with pressboard, and the two layers separated by a pressboard spacer. Various recent developments in epoxy resins have made possible the use of better insulants at higher temperatures. Typical slot sections are shown in *Figure 20.27*. The coils are contained in the slots by wedges or by steel or glass-cord binding.

In small wire-wound armatures the coil ends are soldered direct into grooves in the commutator sectors. For strip windings, the sectors carry ‘risers’ for connection to the ends of the coils.

Bearings End-shield bearings are usual for ratings up to 250 kW, above which pedestal bearings are employed. Journal bearings are fitted where ball or roller bearings are unsuitable.

Enclosure and ventilation Recent standards define the conditions to be met by machines for a variety of ambient conditions (e.g. drip-proof, splash-proof, hose-proof, weather-proof, and flame-proof).

Cooling air is drawn into the machine directly or through cowls, pipes or screens except in totally enclosed machines, for which there is no communication between the outside air

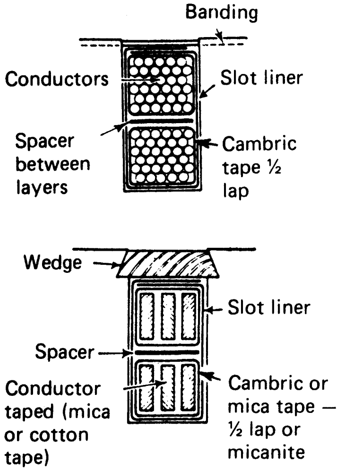


Figure 20.27 Typical slot sections

and the interior of the motor. As cooling must then be solely by dissipation from the outside of the carcass, the rating is limited to about 75 kW. Cooling can be improved by shaft-mounted fans, by inlet and outlet pipes or by closed-circuit ventilation. In the latter case the fan-assisted air circulating through the machine is cooled by passing it through an air/air (c.a.c.a.) heat-exchanger or an air/water (c.a.c.w.) exchanger mounted on the motor frame.

20.3.2.3 Commutation

Modern motors can be made to commutate sparklessly up to 1.5–2 p.u. load. To secure this behaviour, commutating poles are fitted to all motors in the integral-kilowatt range.

Compoles Before the commutator sectors connected to a particular armature coil reach a brush, the coil will be carrying current in a certain direction; while the sectors connected to the coil are passing under the brush the coil will be short circuited by the brushes, and after leaving the brushes the coil will be carrying current in the opposite direction. The current must thus be reversed during the time for which the coil is short-circuited (the time of commutation). At normal commutator peripheral speeds of 10–30 m/s this time will usually lie between 2.5 and 0.2 ms. Owing to the inductance of the coil, the current cannot reverse in this time without some external assistance; it is necessary to induce in the short-circuited coil an e.m.f. to assist the change of current, i.e. an e.m.f. in a direction opposite to that of its e.m.f. when, after leaving the commutating zone, it enters to next pole region. Compoles are therefore fitted to influence the coil-sides undergoing commutation; the compoles have an excitation proportional to the armature current and the polarity of the successive main pole. The arrangement (i) neutralises the main armature m.m.f. in the commutating zone and (ii) produces the necessary commutation flux density there. The compole flux required is proportional to the armature current, and the compole windings are therefore connected in series with the armature.

Commutation in a machine not fitted with compoles can be effected by brush-shifting backward (against the direction of rotation) so that the commutating flux is provided by the succeeding pole. This cannot be done if the motor is required to run in both directions.

Sparking Sparking causes burning and pitting of the commutator surface, so intensifying the trouble. The origins of sparking, and the remedies, are as follows.

Mechanical defects The chief causes are: the sticking of brushes in holders; 'high' or 'low' commutator bars; flats, irregularities or dirt on the commutator surface; badly bedded brushes. The commutator can be cleaned while running with a commutator stone, but irregularities make it necessary to grind the commutator. Small sparks between sectors, starting a few centimetres away from the brush, are probably due to partial short circuits caused by dirt on the mica surfaces. Correct bedding is essential to ensure that the brushes carry current over the whole of their contact surface. It can be carried out by adjusting the brush springs to give a fairly heavy tension and passing, first coarse and finally fine, glass-paper between the brush and the commutator; care must be taken to remove all trace of dust after the operation.

Incorrect brush position The correct position of the brush rocker is usually marked by the manufacturer, but it may tend to move in service. If the marking is obliterated, the correct position can be found by determining the neutral position. If the machine has no compoles the brushes will have to be moved backward from the neutral position by two or three sectors, the best position being found by trial. With a compole motor the brushes should be almost exactly on the neutral position, although they may be moved forward by a small distance so that the compole flux adds to that of the preceding main pole and prevents any tendency of the speed to rise as the load comes on, the compole winding acting as a series stability winding. Incorrect spacing between adjacent brush arms may occur; this should be checked very carefully by a steel tape, not by counting the sectors.

Winding defects Open- and short-circuited coils in the armature will cause severe sparking. An open-circuited coil will cause sparks to go all round the commutator with severe burning at the bars connected to the open-circuited coil. A short-circuited coil will cause overheating of the faulty coil and segments and is often due to molten solder falling between the commutator risers. In either case the presence of a fault can be verified by carrying out a drop test.

Incorrect compole excitation The compole strength can be checked by the brush-drop or black-band tests. Incorrect strength can be remedied by inserting or removing thin steel shims between the back of the compole and the yoke: removing a shim increases the air gap and weakens the compole. Weakening can also be obtained by shunting a resistor diverter across the compole winding, but this is not fully effective in transient conditions.

Thyristor-assisted commutation The speed and output limitations imposed by commutation have encouraged attempts to use thyristors to perform the switching function, leaving the brush to act as a simple current collector. Success would enable the voltage per sector (30–40 V peak, 15–20 V mean) to be raised, fewer sectors could be employed and high-power high-voltage d.c. machines achieved. Promising results have been obtained with the arrangement of *Figure 20.28*. The armature coils are connected to two separate commutators with alternately 'live' and 'dead' sectors, the latter shown shaded. In the diagram, brush D has just come fully into contact with active sector 2, and thyristor T_x has been turned on; the e.m.f. E_c in the coil that has just been commutated must, acting through brush A, be enough to turn off thyristor T_y. Separation of the brushes into two parts, AB and CD, is

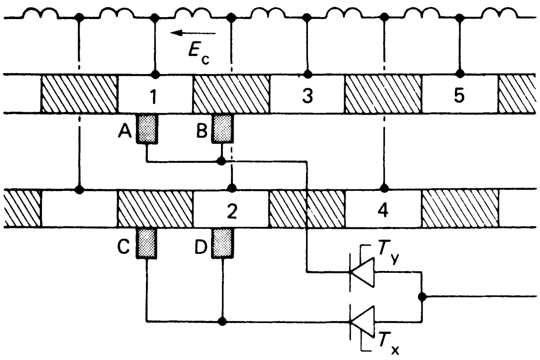


Figure 20.28 Thyristor-assisted commutation

necessary to ensure that the part-brush is fully on to an active sector before it begins to carry current; otherwise contact damage occurs. Although the switching procedure is satisfactory, current collection difficulties, causing commutator damage and arising from the transient current changes, have not yet been overcome.

20.3.2.4 Starting

If a d.c. motor is to be started from a constant-voltage supply, its normally low resistance must be augmented to limit the current to a safe value, e.g. 1.5–2 times full-load value. The starting rheostat is cut out as the motor speed rises and the counter-e.m.f. imposes a limit on the current.

Shunt motor For maximum starting torque, the field must be fully established at starting; the starting rheostat is therefore connected only in the armature circuit. The total starting resistance R is determined by the maximum starting current $I = V/(R + r_a)$, where the armature resistance r_a has the typical values:

Motor rating (kW)	1	2	5	10	50	100
Armature resistance (Ω)						
at 110 V	1.4	0.8	0.2	0.1	0.08	0.006
at 230 V	6.0	3.0	0.8	0.5	0.10	0.025
at 460 V	22	13	3.0	2.0	0.50	0.010

Traditional face-plate starters are obsolescent. Usually a fully automatic push-button system is employed.

Series motor The starting rheostat is in series with the motor, the resistance of which is about twice the value given for r_a in the table above. Industrial series motors are often used in cranes and hoists, and the starting resistance is commonly utilised also for speed control.

20.3.2.5 Speed control: standard motors

A prime reason for the continued use of d.c. motors is the possibility of simple and economic speed control over a wide range. Reference to the expression in Section 20.3.2.1 shows that speed can be controlled by varying the applied voltage, the flux or the armature resistance.

Shunt motor In a shunt machine the variation of the supply voltage does not greatly affect the speed because the result is also a comparable change in flux.

Field control The field current (and therefore the flux) is varied by adding resistance into the field circuit (*Figure 20.29*). For a given setting of the field regulator the speed is approximately independent of the load, giving a series of flat speed–torque curves. The upper limit of speed for a standard motor is about 30–50% above normal, fixed partly by mechanical considerations and partly by weak-field flux distortion. However, a 3:1 range can be obtained by suitable design, although very low speeds cannot be obtained in this way. For a given armature current, a flux reduction raises the speed but reduces the torque to yield a *constant-power* characteristic, *P* in *Figure 20.29*. The loss in the field-regulator resistance is small, so that the efficiency of the machine is not affected.

Armature-circuit resistance control The speed for a given value of resistance added into the armature circuit falls with the load, giving a group of speed–torque characteristics (*Figure 20.29*). The flux remains constant so that for a given current the torque will not vary with speed (*constant-torque* characteristic); power output therefore falls proportionately with speed. Owing to the losses in the added resistance the efficiency is low and approximately proportional to the speed: e.g. with a 60% speed drop the efficiency will be a little less than 40%. The resistance required is $R = x(V - I_a r_a)/I_a$ where x is the desired fractional speed reduction and I_a is the armature current at the reduced speed, the latter depending on the type of load.

Diverter control With series armature-circuit resistance control a large resistance is required to obtain low speed on small load, and the machine is unstable in that there is a large change of speed with load. This can be overcome by adding a variable resistor in parallel with the armature circuit. The efficiency is low, and the method is justified only as a temporary measure or with very small motors.

Ward–Leonard control The main d.c. motor M is separately excited with a constant field current, the armature being supplied with a controlled variable voltage obtained from a d.c. generator G driven by a constant-speed motor (*Figure 20.30*). Control of the generator field varies the main motor armature voltage and consequently the armature speed: a range of 25:1 is obtainable, and reversal is possible if the generator field excitation can be reversed. Each setting of the generator field provides for a ‘shunt’ operating characteristic in the main motor, with a torque proportional to the armature current. The method is economical in energy and is applicable to mine winding gear and machine-tool drives, but it is high in capital cost; consequently for smaller ratings the motor–generator is replaced by a thyristor bank.

Series motor The three basic methods of speed control are shown in *Figure 20.31.*

Field control This is obtained by a diverter rheostat in parallel with the field circuit. Only on small machines is a continuous speed variation obtainable, and the diverter must generally be varied in one or two discrete steps. An alternative, commonly used with traction motors, is to tap each field winding so that part of the winding is cut out to reduce the field m.m.f. and raise the speed. Both methods can give only a speed rise. In some cases it is possible to arrange the field windings in two groups, which can then be connected in series or parallel, the latter giving 20–30% higher speed for a given current than the former.

Resistance control A variable resistor in series with the motor reduces its terminal voltage and lowers the speed.

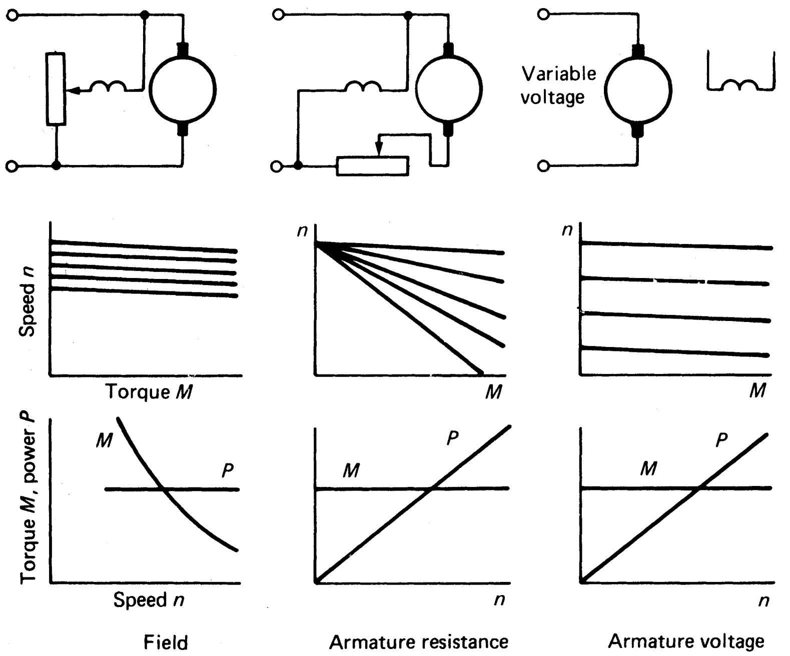


Figure 20.29 Shunt motor: speed control

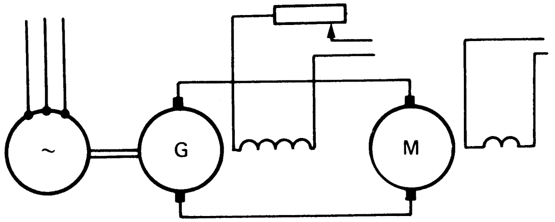


Figure 20.30 Ward–Leonard control

Although the method involves I^2R loss, it is commonly employed for cranes, hoists and similar plant, the resistance steps being used also for starting.

Series/parallel control If two series motors are connected mechanically to ensure the same speed for each (as is usual in d.c. traction systems) series/parallel voltage control can be obtained by connecting the motors electrically in parallel or in series, the former giving full voltage and the latter one-half voltage to each motor. Intermediate speeds can be obtained by resistance or field control. The full-parallel speed is, for a given motor current, approximately twice that in full-series.

A scheme known as *parallel/series* control is sometimes applied to battery vehicles, the battery being arranged in halves that can be paralleled for starting and low speed, and in series for full speed.

20.3.2.6 *Speed control: thyristor-fed motors*

A separately excited motor may be fed from a constant-voltage a.c. supply through thyristors, which rectify the current and also (by delaying the commutation angle α by controlled gate signals) furnish a variable-voltage supply to the armature. The field current is obtained from the a.c. supply through semiconductor diodes or thyristors. The main thyristor equipment thus replaces the motor-generator set of the Ward–Leonard speed control in ratings up to about 500 kW and, being static, is more economical and commercially viable even down to fractional-kilowatt sizes.

Connections The choice of thyristor circuit is a compromise between cost (i.e. the fewest thyristors, which with their firing circuits are more expensive than diodes) and operational difficulties arising from harmonic production or poor commutation, both accentuated by the use of a small number of thyristors. For economy, half-controlled circuits in which half the units are thyristors and half are diodes are common, but such circuits cannot regenerate.

The thyristors may be fed direct from an a.c. supply of r.m.s. voltage V_a as in *Figure 20.32*. The mean direct voltage available with zero commutation delay ($\alpha = 0$) is $V_{a0} = 0.9\,V_a$ for the one-phase and $V_{d0} = 0.43\,V_a$ for the three-phase arrangement. With commutation delayed by angle α the mean direct voltages are

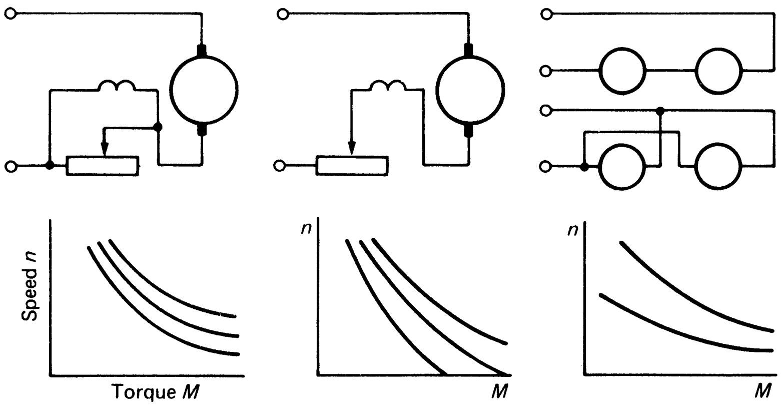
Fully controlled $\quad V_d = V_{d0} \cos \alpha$

Half-controlled $\quad V_d = V_{d0} \frac{1}{2}(1 + \cos \alpha)$

The one-phase half-controlled arrangement of *Figure 20.32*(*a*) is very widely used for ratings up to about 5 kW. The current tends to be discontinuous, causing bad commutation, and sufficient inductance should be included in the circuit to avoid this except at large delay angles; a separate inductor may be used or, in the smaller ratings, the motor winding may be designed to have a sufficiently high inductance.

The three-phase half-controlled circuit (*b*) has been used for outputs up to about 200 kW but recent practice tends towards the use of the fully controlled circuit from about 25 kW up to 1000 kW. The fully controlled circuit (*c*) gives rise to a 300 Hz ripple on the direct voltage but the half-controlled circuit gives 150 Hz over most of its range, rising to 300 Hz when α approaches zero.

The transformer-fed networks in *Figure 20.32* give the designer a free hand in choosing the operating voltage of the d.c. motor. The one-phase centre-tap connection (*d*) has



Field diverter   Resistance   Series/parallel

Figure 20.31 Series motor: speed control

the advantage over the one-phase bridge of making regeneration possible.

By connecting in series two six-pulse units, the supplies to which are obtained respectively from delta- and star-connected windings of a three-winding transformer (*Figure 20.32(e)*) a 30° phase-shift is obtained giving a 12-pulse operation and a 600 Hz ripple of small amplitude. Such an arrangement is generally desirable for outputs above 1 MW and up to 5 MW (the upper limit for a d.c. motor).

The field supply for the motor is generally obtained from the a.c. supply through a 1-phase bridge or centre-tap connection using diodes. If field control is desired in addition to armature voltage control it can be achieved by a conventional field resistor or by using thyristors instead of diodes.

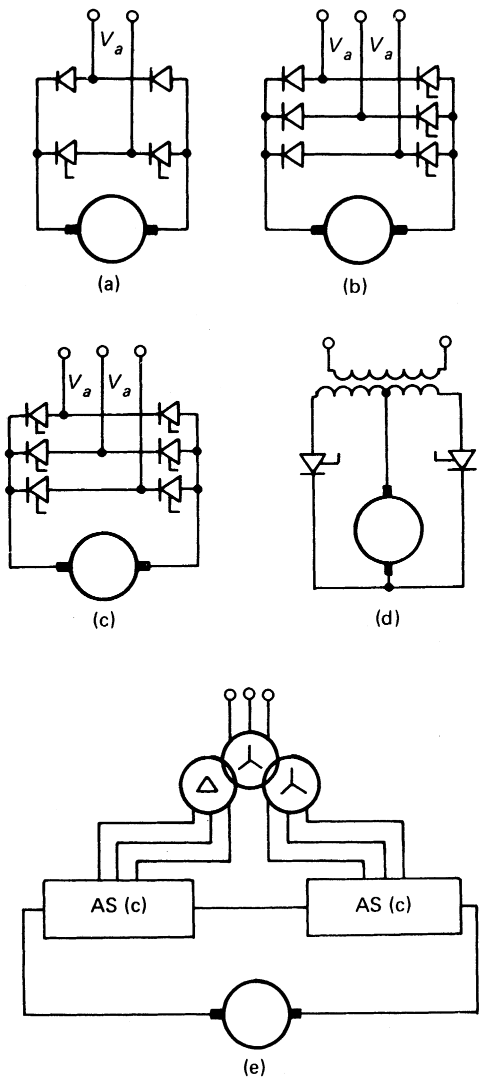
Harmonics In addition to the harmonics on the d.c. side which may interfere with commutation, harmonics also appear on the a.c. side and can cause interference with communication and control circuits and difficulties with other plant connected to the system, particularly overloading of shunt capacitors and possible resonances at the 11th and 13th harmonics. For these reasons the rating of plant connected to a single point on the supply is limited by the Supply Authority, typically from 250 kW at 0.4 kV to 3 MW at 33 kV for six-pulse units, and from 750 kW to 7 MW for 12-pulse units.

Motor construction As a result of harmonics on the d.c. side, a motor may have to be derated by 15–20%. To minimise commutation troubles, compoles and main poles and yokes may all be laminated.

Thyristor firing The signals applied to the thyristor gates are usually high-frequency pulses (2–8 kHz). For large units the pulse current may be 1–2 A in amplitude with a rise-time less than 1 μs. The electronic equipment to generate the signals comprises a power supply, timer and phase-shift units, pulse generator and amplifier, and output unit. The control of the gate pulse may be manual, or by signals from a closed-loop control system.

20.3.2.7 *Braking*

Rheostatic, regenerative and plug braking are applicable, but they cannot hold a motor at rest: for that a mechanical brake is necessary.



(a) (b) (c) (d) (e)

Figure 20.32 Thyristor control

Shunt motor The electric braking methods are shown in *Figure 20.33*.

Rheostatic braking The field connection to the supply is maintained but the armature is disconnected and then reconnected on to a resistor. The machine generates, dissipating power in the resistor. The braking effect is controlled by varying the field current. For a total armature-circuit resistance R, the armature current is $I = E/R = k_1 n\phi/R$ and the braking torque is $M = k_2 EI/n = k_3 n\phi^2/R$. If the excitation is constant, then the braking torque is directly proportional to the speed n and decreases as the motor speed falls.

Plugging The armature connections are reversed, and the motor torque tends to retard the machine and then run it up in the opposite direction. The applied voltage and the armature e.m.f. are additive, so that a resistance of about twice starting value must be included to limit the current. For a total armature-circuit resistance R, the armature current is $I = (V + E)/R = (V + k_1 n\phi)/R$ and the braking torque is $M = (k_4\phi + k_3 n\phi^2)/R$. With constant excitation the braking torque is $k_5 + k_6 n$. Braking by plugging gives a greater torque and a more rapid stop, but current is drawn from the supply during the braking period, and this energy together with the stored kinetic energy has to be dissipated in resistance. A relay must be provided to open-circuit the motor at rest in order to prevent it from running up in reverse.

Regenerative braking If a load (such as a descending hoist) overruns the motor at a speed higher than normal, the counter-e.m.f. E exceeds the terminal voltage V and the machine generates. This is a very convenient method of 'holding' a load, but not at low speeds unless excessive field current is supplied.

With *thyristor-controlled* motors regeneration is possible if fully controlled connections are used. It is necessary to reverse the polarity of the motor terminals relative to those of the thyristor unit. This can be done by reversing either the armature or the field terminals at the moment of entering regeneration, the reversal being effected by conventional reversing contactors. Armature reversal requires about 150 ms but field reversal may require up to 2000 ms. Where a very rapid reversal is required it is necessary to employ separate thyristor units for motoring and regenerating, these being connected in 'anti-parallel'. Only switching of the gate pulses is then required, a process that can be achieved in about 10 ms.

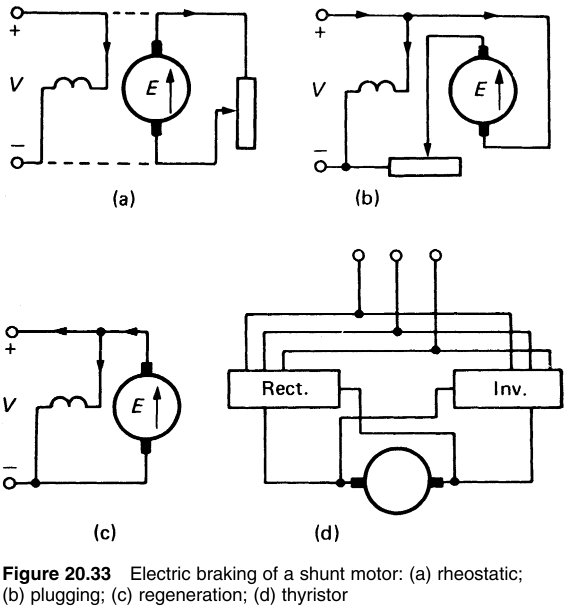


Figure 20.33 Electric braking of a shunt motor: (a) rheostatic; (b) plugging; (c) regeneration; (d) thyristor

Series motor The electric braking methods are shown in *Figure 20.34*.

Rheostatic braking The motor acts as a series generator loaded on resistance. It is necessary to reverse the field connections at the instant of changing from motoring to braking. Further, the load resistance must be below a critical value if the machine is to self-excite. In practice the starting resistance, the value of which is well below the critical, is used for braking. The braking torque is approximately $M = kI^2$.

Plugging The conditions are generally similar to those in the shunt machine.

Regenerative braking This is not practicable with series motors. In traction, regeneration is sometimes effected by separately exciting the motors.

20.3.2.8 Design data

Some general and typical data are given here. Small motors up to 5 kW at 1000 rev/min normally have two poles, up to 50 kW four poles and up to 200 kW six poles. Larger motors have more, the number being such that the 'speed frequency' is 20–30 Hz and the pole-pitch between 45 and 55 cm.

The rated output power P is related to the main dimensions (diameter D, length l), the speed n and the specific magnetic and electric loadings B and A by

$$D^2 ln = P/\pi^2\eta BA$$

where B is the mean gap flux density of value up to about 0.55 T and A varies between 5 and 35 kiloampere-conductors per metre of armature periphery.

The maximum safe peripheral speed is 30–40 m/s for the armature and 25–30 m/s for the commutator. The number of slots is about six per pole for fractional-kilowatt motors and 10–14 per pole for machines of 100 kW. The total current per

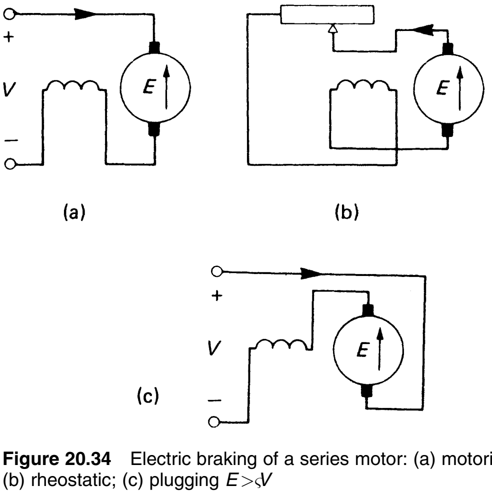


Figure 20.34 Electric braking of a series motor: (a) motoring, $E < V$; (b) rheostatic; (c) plugging $E > V$

slot is about 700 A for the latter rating. The number of slots must suit the winding required. A simple wave winding is used for nearly all industrial motors except in ratings of several hundred kilowatts for which lap (parallel-circuit) windings with equalising connectors may be used. The number of conductors is $Z = E/(p/a)n\phi$; the number of coils and commutator sectors is $2pV/e$, where the mean voltage e between sectors is usually between 12 and 15 V.

The field m.m.f. is 1.15–1.25 times the full-load armature m.m.f., and in designing the magnetic circuit the pole pleakage (about 15%) must be taken into account. The length of the air gap is governed by the requirement that the field distortion shall not be excessive.

20.3.2.9 Disc motors

The disc-armature machine has been developed for special applications such as small pumps, freezer-compressor drives, domestic sewing machines, computer spool drives and light battery vehicles. The basic geometry can be realised in the diagrams shown in *Figure 20.35*. A ring of permanent magnets of alternate polarity provides the axial field. In (i), two magnet rings M flank the armature A. In (ii), one of the magnet rings is replaced by a flux-conveying steel 'yoke' Y. In (iii), the yoke is carried by the armature, increasing its inertia but reducing the effective air gap length. The wave or lap winding (of which a single turn is shown) has an angular span equal to the angular pole-pitch. A multipolar structure, in which a front-face conductor is paired with another on the back of the armature disc, is employed to reduce the length of the 'overhang'. In small low-inertia motors the winding is punched from a copper or aluminium sheet and then placed on either side of an epoxy-resin-impregnated glass-fabric disc. The resin is then given a completion cure. Larger motors have wire-wound armatures with face or barrel commutators.

The magnets are usually of sector rather than circular shape to increase the working flux density towards the outer radius, where it is the most effective. Consider a machine in which the magnets, set between outer and inner radii R and r, produce in this annulus a uniform mean flux density B. Then for an angular speed n and a current loading A (in A/rad) the e.m.f. in a conductor and the torque developed are respectively.

$$E_r = \pi(R^2 - r^2)Bn \quad \text{and} \quad M_e = \tfrac{1}{2}(R^2 - r^2)BA$$

showing the influence of the outer radius in increasing the output. However, R is limited by rotational stress, flexure of the disc and any operational constraints on the armature inertia.

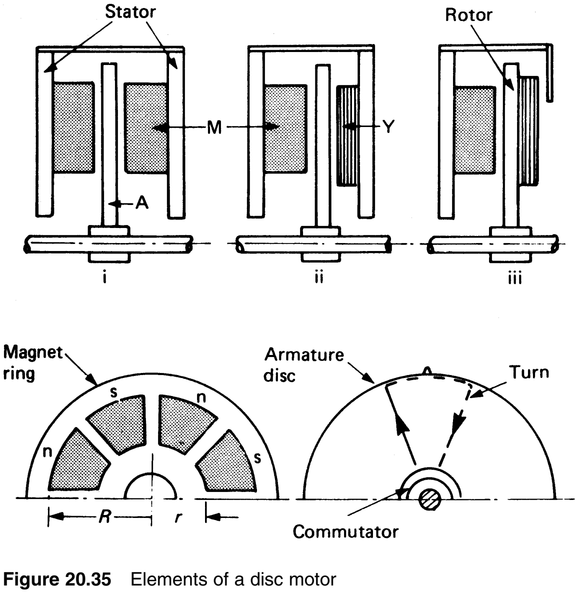


Figure 20.35 Elements of a disc motor

Homopolar motors A homopolar d.c. motor is an embodiment of the Faraday disc. The voltage is inevitably low (less than 200 V) and the current high. Such machines as generators have been built to supply electromagnetic pumps and electrochemical processes requiring very large currents. Homopolar motors have also been built, notably a 2.5 MW 200 rev/min machine for a power-station pump where the motor has a superconducting field system developing a working flux density of nearly 4 T by means of an m.m.f. of 3×10^6 A-t.

20.3.3 Three-phase induction motors

The great majority of industrial, commercial and agricultural electric motors above the fractional-kilowatt size are three-phase induction machines, on account of their simple, cheap and robust construction and the almost universal availability of three-phase supplies.

The induction motor has a 'shunt' speed–torque characteristic, the operating speed n falling slightly below the synchronous speed $n_s = f/p$, where f is the supply frequency and p is the number of pole-pairs for which the three-phase stator winding is arranged. The drop in speed below n_s is the *slip*, given by $s = (n_s - n)/n_s$. The slip increases from nearly zero on no load to 0.03–0.05 on full load. Most industrial motors are four-pole machines with a synchronous speed of 1500 rev/min, but two-pole machines ($n_s = 3000$ rev/min) are sometimes of use, and lower speeds obtained with six, eight or more poles may be used in large ratings.

The *cage* motor has a rotor winding internally short circuited on itself with no external access. The *slip-ring* motor has the polyphase rotor winding brought to three slip-rings so that connection can be made to it for starting or speed control; for normal operation, however, the rotor winding is short circuited. In each type the stator carries a conventional three-phase winding, fed from the main supply and generally delta-connected. Almost all induction motors are of the cage type, with slip-ring machines used normally only in ratings above about 100 kW.

20.3.3.1 Operating principle

Currents in the stator windings set up an air gap travelling-wave magnetic field of almost constant magnitude and moving at synchronous speed. The field cuts the rotor conductors at slip speed, inducing a corresponding e.m.f. and causing currents to flow in the short-circuited windings. The interaction of these currents with the travelling-wave field produces torque to turn the rotor in the direction of the field. The magnitude of the rotor currents depends on the slip and on the impedance (comprising resistance, and inductive reactance proportional to slip) of the rotor windings. With the rotor running at synchronous speed, the rotor slip is zero, the rotor inductive reactive vanishes and, as the gap flux does not cut the rotor conductors, the induced e.m.f. is zero; as a consequence there is no rotor current and no torque is developed. Since there must always be a small torque to overcome mechanical loss, the motor cannot quite achieve synchronous speed. As the motor is loaded, it slows so that the slip becomes a small finite quantity, rotor e.m.f. is developed and rotor current flows; the rotor circuits are mainly resistive but a small inductive reactance is introduced. The various interactions yield the torque-speed curve A in *Figure 20.36*. It has been taken into the region of reverse rotation (slip greater than unity), for which the machine acts as a brake. The normal working range of the machine as a motor is the region for small positive slips: here the torque-speed relation is almost linear, corresponding to that of the d.c. shunt motor.

Reversal of the direction of rotation of the motor is obtained by interchanging two of the stator terminal connections, thus reversing the direction of the travelling-wave field.

20.3.3.2 Equivalent circuit

The performance is most readily predicted with the aid of an equivalent circuit (*Figure 20.37*) assembled from the various resistances and inductances (i.e. magnetising and leakage) of the machine, taken as independent of current, frequency and saturation conditions. The essential parameters are as follows, it being assumed for convenience that the rotor and stator windings are identical; all electric-circuit quantities are *per phase*:

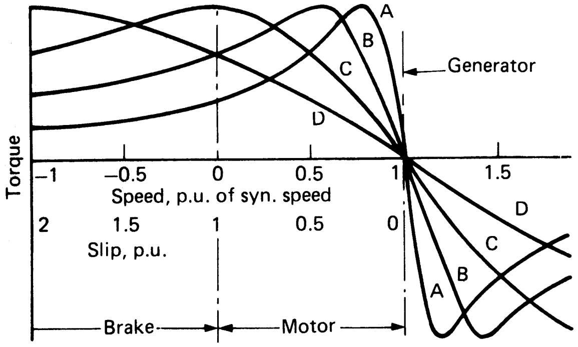


Figure 20.36 Three-phase induction motor: torque–speed characteristics

V_1	stator applied voltage
E_1, E_2	stator e.m.f., rotor e.m.f. at standstill
r_1, x_1	stator resistance and leakage reactance
r_2, x_2	rotor resistance and leakage reactance at supply frequency (corresponding to standstill, $s = 1$)
r_m, x_m	resistance representing core loss, magnetising reactance
I_1, I_2	stator current, rotor current
I_0	no-load current given by $\sqrt{(I_m^2 + I_c^2)}$, where $I_c = \Leftarrow E_1/r_m$ and $I_m = E_1/x_m$

The basic equivalent circuit is shown in *Figure 20.37*(*a*): it is similar to that of a transformer on short circuit except that the transformer ratio varies with slip (and therefore with the speed). With unity turns-ratio and a division of the rotor parameters by the slip s, equivalent circuit (*b*) is obtained. For easier calculation the approximate circuit (*c*) can be used, with an error not exceeding 2% or 3% provided that the operating conditions are not abnormal. The rotor resistance r_2/s has been split into r_2 for the rotor I^2R loss, and $r_2(1-s)/s$ in which the I^2R value represents the power conversion to the mechanical form.

Equations The following relations can be developed from the approximate equivalent circuit:

Rotor current: $I_2 = V_1/[(r_1 + r_2/s)^2 + (x_1 + x_2)^2]^{1/2}$
Stator current: $I_1 = I_2 + I_0$
Power division: rotor input : rotor I^2R: gross output $= 1 : s : (1 - s)\Leftarrow$
Gross torque: $M = V_1 I_2 (r_2/s)/2\pi n_s$

The peak torque, which is independent of the actual value of rotor resistance, is approximately

$M_m = \Leftarrow V_1^2/(x_1 + x_2)4\pi n_s$ at $s = \Leftarrow r_2/(x_1 + x_2)\Leftarrow$

and its value is normally 2–2.5 times the full-load torque.

The losses in the machine comprise the core loss, stator and rotor I^2R loss, and mechanical loss in windage and friction. The no-load input current is 0.25–0.3 of the full-load current, at a lagging power factor in the range 0.15–0.2. At standstill on normal voltage the stator current is 4–6 times full-load current. Typical full-load values for conventional induction motors are:

10 kW motor:	p.f., 0.87 (4-pole), 0.75 (12-pole); efficiency, 0.85
1000 kW motor:	p.f., 0.94 (4-pole), 0.91 (12-pole); efficiency, 0.95

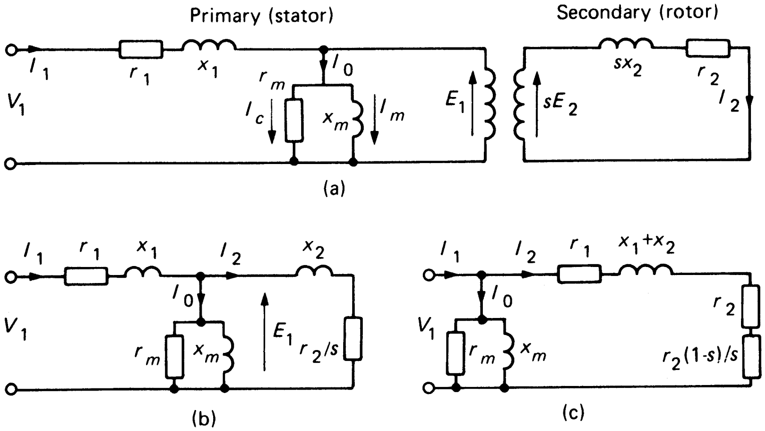


Figure 20.37 Three-phase induction motor: equivalent circuit per phase

20.3.3.3 Construction

As the working flux is alternating, the stator and rotor cores must be laminated, using plates 1.0–1.5 mm thick.

Stator The core comprises annular stampings for small and segmental plates for large machines; it is mounted in a welded steel frame that does not form part of the magnetic circuit.

The voltage for which the stator is wound is normally between 380 and 440 V for motors up to 250 kW. Larger machines are wound for higher voltages, the minimum economic sizes being about 250 kW for 3.3 kV, 400 kW for 6.6 kV and 750 kW for 11 kV.

Rotor Slip-ring rotors may be wound to develop an e.m.f. when stationary of about 100 V for small and up to 1 kV for large machines, with insulation to correspond. The winding of a cage rotor comprises copper or aluminium bars located in slots (usually without insulation) and welded or brazed at each end to a continuous end-ring. The joints between bars and end-ring may prove to be points of weakness unless carefully made. Small cage motors generally have aluminium bars and rings cast into the rotor in one piece, with the end-rings shaped to form simple fan blades. To minimise the magnetising current the air gap is made as small as is mechanically practicable, e.g. from 0.25 mm for small motors up to about 3 mm for large motors.

Enclosure This follows standard practice. The *flameproof* construction is available for motors used in hazardous atmospheres classified as 'Division 1 Areas' (mines, petroleum plants, etc.): it is a total enclosure with all joints flanged so that any flame generated by an internal explosion will be cooled by its passage through the joint and will not ignite external explosive gases. Cage motors, which have no slip-rings, are well suited to such situations, and they may be used without flame-proofing in 'Division 2 Areas' where flammable gas is not present unless there is some breakdown in the plant.

20.3.3.4 Starting

The factors of importance are (i) the starting torque and (ii) the starting current drawn from the supply. If the motor is to be started on full load, the starting torque must be 50–100% above full-load torque to overcome static friction and to ensure that the motor runs up in a reasonably short time to avoid overheating. Lower values are acceptable if the motor is always to be started on no load, and may be desirable in order to avoid too abrupt a start. The starting current should be as low as practicable to avoid overheating. Occasionally, supply authorities limit the starting current that may be drawn from the supply in order to avoid excessive voltage drops interfering with other consumers; it must also be remembered that, since torque is proportional to (voltage)^2, any voltage drop will significantly reduce the available starting torque. The torque and current may be expressed in terms of full-load values, but a more significant comparison between motors of different efficiencies and power factors is had by expressing the starting kVA in terms of the full-load output in kilowatts.

The curves A, B, C, D in *Figure 20.36*, drawn for different resistances, show that the rotor resistance has a major effect on the starting torque.

Slip-ring motor By adding external resistance to the rotor circuit any desired starting torque, up to the maximum-torque value, can be achieved; and by gradually cutting out the resistance a high torque can be maintained throughout the starting period. The added resistance also reduces the starting current so that up to 2–2.5 times full-load torque can be obtained with 1–1.5 times full-load current.

Cage motor For high efficiency, the rotor resistance must necessarily be low, and the starting torque on direct starting (when the motor is switched on to full voltage) is likely to be 0.75–1 times full-load torque with a stator input of 4–6 times full-load current. The table below gives typical values of the ratio (starting kV-A)/(full-load kW) for a range of conventional motors with direct starting are:

Range, kW:	1–6	6–40	40–250	250–500	500–1500	1500–4000
kV-A/kW:	10	9	8	7.7	7.4	7.2

These values are acceptable to supply authorities in most cases.

From the relations already given for the division of the power input to the rotor, the starting torque M_s for a starting current I_s is, in terms of the full-load values M_1 and I_1 and the full-load slip s_1, given by

$$M_s/M_1 = (I_s/I_1)^2 s_1$$

If direct starting is not admissible by reason of the initial current and/or the impulsive torque, then the voltage must be reduced for starting, bearing in mind that torque is proportional to (voltage)^2.

Series resistance A resistor in each line to the stator terminals can reduce the current to any fraction x of the direct-starting value, but the torque will be the fraction x^2. Although cheap and simple, this method is acceptable only for motors that start on no load.

Autotransformer Usually no more than three tappings (50%, 70% and 80%) are provided. With a tapping giving the fraction x of normal voltage, both current and starting torque are reduced to x times the direct-start values. The cost of the transformer and contactors is high, especially if special connections are used to avoid disconnection of the supply when tap-changing.

Star–delta switch This starts the motor in star connection, and changes it to the normal delta connections as the speed approaches normal. Both current and torque at starting are one-third of their respective direct-start values. All six stator phase-ends must be available. The momentary disconnection can cause significant transient effects during the change-over.

Current displacement This relies on the change of leakage reactance with rotor frequency to give enhanced starting and run-up torque, but with some sacrifice in pull-out torque. The inner cage in (*a*) and the inner region of the deep bar in (*b*) of *Figure 20.38* have a high leakage inductance, and at low speed (i.e. at higher rotor frequency) current is forced mainly into the outer cage in (*a*) and the upper region of the deep bar in (*b*). Thus most of the rotor current flows in the higher resistance outer cage or in the constricted region of the deep bar.

Solid state For loads with a rising torque/speed load demand, a 'soft' start can be obtained by including anti-parallel thyristors in the stator circuits, using phase control. At starting, stator voltage control can be obtained by appropriate triggering, or stator impedance control by use of a gapped-core inductor.

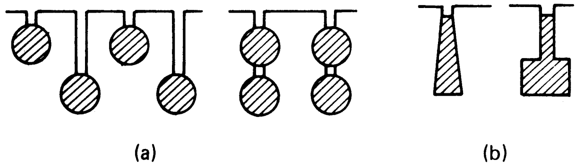


Figure 20.38 High-torque rotor cages: (a) double cage; (b) deep bar

20.3.3.5 Speed control

The speed for an induction motor is $n = (1 - s)f/p$. The frequency f is normally fixed, the machine is built with p pole-pairs and, as the operating slip s lies generally between the limits 0.03 and 0.05, the motor is a substantially constant-speed machine with a working range as shown on curve A of *Figure 20.36*. Speed variation is often needed and, with some additional cost and complication, can be achieved by varying the slip, the number of poles or the frequency. For small motors a limited control can also be obtained by varying the applied voltage.

Slip control This, applicable only to slip-ring motors, requires connection into the rotor circuit of a device producing an adjustable volt drop or counter-e.m.f. The rotor induced e.m.f. sE_2 must overcome this to enable torque-producing current to flow, so that the slip must change in accordance with the magnitude of the injected volt drop or e.m.f.

Resistance control A variable volt drop is set up in the rotor circuit by variable resistors in each phase, giving a series of torque–speed characteristics (A, B, C, D in *Figure 20.36*). Only the low-slip parts can give normal operation, as the lower speeds may be unstable. The starting resistors, if continuously rated, can be used for speed control. The method is cheap and simple, but results in high I^2R ('slip energy') especially for low speeds. The efficiency is a little less than $1 - s$, the speed varies widely with load, and low speeds are not obtainable at low loads. This form of control is used only where small or infrequent speed reductions are called for.

Slip-energy recovery Here the slip energy is not dissipated in resistance but is returned to the supply (constant-torque drive, as with resistance control) or added to the shaft output of the main motor (constant-power drive). Commutator machines have in the past been employed to deal with the slip energy. A modern method (*Figure 20.39*) is to rectify the slip-frequency currents in a diode bridge network; the unidirectional output current is smoothed and passed on to a three-phase line-commutated inverter at a rate depending on the supply voltage, the rectified direct voltage and the thyristor firing angle. The inverted current has a fixed waveform and a constant conduction angle of $2\pi/3$ rad. The onset of conduction with respect to the phase-voltage zero is controlled by the firing angle. As power flow through the rectifier is unidirectional, only subsynchronous speeds are feasible.

Pole changing Switching the stator winding to give two (sometimes three) different numbers of pole-pairs gives two (or three) alternative running speeds. Cage rotors are normally employed, as slip-ring windings must be pole-changed to correspond always to the stator. Pole-changing motors with a 2:1 ratio have been used for many years, a typical arrangement being that in *Figure 20.40*. A more recent innovation is the pole-amplitude-modulated method.

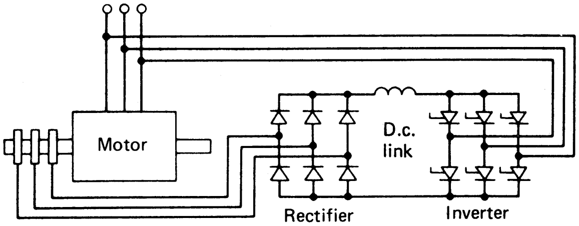


Figure 20.39 Thyristor slip-energy recovery

Pole-amplitude modulation The m.m.f. (and flux) distribution around the air gap produced by one phase of a conventional machine can be expressed as $F(\theta) = A \sin p\theta$ςwhere θς represents an instantaneous angular position. If $F(\theta)$ is modulated by making $A = C \sin k\theta$, the m.m.f. becomes

$$F(\theta) = \tfrac{1}{2}C[\cos(p - k) - \cos(p + k)]\Leftarrow$$

which is an m.m.f. comprising two superimposed waves of pole-number $p - k$ and $p + k$. One wave can be eliminated by adjusting the chording and relative position of the phase windings so that the machine has p pole-pairs (unmodulated) and either $p - k$ or $p + k$ pole-pairs (modulated). The modulation is effected by reversing half of each phase winding and, in some cases, isolating certain coils. Two basic forms of connection are given in *Figure 20.41*. The coils that are isolated for the unmodulated connection are in the sections A′A, B′B and C′C. For designs that do not need coil isolations the phase terminals during modulation are A′B′C′⇐

In general, when the distribution of coil groups per pole and per phase is not uniform, some coils are isolated for modulation; with a more uniform distribution, however, a simple reversal of the second half of each phase winding is sufficient. The latter is more suitable for power outputs that are required to be similar at the two speeds.

This simple theory led to the design of successful industrial motors with close speed ratios from, for example, 8/10 or 12/14 poles. The recent development of a more general theory has made possible wide ratios from, for example, 10/2 and 16/4 poles, and also three-speed motors. There is virtually no limit to the size and speed ratio available, even in fractional-kilowatt ratings. The pole-amplitude-modulated (p.a.m.) motor is thus superseding the two-winding change-speed motors commonly used in the past, as the starting torques, power factors and efficiencies of p.a.m. machines are comparable with those of single-speed standard machines and have optimum performance on all ranges. The rating for a given frame-size is about 90% of that of a normal single-speed motor.

Frequency variation Development of the static thyristor inverter has made possible the provision from a three-phase supply of an a.c. source of controllable voltage and of frequency infinitely variable from zero up to three or more times the supply frequency. Wide speed variation with such a source is possible with a cage motor. To maintain a constant motor flux, the voltage applied to the motor must be proportional to the frequency. The speed-torque characteristics (*Figure 20.42*) show that the peak torque is approximately the same at all operating frequencies.

D.c.-link converter A rectifier converts the a.c. of mains frequency to d.c., and a thyristor inverter converts this to a.c. of the desired frequency in the d.c.-link converter

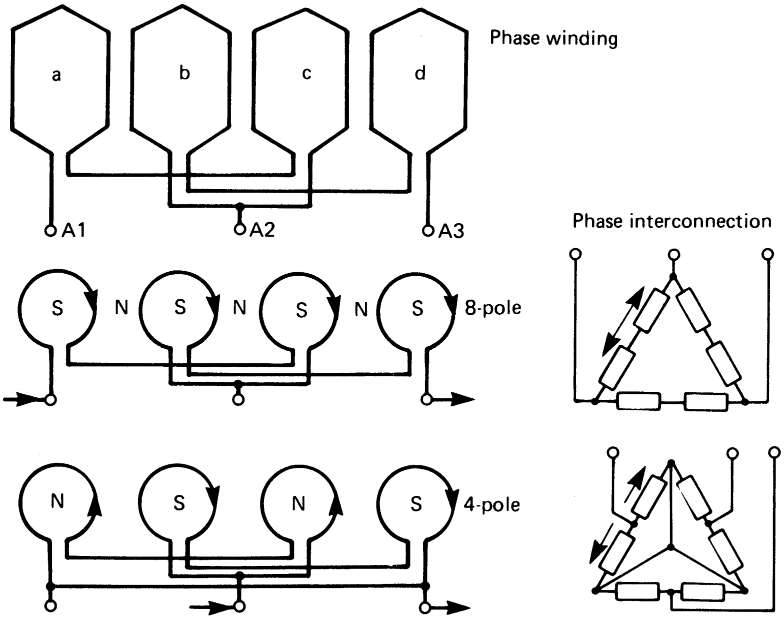


Figure 20.40 Pole change in the ratio 2:1

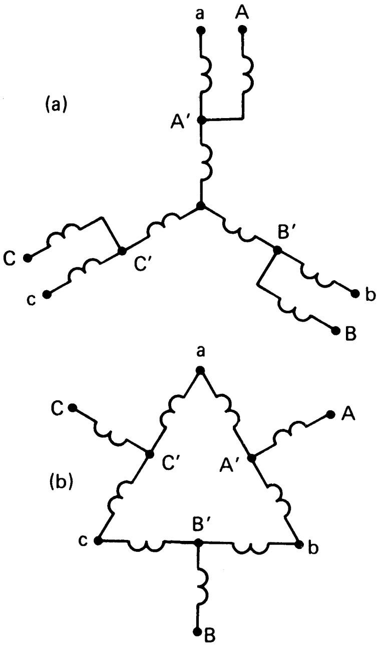


Figure 20.41 Pole-amplitude modulation. Unmodulated/modulated connections: (a) parallel-Y/series-Y; (b) parallel-Y/series-delta. Poles $2p$: supply to ABC with abc joined. Poles $2(p \pm 1)$: supply to abc with ABC isolated

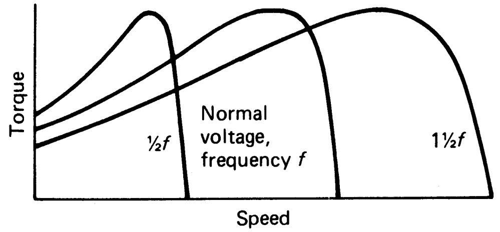


Figure 20.42 Torque–speed relationships for variable frequency

(*Figure 20.43*). The direct voltage is varied by a thyristor chopper instead of by a controlled thyristor equipment in order to mitigate harmonics in the supply. The inverter produces a waveform with about 20% of fifth and 14% of seventh harmonic, and these may cause small additional losses in the motor; however, the impairment may be reduced by series inductors in the motor circuit. It must be remembered that although the motor has the usual overload capacity, the converting equipment has no such reserve and must be rated for the peak power.

An upper frequency of about 150 Hz is usual, permitting the speed range of a two-pole machine to be from zero to 9000 rev/min. For small machines and with thyristors having very fast switching characteristics, higher frequencies can be generated and speeds up to 100 000 rev/min achieved.

Cyclo-converter In this alternative equipment, sections of the normal mains-frequency wave are selected and used to build up an outgoing wave of lower frequency, usually not higher than about two-thirds of the mains frequency, in a single static unit. The equipment has an efficiency higher than that of the d.c.-link converter, but a considerably more complicated control network: basically 18 thyristors are required for a three-phase motor. The cyclo-converter employs natural commutation and may be more reliable than the forced commutation in the d.c.-link converter under abrupt changes of load; it can also be made reversible for regenerative braking without the additional complication that a d.c.-link converter would involve for this duty.

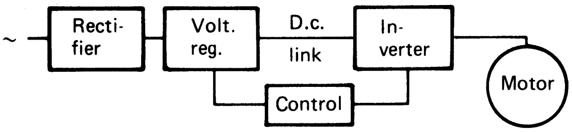


Figure 20.43 Basic d.c.-link converter system

Voltage variation As the peak torque of an induction motor is proportional to the square of the voltage, a limited speed control can be effected by voltage variation. Typical speed–torque curves for a small motor operating at 100% and 70% of rated voltage are given in *Figure 20.44*. For a constant-torque load the speed is reduced from A to B; for a torque proportional to the square of the speed, as for a fan, the greater reduction from C to D is obtained. Greater variation occurs if the motor has a high rotor resistance/reactance ratio, so that the method is ineffective for motors of rating more than a few kilowatts; it is, in fact, used chiefly with small mass-produced motors. The voltage variation may be achieved by means of series rheostats (cheap but wasteful) or by thyristors in the supply circuit (expensive, efficient and harmonic-producing).

20.3.3.6 Braking

Braking may be required to bring a motor and its load to rest rapidly in an emergency or as part of a production process, or to hold the motor at a set speed against gravity, as in a descending hoist. Mechanical brakes must be used if a motor is to be held at rest, but motional braking can be electrical and may not need much auxiliary control equipment.

Plugging If a motor, operating with a small slip s, has a pair of its supply leads interchanged, the direction of its travelling-wave air-gap field reverses and the slip becomes $2 - s$. A braking torque is developed, retarding the motor towards standstill, $s = 1$. The braking torque normally varies from about one-half the value of the starting torque initially, up to the starting torque when the machine comes to rest, and the braking current throughout is approximately equal to the starting current. Unless the motor is disconnected when it stops, it will start up again in the opposite direction.

D.c. injection The three-phase supply to the stator is disconnected and immediately replaced by a d.c. supply giving a stationary field in the air gap. The rotor conductors, moving in this field, develop e.m.f.s and currents that exert a braking torque, initially about equal to the starting torque but falling with the speed and vanishing as the motor is brought to rest by friction. The direct current is fed into the stator by two terminals, leaving the third isolated if the stator winding is star-connected.

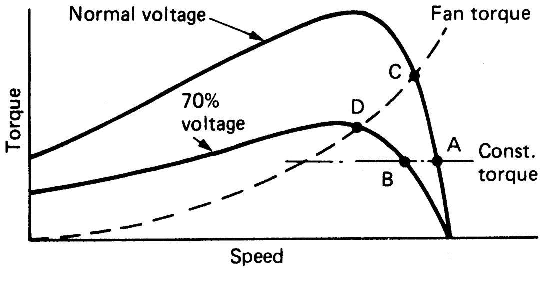


Figure 20.44 Torque–speed reationships for variable voltage

Regeneration When a load overdrives the motor to a speed exceeding synchronous (i.e. with negative slip) as with a descending hoist load, the machine acts as an induction generator and sets up a braking torque (see *Figure 20.36*). Such braking cannot bring the motor to rest but it can limit the speed to a value a little above synchronous, the power from the load being partly returned to the supply. With two-speed pole-change motors a high braking torque can be obtained if, when running at the higher speed, the motor is switched to the larger pole-number.

20.3.4 Three-phase commutator motors

The recovery of slip energy can be achieved in a single machine by incorporating in the rotor a commutator winding. The *Schrage* (rotor-fed) and *doubly fed* (stator-fed) motors have some commercial importance. Both have a 'shunt' speed–torque characteristic and can operate both above and below synchronous speed. A three-phase commutator motor with a 'series' characteristic is also available.

20.3.4.1 Schrage motor

The Schrage motor has its primary winding on the rotor, connected to the supply through slip-rings and brushes. The rotor also carries a low-voltage commutator winding with conductors located in the same slots as, and above, those of the primary. The secondary winding is on the stator. The primary and secondary windings are similar to those of a conventional induction motor. The brushgear comprises two movable rockers, each fitted with three brush spindles per pair of poles. The two rockers are geared together, each being fitted with a toothed segment. The two segments mesh with pinions fitted to a short shaft to which either a handwheel or a small pilot motor is connected. The gearing is so arranged that the movement of the handwheel or pilot motor causes the brush rockers to move in opposite directions.

The brushes attached to each rocker move over separate portions of the commutator surface to enable these brushes to be placed 'in line' or to be moved in either direction, so that more or fewer commutator sectors are included between a brush on one rocker and the corresponding brush on the other.

The bus-rings of each rocker are connected to the secondary (stator) winding as in *Figure 20.45*, which represents a two-pole machine.

As the primary winding is on the rotor, e.m.f.s of slip frequency are induced in the secondary (stator) winding. The e.m.f.s at the brushes are also of this frequency. The e.m.f. induced in each coil of the commutator winding is constant at all speeds, and therefore the e.m.f.s injected, via the brushes, into the secondary winding are proportional to the number of commutator sectors included between corresponding brushes on the two rockers, i.e. between the brushes connected to a particular phase of the secondary. Thus, when these brushes are in line, the secondary winding has no e.m.f. injected into it and the motor will run with its natural slip. When the brushes are moved so that the injected e.m.f. opposes the current, the speed is reduced (slip positive); for the opposite movement the speed is increased (slip negative).

Performance In a machine with a synchronous speed n_s and a brush-separation electrical angle θ, the no-load speed is

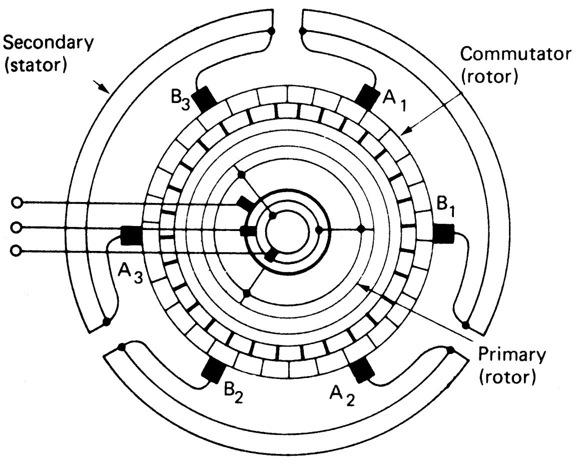


Figure 20.45 Schrage motor; stator and rotor circuits

$$n = n_s(1 - k \sin \tfrac{1}{2}\theta)$$

where k is a constant depending on the numbers of turns in the secondary and commutator windings. A typical relation between n and θ is given in *Figure 20.46*(*a*) for a motor with a 4:1 speed range—about the practical limit. For a given brush position the speed is nearly constant up to 1.5–2 times full-load torque, as shown in *Figure 20.46*(*b*). The speed drop with increasing load is greater than for a plain induction motor because of the brush resistance and the impedance of the tertiary commutator winding. At synchronous speed, when $\theta = 0$ and the secondary is short circuited through the brushes, the overall efficiency is similar to that of an induction motor. At other speeds the efficiency is perhaps 5% lower on account of the tertiary I^2R loss and the stator core loss. The power factor approaches unity at speeds above synchronous as the negative slip results in a capacitive effect. At subsynchronous speeds the power factor falls; however, it can be raised in non-reversing motors by arranging that the brush movement is asymmetric with respect to the 'in-line' position, the axis bisecting a corresponding pair of brushes being progressively displaced in a direction opposite to that of the rotor rotation.

Starting can be effected by direct switching with the brushes set in the lowest-speed position, the starting torque being about 1.5 times full-load value with 1.5–2 times full-load current. Commutation limits the output to about 20 kW/pole and the speed range to 4:1. The maximum output for a motor is thus about 200 kW. With limited speed range, rather higher ratings, e.g. 350 kW and 1.5:1, can be achieved. As the primary winding is supplied through slip-rings the supply voltage is restricted to about 600 V.

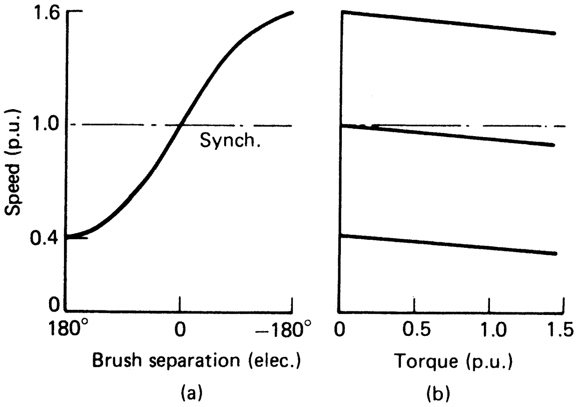
20.3.4.2 Doubly fed motor

The stator resembles that of a conventional induction motor; it is fed from the supply at any desired voltage up to 11 kV. The rotor carries a commutator winding and has six (occasionally three) brushes per pole-pair (see *Figure 20.47*). The rotor voltage is necessarily low (200–300 V) and the brushes are connected to the supply through a variable-ratio transformer. The gap flux travels at synchronous speed, and as a result of the commutating function the e.m.f.s at the brushes are of supply frequency at any speed. A variable e.m.f. can thus be obtained from the transformer and injected into the rotor circuits to give speed control from standstill to 1.5–2 times synchronous speed. In practice the variable-voltage transformer is an induction regulator to give smooth speed control. It must be a double regulator to avoid changing the phase angle of the injected voltage.

Performance At the zero-voltage position of the regulator the brushes are short circuited through the regulator winding and the machine operates as an induction motor, though with a higher rotor effective impedance. Moving the regulator in one or other direction introduces an e.m.f. into the rotor circuit, to give speed–torque relations similar to those of the Schrage motor. The overall power factor tends to be low owing to the magnetisation of the regulator, but it can be improved by special means.

With the regulator in the lowest-speed position the motor can be direct-started to give about 1.5 times full-load torque with 1.5–2 times full-load current. The regulator must carry the slip power, so that speed variation down to zero requires it to be of a physical size comparable to that of the motor. Machines of some thousands of kilowatts can be economically built for speed ranges of ±15% or 20% of synchronous speed.

Unlike the Schrage motor the stator-fed commutator motor is not self-contained, but it can be made in larger ratings and for higher voltages. Again, the simpler brush arrangements makes the machine economic in ratings down to 2 or 3 kW.



(a) (b)

Figure 20.46 Schrage motor: characteristics

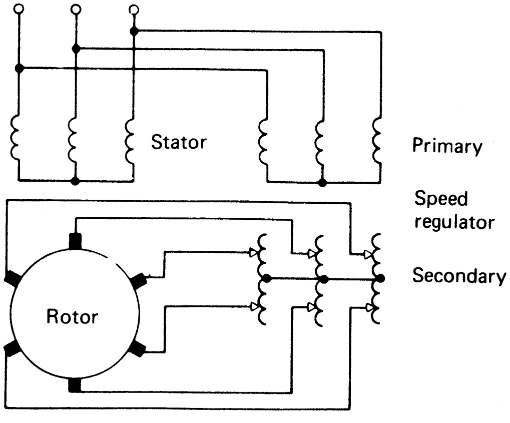


Figure 20.47 Stator-fed three-phase commutator motor

20.3.4.3 Three-phase series motor

It is possible to connect the rotor brushes in series with the stator winding to give a machine with a 'series' speed–torque characteristic and with speed variation by moving the brush position. To limit the rotor voltage, however, a transformer is necessary between stator and rotor. If the transformer is that of the induction-regulator type the brushes can be fixed and the speed adjusted by means of the regulator. The series commutator motor is uncommon, but if a steeper characteristic than that furnished by the stator-fed machine is desirable or acceptable, as for fan drives, there is some economy because of the smaller losses and the simpler regulator.

20.3.5 Synchronous motors

Any synchronous generator will operate as a motor and run at precisely synchronous speed up to its pull-out torque of 2–2.5 times full-load torque. Other significant features of the motor are the controllability of its power factor up to unity or leading values, the necessity of a d.c. excitation circuit and the fact that the motor is not inherently self-starting. In ratings above 300–500 kW, however, the synchronous motor, although more expensive than the induction motor, has a higher efficiency and lower running cost and, therefore, often gives a more economic drive. Except for 3000 rev/min motors the salient-pole construction is generally adopted.

20.3.5.1 Starting

If the motor is always to be started on no load it can be run up to speed by a small pony motor, usually an induction motor, and then allowed to pull into synchronism when the excitation is switched on. If the motor has solid poles and pole-shoes it may be possible to start it by induction-motor action resulting from eddy currents induced in them when the supply is switched on.

Induction start Most synchronous motors are started by use of a cage winding embedded in the pole-faces to give an induction-motor torque when the stator is energised, by direct switching on through an autotransformer. When the speed approaches synchronous the d.c. excitation is applied and the motor synchronises. During starting, high voltages may be induced in the field winding, and it is usual to short-circuit this winding during the start through a resistor which is disconnected after the machine has pulled into step. The current in the field winding adds significantly to the starting torque.

To ensure that the motor closely approaches synchronous speed at the end of an induction start, the resistance of the cage winding should be low; however, for good torque production at low speeds the cage resistance should be high, so some compromise is required.

With direct switching the starting torque is about one-half of full-load torque with 2–3 times full-load current. For machines rated above 200 kW an autotransformer is needed for starting.

20.3.5.2 Excitation

The conventional method of excitation is by a shunt-connected d.c. exciter mounted on the motor shaft, control being effected by variation of the exciter field current; the exciter should, however, be disconnected from the motor field winding during starting on account of the alternating currents that would otherwise be induced in it and which could destroy residual magnetism and prevent the build-up of the excitation.

An a.c. mains-fed rectifier with d.c. output fed to the rotor through the slip-rings could replace the d.c. exciter, but modern practice favours *brushless excitation*, in which an a.c. exciter feeds the rotor field winding through rectifiers, the whole arrangement being incorporated in the rotor. The a.c. exciter field is energised by means of a small permanent-magnet generator to ensure build-up of the excitation under all conditions.

Excitation control can be made automatic, but it is more usually pre-set to give unity or leading power factor at full load. The increased reactive leading power at lighter loads helps to raise the overall system power factor and to improve the transient stability of the motor when it is subject to disturbances.

20.3.5.3 Synchronous-induction motor

Where high starting torques (e.g. 2–2.5 times full-load) are required, the synchronous-induction motor is suitable. It resembles a slip-ring induction motor and is started on resistance when it is up to nearly synchronous speed, d.c. excitation is switched on to the rotor through the slip-rings and the machine synchronises.

The air gap of the synchronous-induction motor is longer than that of the normal induction motor in order to achieve synchronous stability, and the rotor winding resistance is lower in order to ensure pulling into step. An exciter must also, of course, be provided. Another difficulty is the adaptation of the rotor for the dual purpose of starting and excitation; the direct current is normally fed into two of the slip-rings, the third being isolated so that the winding is not all usefully employed during running. Moreover, the relatively few turns and the large current for which the winding is usually designed necessitate an abnormal design for the exciter, i.e. a low-voltage, high-current machine. The starting performance is similar to that of a slip-ring induction motor, but the running performance is better in that the efficiency is 1–2% higher and the power factor may be made unity or leading. The pull-out torque in the synchronous mode is about 1.5 times full-load torque, but if the machine pulls out it can continue to run as an induction motor with a peak torque up to 2.5 times full-load value.

Where the compromise characteristics of the synchronous-induction motor are inadequate, large salient-pole synchronous motors may be built with a slip-ring pole-face winding, so that the starting and synchronous functions may each be optimised.

20.3.6 Reluctance motors

The reluctance motor is a cheap and reliable synchronous motor that requires no d.c. excitation. Commercial motors are available in ratings of 20 kW or more. The machine has a three-phase stator winding similar to that of an induction motor, and a rotor without windings. For a given frame the output is about 60% of that of an induction machine, and the motor has a slightly lower efficiency. It has, however, advantages for drives such as the accurate positioning of nuclear-reactor rods, the operation of rotating stores in computers, and in synchronised multi-motor drives.

The essential feature of the rotor is a strong 'saliency' effect obtained in ways such as those illustrated in *Figure 20.48*. The obvious saliency in (*a*) is, in modern machines, replaced by designs based on studies of flux patterns to give the greatest difference in the reluctance offered respectively in the direct and quadrature axes—a condition of good saliency effect. The rotor iron may be solid but is more usually laminated.

Reluctance torque is maintained only at synchronous speed, so that some form of cage winding must be incorporated for starting. Direct switching is employed, giving starting currents

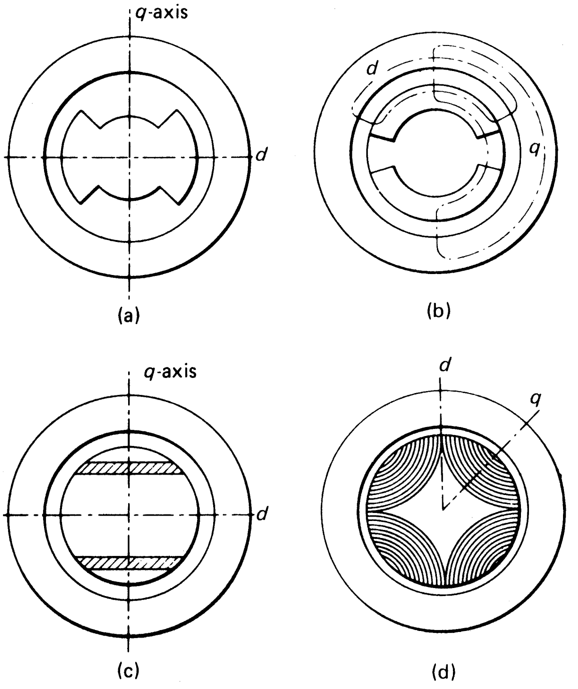


Figure 20.48 Reluctance motors: (a) salient two-pole; (b) segmental two-pole; (c) flux-barrier two-pole; (d) axially laminated four-pole

up to 4–6 times full-load current. The effective rotor resistance has an important influence on the starting and pull-in torques.

The requirements for a satisfactory motor are good synchronous performance (efficiency, power factor and pull-out torque), good pull-in torque (especially for high-inertia loads) and stability. Some of these conflict: increasing the ratio of *d*- and *q*-axis reluctances gives higher output and pull-out torque, but lower pull-in torque and impaired stability. The design is influenced particularly by the load inertia. For a motor of 5 kW rating typical data are: reluctance ratio, 3–6; efficiency, 70–80%; power factor, 0.6–0.75 lagging; pull-out torque, 2–2.5 p.u.; pull-in torque, 0.9–1.2 p.u.

Motors with change-speed windings are possible, and motors can be built for variable frequency (20–200 Hz) but these are liable to instability at the lower end of the range.

20.3.7 Single-phase motors

Single-phase motors are rarely rated above 5 kW. Fractional-kilowatt motors, most of which are one-phase, account for 80–90% of the total number of motors manufactured and for 20–30% of the total commercial value. A typical modern home may have 10 or more one-phase motors in its domestic electrical equipment.

20.3.7.1 Series motor

As the direction of rotation and of torque in a d.c. series motor are independent of the polarity of the supply, such a motor can operate on a.c. provided that all ferromagnetic parts of the magnetic circuit are laminated to minimise core loss.

Universal motor In the fractional-kilowatt sizes the series motor has the advantage, since it is non-synchronous, of being able to run at speeds up to 10 000 rev/min. It is very well adapted to driving suction cleaners, drills, sewing machines and similar small-power rotary devices. Its facility of operating on d.c. and a.c. is not now important, but is the origin of the term 'universal'. The machine has a 'series' speed–torque characteristic, the no-load speed being limited by mechanical losses. The power factor is between 0.7 and 0.9 (mainly the result of armature inductance), but this is of no significance in small ratings. Typical characteristics for a motor for d.c. and 50 Hz supplies of the same nominal voltage are shown in *Figure 20.49*.

In all a.c. commutator motors the commutation conditions are more onerous than on d.c. because the coils undergoing commutation link the main alternating flux and have e.m.f.s induced of supply frequency. The e.m.f.s are offered a short-circuited path through the brushes and contribute to sparking at the commutator. As the e.m.f.s are proportional to the main flux, the frequency and the number of turns per armature coil, these must be limited; a further limit on the current in a short-circuited coil is provided by high-resistance carbon brushes.

Compensated motor Series a.c. commutator motors up to 700–800 kW rating are used in several European railway traction systems. For satisfactory commutation the frequency must be low, usually 16$\frac{2}{3}$ Hz, and the voltage must also be low (400–500 V), this being provided by a transformer mounted on the locomotive. The inductance of the armature winding is necessarily rather high, so that a compensating winding must be fitted to neutralise armature reaction in order to ensure a reasonable power factor.

Motors of this type have been built, of limited output, for operation on modern 50 Hz traction systems but have now been superseded by rectifier- or thyristor-fed d.c. motors.

20.3.7.2 Repulsion motor

The repulsion motor is a form of series motor, with the rotor energised inductively instead of conductively. The commutator rotor winding is designed for a low working voltage. The brushes are joined by a short circuit and the brush axis is displaced from the axis of the one-phase stator winding

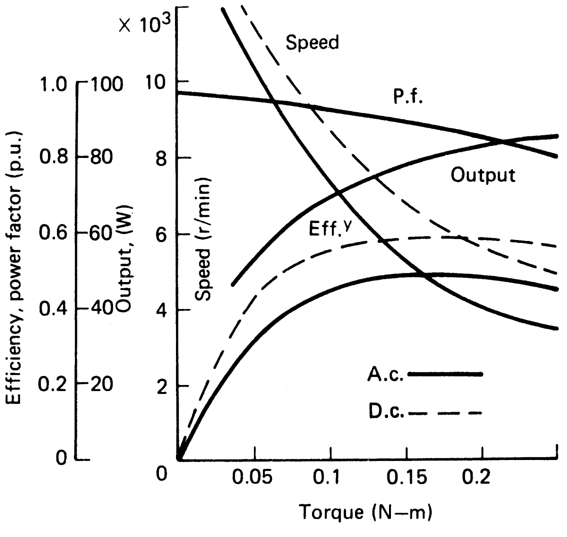


Figure 20.49 Characteristics of a 75 W universal motor

(*Figure 20.50*). With non-reversing motors (*Figure 20.50(a)*) a single stator winding suffices; however, for reversing motors the stator has an additional winding, connected in one or other sense in series with the first winding to secure the required angle between the rotor and effective stator axes for the two directions of rotation, as in *Figure 20.50(b)*.

A stator winding of N_1 turns as in (*a*) can be resolved into two component windings respectively coaxial with and in quadrature with the axis of the rotor winding, and having respectively turns N_1 sin α and N_1 cos α. Windings (*b*) give the two axis windings directly, although here the turns can be designed for optimum effect. The coaxial winding induces e.m.f.s and currents in the rotor, and these currents lying in the field of the other stator winding develop torque; since both stator and rotor currents are related, the motor has a 'series' characteristic.

When the motor is running, the direct and quadrature axis fluxes have a phase displacement approaching 90°, so producing a travelling-wave field of elliptical form which becomes nearly a uniform synchronously rotating field at speeds near the synchronous. Near synchronous speed, therefore, the rotor core losses are small and the commutation conditions are good.

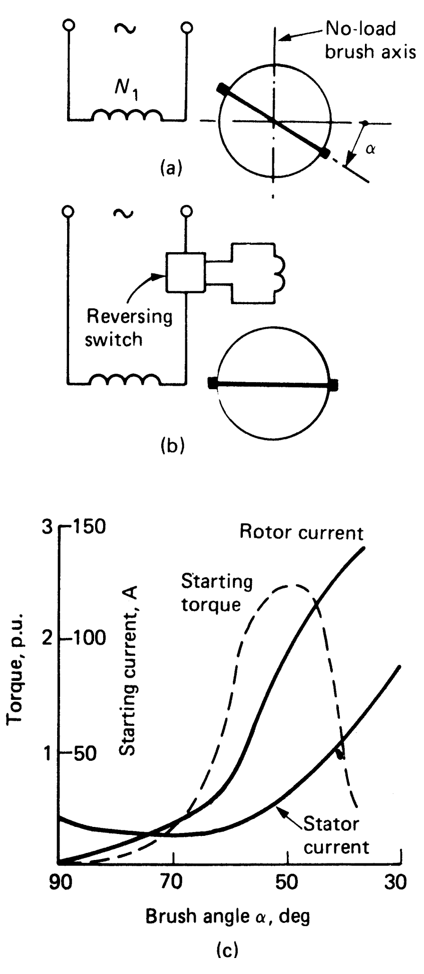


Figure 20.50 Repulsion motor: alternative forms and starting characteristics

Small motors can readily be direct-switched for starting, with 2.5–3 times full-load current and 3–4 times full-load torque. The normal full-load operating speed is chosen near, or slightly below, synchronous speed in order to avoid excessive sparking at light load. Repulsion motors are used where a high starting torque is required and where a three-phase supply is not available. For small lifts, hoists and compressors their rating rarely exceeds about 5 kW.

20.3.7.3 Induction motors

The one-phase induction motor is occasionally built for outputs up to 5 kW, but is normally made in ratings between 0.1 and 0.5 kW for domestic refrigerators, fans and small machine tools where a substantially constant speed is called for. The behavior of the motor may be studied by the rotating-field or the cross-field theory. The former is simpler and gives a clearer physcial concept.

Rotating-field theory The pulsating m.m.f. of the stator winding is resolved into two 'rotating' m.m.f.s of constant and equal magnitude revolving in opposite directions. These m.m.f.s are assumed to set up corresponding gap fluxes which, with the rotor at rest, are of equal magnitude and each equal to one-half the peak pulsating flux. When the machine is running, the forward field component *f*, i.e. that moving in the same direction as the rotor, behaves as does the field of a polyphase machine and gives the component torque–speed curve marked 'forward' in *Figure 20.51*; the backward component *b* gives the other torque component, and the net torque is the algebraic sum. At zero speed the component torques cancel so that the motor has no inherent starting torque, but if it is given a start in either direction a small torque in the same direction results and the machine runs up to near synchronous speed provided that the load torque can be overcome.

The component torques in *Figure 20.51* are, in fact, modified by the rotor current. Compared with the three-phase induction motor, the one-phase version has a torque falling to zero at a speed slightly below synchronous, and the slip tends to be greater. There is also a core loss in the rotor produced by the backward field, reducing the efficiency. Moreover, there is a double-frequency torque pulsation generated by the backward field that can give rise to noise. The efficiency lies between about 40% for a 60 W motor and about 70% for a 750 W motor, the corresponding power factors being 0.45 and 0.65, approximately.

The equivalent circuit of *Figure 20.52* is based on the rotating-field theory, using parameters generally similar to

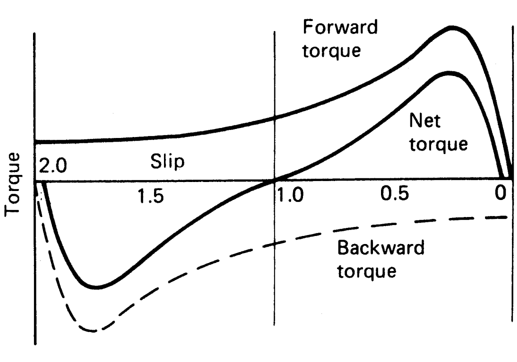


Figure 20.51 Torque components in a single one-phase induction motor

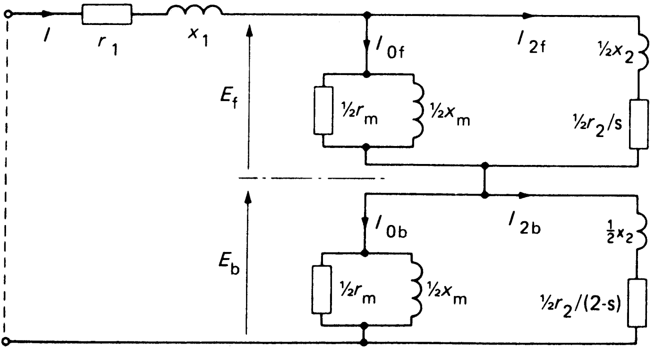


Figure 20.52 Simple one-phase induction motor: equivalent circuit

those for the three-phase machine. The e.m.f.s E_f and E_b are generated respectively by the forward and backward field components and are proportional thereto. The respective component torques are proportional to $I_{2f}^2 r_2/2s$ and $I_{2b}^2 r^2/[2(2-s)]$, the next torque being their difference.

Starting To start a one-phase induction motor, means are provided to develop initially some form of travelling-wave field. The arrangements commonly adopted give rise to the terms 'shaded-pole' and 'split-phase'.

Shaded-pole motor The stator has salient poles, with about one-third of each pole-shoe embraced by a shading coil. That flux which passes through the shading coil is delayed with respect to the flux in the main part of the pole, so that a crude shifting flux results. The starting torque is limited, the efficiency is low (as there is a loss in the shading coil), the power factor is 0.5–0.6 and the pull-out torque is only 1–1.5 times full-load torque. Applications include small fans of output not greatly exceeding 100 W.

Resistance split-phase motor The additional flux is provided by an auxiliary starting winding arranged spatially at 90° (electrical) to the main (running) winding. If the respective winding currents are I_m and I_s with a relative phase angle α, the torque is approximately proportional to $I_m I_s \sin \alpha$. At starting, the main-winding current lags the applied voltage by 70–80°. The starting winding, connected in parallel with the main winding, is designed with a high resistance or has a resistor in series so that I_s lags by 30–40°. The effect of this resistance on the starting characteristic is shown in *Figure 20.53*(*a*). With given numbers of turns per winding and a given main-winding resistance, then for a specified supply voltage and frequency there is a particular value of starting-winding resistance for maximum starting torque. The relation can be obtained from the phasor diagram. *Figure 20.53*(*b*), in which V_1 is the supply voltage and I_m at phase angle ϕ_m is the main-winding current. The locus of the starting-current phase I_s with change in resistance is the semicircle of diameter OD (which corresponds to zero resistance). The torque is proportional to $I_m I_s \sin(\phi_m - \phi_s)$ and is a maximum for the greatest length of the line AC. From the geometry of the diagram it can be shown that for this condition $\phi_s = \frac{1}{2}\phi_m$.

Direct switching is usual. To reduce loss, the auxiliary winding is open-circuited as soon as the motor reaches running speed. The starting torque for small motors up to 250 W is 1.5–2 times full-load torque, and that for larger motors rather less, in each case with 4–6 times full-load current. The operating efficiency is 55–65% and the power factor 0.6–0.7.

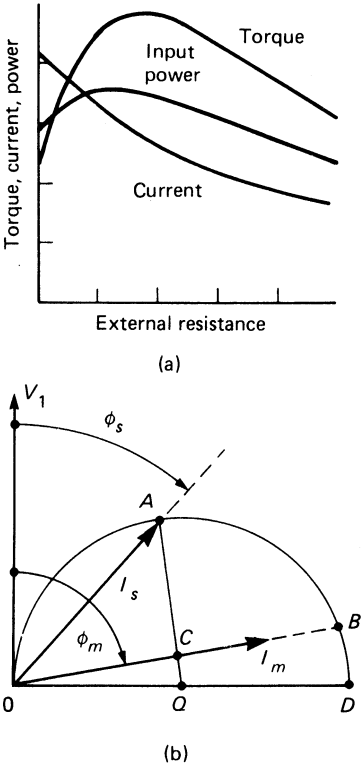


Figure 20.53 Single-phase induction motor: split-phase resistance start

Capacitor split-phase motor A greater phase difference $(\phi_m - \phi_s)$ can be obtained if a series capacitor is substituted for the series resistor of the auxiliary winding. Maximum torque occurs for a capacitance such that the auxiliary current leads the main current by $(\frac{1}{2}\pi - \alpha)/2$. The capacitor size is from 20–30 μF for a 100 W motor to 60–100 μF for a 750 W motor. For economic reasons the capacitor is as small as is consistent with producing adequate starting torque, and some manufacturers quote alternative sizes for various levels of starting torque.

If the capacitor is left in circuit continuously (*capacitor-run*) the power factor is improved and the motor runs with less noise. Ideally, however, the value of capacitance for running should be about one-third of that for the best starting. If a single capacitor is used for both starting and running, the starting torque is 0.5–1 times full-load value and the power factor in running is near unity.

Repulsion-induction motor Machines have been designed to combine the high starting torque capability of the repulsion motor with the constant-speed running characteristic of the induction motor.

Repulsion-start motor This motor has a stator winding like that of a repulsion motor and a lap commutator winding, with the addition of a device to short circuit the commutator sectors together by centrifugal action when the speed reaches about 75% of normal. The device may also release the brushes immediately thereafter. Thus the commutator rotor winding becomes, in effect, a short-circuited 'induction'-type winding for running. Small motors direct-switched give 3–4 times full-load torque with about three times full-load current. A lower starting current is obtained by connecting a graded resistor in series with the stator winding.

Repulsion-induction motor The machine has a repulsion-type stator winding but the change from the repulsion-mode to the induction-mode operation is gradual as the machine runs up to speed. The rotor has two windings in slots resembling those of a double-cage induction motor. The outer slots carry a commutator winding with brushgear, the inner slots contain a low-resistance cage with cast aluminium bars and end-rings, and its deep setting endows it with a high inductance. During acceleration the reactance of the cage falls and its torque increases, tending to counterbalance the falling torque of the commutator winding. At speeds above synchronous the cage torque reverses, giving a braking action which holds the no-load speed to a value only slightly above synchronous speed. The commutation is better than that of a plain repulsion motor, and the motor is characterised by a good full-load power factor (e.g. 0.85–0.9 lagging). With direct switching the starting torque is 2.5–3 times and the current 3–3.5 times full-load value.

20.3.8 Motor ratings and dimensions

Motors of small and medium rating are built to the standards of IEC 72, which list a coherent range of main structural dimensions with centre heights between 56 and 1000 mm. BS 3939 gives the standard ratings for the UK. *Table 20.5* lists data for rotating machines up to 1 MW. Approximate values of rated current for three-phase, one-phase and d.c. machines are set out in *Table 20.6* for machines up to 150 kW and for a range of supply voltages. Voltages for d.c. machines correspond to nominal values obtained from rectified a.c. supplies.

Table 20.5 Rotating machines: recommended ratings and shaft heights

Rating (kW)
0.06, 0.09, 0.12, 0.18, 0.25, 0.37, 0.55, 0.75, 1.1, 1.5, 2.2, 3.7, 5.5, 7.5, 11, 15, 18.5, 22, 30, 37, 45, 55, 75, 90, 110, 132, 150, 160, 185, 200, 220, 250, 280, 300, 315, 335, 355, 375, 400, 425, 450, 475, 500, 530, 560, 600, 630, 670, 710, 750, 800, 850, 900, 950, 1000
Shaft-centre height (mm)
56, 63, 71, 80, 90, 100, 112, 132, 160, 180, 200, 225, 250, 280, 315, 355, 400, 450, 500, 630, 710, 800, 900, 1000

20.3.9 Testing

Tests on machines after manufacture or after erection on site are made in accordance with standard Specifications. They cover (i) insulation resistances, (ii) winding resistances, (iii) temperature rise, and (iv) losses. Further tests on particular types of machine (e.g. commutation, starting) are required to meet customers' requirements or to obtain design data. Where a batch of similar machines is concerned, a 'type test' on one for detailed performance is usually acceptable.

20.3.9.1 Insulation

'Megger' testing of the insulation resistance between windings, and from windings to frame, must be performed before any live connections are made. The insulation resistance (in mega-ohms) should not be less than 1, or less than $V/(1+S)$ for a machine of rating S (in kilovolt-amperes), where V is the rated voltage. It may be necessary, if the insulation resistance is low, to 'dry out' the machine. The winding continuity having been checked, an insulation test should immediately precede a h.v. test and also be made prior to energising the machine for the first time.

20.3.9.2 Resistance

Measured resistances of the windings check the design figures and are required for calculating losses: winding temperatures must be noted. An ammeter/voltmeter method is usual; alternatively a bridge method may be used if there are no brush contacts in the circuit. For a commutator winding, the volt drop is taken between the commutator sectors under the brushes, the brush drop being taken separately.

20.3.9.3 Temperature rise

The permissible temperature rise of a machine on rated load depends on the insulation class. Temperatures are measured at a number of points (particularly at or near likely 'hot spots') and, where possible, during a heat run with the machine operating at rated load, until a steady temperature has been reached. The rated load may be a *maximum continuous rating*, or a *short-time rating*, or some special rating based on a duty cycle. The duration of the heat run varies from 2 h for small machines to 8 h or more for large ones. Temperature readings are taken (where feasible) every 15 or 30 min. The following three methods of temperature measurement are in use.

Thermometer Mercury or alcohol thermometers may be used, the latter being preferable especially on large machines, as eddy currents in the mercury caused by stray fluxes may cause high readings; also mercury from a broken thermometer can cause damage to certain alloys. Good contact must be made between the thermometer bulb and the surface concerned, and the bulb should be well covered by a non-heat-conducting material such as felt or putty.

The thermometers should be located on the stator core and windings and read at intervals throughout the run, and then affixed to the rotor core and windings and to the commutator, if any, as quickly as possible after shut-down. They should be placed where temperatures are likely to be highest; however, only the surface temperature is measured, and not the true hot-spot temperature.

Resistance The resistance–temperature coefficient of the winding material can give an average winding temperature.

Table 20.6 Approximate rated motor currents (A)

	Induction motors					*D.c. motors*			
	Three-phase			*One-phase*					
Rating (kW)	*350 V*	*415 V*	*500 V*	*240 V*	*415 V*	*170 V*	*290 V*	*440 V*	*570 V*
0.06	—	—	—	0.7	0.4	0.6	0.4	0.3	0.2
0.09	—	—	—	0.9	0.5	0.9	0.5	0.4	0.3
0.12	—	—	—	1.2	0.7	1.1	0.6	0.5	0.4
0.18	—	—	—	1.5	0.9	1.6	0.9	0.6	0.5
0.25	—	—	—	1.9	1.1	2.0	1.2	0.8	0.6
0.37	1.3	1.1	0.9	2.8	1.6	2.8	1.7	1.1	0.9
0.55	1.8	1.6	1.3	4.0	2.3	4.1	2.4	1.6	1.2
0.75	2.4	2.0	1.7	5.0	2.9	5.3	3.0	2.0	1.6
1.1	3.4	2.9	2.4	7.2	4.1	7.5	4.4	3.0	2.3
1.5	4.5	3.7	3.1	9.3	5.4	10	6.0	4.0	3.1
2.2	6.3	5.3	4.4	13	7.8	15	8.7	5.8	4.4
3.7	10	8.4	7.0	22	13	24	14	9.6	7.4
5.5	13	12	10	32	18	36	21	14	11
7.5	18	16	13	43	25	49	29	19	15
11	28	23	19	62	36	—	42	28	21
15	35	30	25	82	48	—	57	38	29
18.5	45	38	31	100	58	—	70	47	36
22	53	44	37	118	68	—	83	56	43
30	68	56	48	—	—	—	114	76	58
37	85	72	60	—	—	—	138	92	71
45	100	85	71	—	—	—	169	112	86
55	125	105	88	—	—	—	206	137	106
75	165	137	113	—	—	—	280	187	144
90	185	150	125	—	—	—	337	225	173
110	210	175	145	—	—	—	408	270	210
130	245	215	180	—	—	—	480	320	247
150	320	250	210	—	—	—	555	370	285

For copper, the temperature θ_2 corresponding to a measured resistance R_2 is related to the resistance R_1 at θ_1 by

$$\theta_2 = (R_2/R_1)(\theta_1 - 235) - 235$$

Embedded detector Thermocouples or resistance thermometers can be embedded in the core and windings during manufacture, a site between the coil-sides of a double-layer winding being common. At least six detectors, suitably distributed, should be installed. When well sited, detectors can give a closer estimate of hot-spot temperatures than other methods.

Temperature limits The three techniques above do not measure the same quantities, nor do they measure actual hot-spot temperatures. The values must therefore be in most cases significantly lower than the limits appropriate to the insulation class. Typical values for various locations as laid down in BS 2613 are given in *Table 20.7*.

20.3.9.4 Losses and efficiency

Efficiency may be determined by direct output/input ratio, by the total loss and either input or output, or by loss summation. As the efficiency of a large machine may have to be guaranteed within 0.01%, accurate determination is essential.

Output/input Electric power is readily measured, but mechanical power measurement requires some form of dynamometer, often of limited accuracy. Test rigs with instrumentation are used for small motors, but for large machines an adequate estimation of efficiency by this method may be impossible.

Back-to-back If two similar machines are available, one as a motor can drive the other as a generator. The net input is then the total loss, which can be accurately measured. This method can be applied for heat runs with comparative economy.

Loss summation A separate determination of each separable loss is made; the items are then summed to give the total loss, and thence the efficiency. The losses to be determined are:

core,
stator I^2R,
rotor I^2R,
load (stray),
brush-contact,
excitation, and
friction and windage.

Apart from the stray loss, each of these can be determined without loading the machine (although it must be run at normal speed). For uniformity the current losses are

Table 20.7 Temperature limits (°C)

Part of machine	*Method**	*Insulation class*				
		A	*E*	*B*	*F*	*H*
A.c. windings						
5000 kVA or core length over 1 m	T, D	60	70	80	100	125
5000 kVA	T	60	75	80	100	125
Commutator windings	R	50	65	70	85	105
	T	60	75	80	100	125
Field windings						
Low-resistance	R, T	60	75	80	100	125
Other windings	R	50	65	70	85	105
	T	60	75	80	100	125
Single-layer windings with bare or varnished metal	R, T	65	80	90	110	135
Permanently short-circuited windings	T	60	75	80	100	125
Iron core in contact with insulated windings	T	60	75	80	100	125
Commutators and slip-rings	T	60	70	80	90	100

Temperature measurement: T, thermometer; R, resistance, D, embedded detector.

found from the measured resistances referred to a standard temperature of 75 °C.

Stray losses These are, in practice, largely proportional to (current)^2 and, although insignificant in small machines up to a few kilowatts, they become very important in large machines. It is possible to estimate the stray loss from certain tests, as described later for particular machines; in other cases, however, they must be estimated from experience.

20.3.9.5 H.v. tests

The final test carried out before shipping a machine is the h.v. test in which a specified voltage at a frequency between 25 and 100 Hz is applied for 1 min between windings and earth and between windings. The specified voltage is usually (twice rated voltage + 1000) volts, although certain exceptions to this are given in BS 2613.

The purpose of the test is to ensure that the insulation has a sufficient factor of safety to guard against fortuitous voltage transients which may occur in practice. The test should, however, only be carried out once as repeated applications may damage the insulation. It may be desirable in some cases to repeat the test after the machine has been assembled on site, in which case a voltage of not more than 80% of the original test voltage should be applied.

20.3.9.6 D.c. motors

The following tests are related specifically to d.c. machines.

Armature volt-drop test This is carried out on the armature winding before assembly or if it has developed a fault. Current is fed into the armature by clamped connectors and the voltage between sectors is measured around the commutator. A placing of the connectors a *pole-pitch* apart is suitable for most industrial two- or four-pole wave-connected armatures: the direction of the volt drop changes at each pole-pitch. A *diametral* connection is suitable for small lap-connected rotors and for six-pole wave-connected rotors: the direction of the drop reverses at the lead-in points. With the *bar-to-bar* connection the current is led into adjacent sectors and repeated all round the commutator, and all measured drops should be the same. An alternative test for wave windings is to lead the current into the commutator at sectors separated by the *commutator pitch* of the winding. This is repeated all round the commutator and all readings of volt drop should be the same: this checks each individual coil.

Neutral setting Adjustment of the brush rocker so that the brushes are in the correct neutral position can most conveniently be done by applying about half normal voltage to the field winding with a low-reading voltmeter connected between the positive and negative brushes. The position of the rocker should be adjusted until there is no kick on the voltmeter when making or breaking the field circuit. It is desirable to remove all brushes except the two being used, and these should be bevelled so that they do not cover more than one segment.

If the exact position of a coil on the armature can be observed, the armature can be moved until this coil is symmetrically placed with its centre line opposite the centre line of a pole; the brush rocker should then be moved until a set of brushes stands on the commutator segment connected to this coil and it will then be in the neutral position. Occasionally one of the armature coils is specially marked by the manufacturers to facilitate this method of setting the neutral.

No-load test The power input to the motor running on no load with normal field current and at normal speed gives the sum of the core, friction and windage losses, together with a small armature I^2R loss which can be calculated and deducted. A *shunt* motor is run at normal field current, the speed being adjusted to normal full-load value by varying the applied armature voltage. The core and mechanical loss so determined is the *fixed* or *constant* loss, so called because it changes only slightly with load. A *series* motor is run with separate excitation to the field, and an armature voltage that will give (for a specified field current) the speed corresponding to load conditions at rated voltage. The voltage to be applied is equal to the counter-e.m.f. under load conditions, i.e. rated voltage less the volt drop in field and armature resistances.

The mechanical loss alone can be estimated by running the motor on no load in normal connection but with a low applied voltage that is adjusted to give the speed required. Then the power input corrected for I^2R loss is substantially the mechanical loss alone.

Loss summation The core, friction and windage losses from the no-load test are added to calculated armature and field I^2R losses (corrected to 75 °C) and to the brush loss taken generally as that due to a total brush volt drop of 2 V. The stray loss is allowed for by a deduction of 1% from the efficiency calculated from the losses mentioned above. If the machine has a compensating winding to counteract the distorting effect of armature reaction, the deduction is 0.5%.

Back-to-back

Shunt motors *Figure 20.54* shows the connections for a back-to-back test for similar machines. The motor M drives the generator G, the excitation of which is adjusted until there is no voltage across switch S; this is then closed. Raising the excitation of G causes it to generate and supply most of the power required for driving M. The two machines operate with both field and armature currents slightly differing in magnitude, but except for small machines the differences are minor.

Series motors Back-to-back tests are more common for series motors, which are not easily loaded with safety. Many series motors are, however, operated in pairs as in traction. The four most common methods are set out in *Figure 20.55*. Method (*a*) is not strictly a back-to-back test as the generator output is dissipated in resistance. Putting the generator field in series with the motor ensures that the generator will always excite. In method (*b*) a variable-voltage booster B is included in the circuit of G to raise its voltage and enable power to be returned to the supply. In both (*a*) and (*b*) it is desirable to boost the supply so that the voltage applied to M can be held at a correct value. Method (*c*) requires an auxiliary drive motor to supply the core and mechanical losses, and a low-voltage booster to supply the I^2R loss. Method (*d*) requires a separate source of excitation, although for small machines the field winding can be connected in series with G and controlled by a diverter.

Commutation The compole setting can be checked by measuring the *volt drop* between the brush and the commutator at three points along the brush width with the machine running on load (*Figure 20.56*(*a*)). The drop should be approximately the same at all points, as in *Figure 20.56*(*b*); if it is greater at the trailing edge, commutation is being delayed and the compoles are too weak. If it is higher at the leading edge the opposite is the case.

The *black-band* test is an alternative method. The compole current is varied independently of the armature current, and at each constant armature current the compole excitation is raised and lowered until sparking occurs. In the zone between these limits (*Figure 20.56*(*c*)), commutation is 'black', i.e. spark-free. The black band should be symmetrical about the axis of armature current.

20.3.9.7 Induction motors

The following are related specifically to three-phase induction motors. Most of the necessary data are obtained from the no-load and short-circuit tests.

No load The motor is run on no load at rated voltage and frequency, and the current and power input are measured. The input power supplies core and mechanical losses plus the stator I^2R loss; the rotor I^2R loss on no load can be neglected. If required, the motor may be operated at voltages above and below normal (*Figure 20.57*) and the power input and current plotted. Extrapolation of the power curve to zero voltage gives the friction and windage loss.

Short circuit (locked rotor) The short-circuit current and power are measured by holding the rotor stationary and applying a low voltage to the stator, it being usual to adjust the voltage to obtain full-load current. In the case of a cage motor the starting torque may also be measured. If the motor is connected for star–delta starting, the actual starting current and torque may be measured direct. The short-circuit current at full voltage is calculated by assuming the current to be proportional to voltage, although on some machines the short-circuit current is actually higher because the leakage reactance is reduced by saturation. This is more noticeable on two- and four-pole motors designed for a high flux density. The starting torque is approximately proportional to (phase voltage)^2.

The short-circuit current is 4–8 times full-load current, depending on the speed and type of motor. A multipolar motor works on a lower flux density to allow a reasonably good power factor. To this end some overload capacity may be sacrificed and the short-circuit current is therefore low.

Parameters The parameters of the equivalent circuit (*Figure 20.37*) can be evaluated from test results. The no-load test gives r_m and x_m if the stator leakage impedance drop is neglected (which is usually justifiable). Neglecting r_m and x_m in the short-circuit test gives the total motor effective resistance

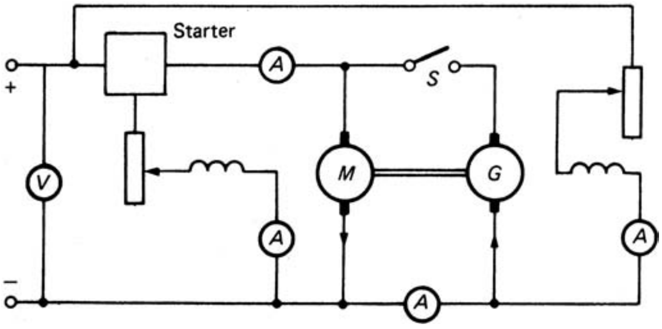


Figure 20.54 D.c. shunt motors: back-to-back test

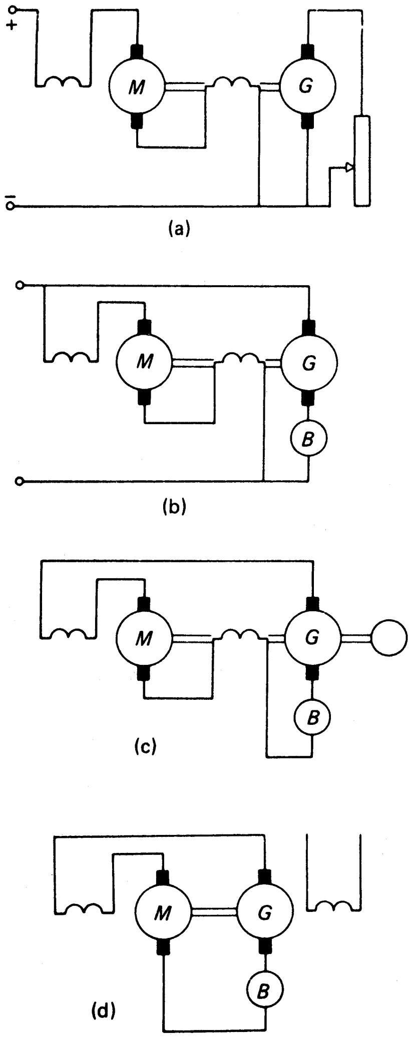

+
−
M
G
(a)
M
G
B
(b)
M
G
B
(c)
M
G
B
(d)

Figure 20.55 D.c. series motors: back-to-back tests

and leakage reactance, r_1+r_2 and x_1+x_2. The stator resistance r_1 can be measured directly and r_2 found; however, it is not possible to separate the two leakage reactances and it is usual to assume that they are equal.

Stray loss Stray loss is included in the short-circuit power. If, as in a slip-ring machine, both r_1 and r_2 (actual and referred) are measurable, the 'true' I^2R loss can be calculated and the stray loss found by subtraction. Where efficiency is calculated by loss summation, a deduction from the calculated efficiency is made (e.g. 0.0625 p.u. at rated load).

Back-to-back For large machines this test is essential. There must be provision for accommodating the speed difference resulting from the positive and negative slips of the motor and generator machines. Methods available are: (i) a close-ratio gearbox where at least one of the machines has a slip-ring rotor for slip control; (ii) a fluid coupling where both are cage machines; and (iii) an adaptation of the automobile differential.

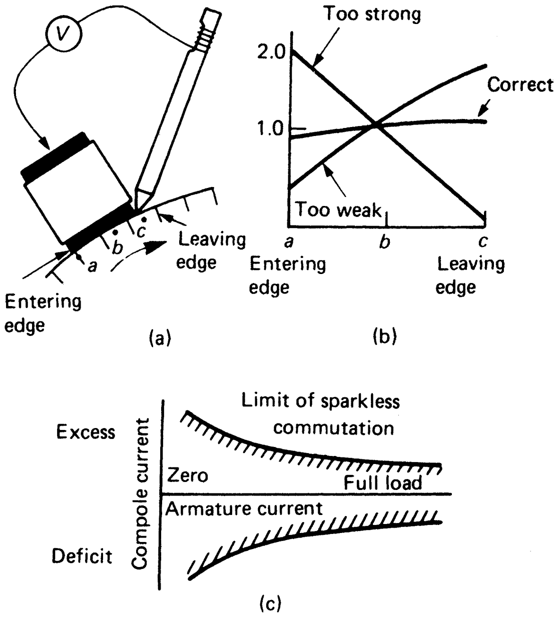

V
a
b
c
Leaving edge
Entering edge
(a)
Too strong
2.0
1.0
Correct
Too weak
a
b
c
Entering edge
Leaving edge
(b)
Limit of sparkless commutation
Excess
Compole current
Zero
Full load
Armature current
Deficit
(c)

Figure 20.56 Commutation tests

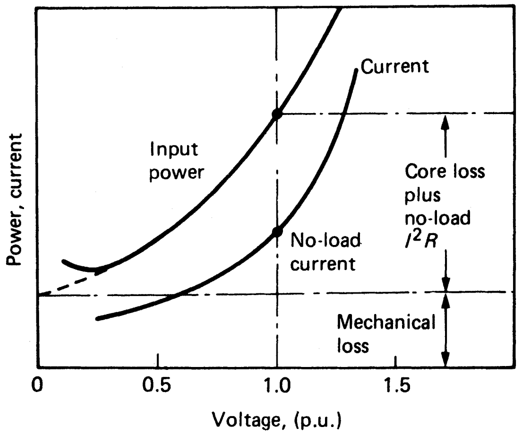


Figure 20.57 Induction motor: no-load characteristics

In (iii), the test machines are coupled to the 'road-wheel' shafts. The torque shaft, driven slowly by a geared auxiliary motor, depresses the speed of the motoring machine and raises that of the generator. The main supply of rated voltage and frequency to the two stators provides the mechanical, core and stator I^2R loss and magnetising currents. The torque shaft controls the drive-power exchange between the two test machines.

20.3.9.8 Synchronous motors

Relevant procedures are given below. Again use is made of the open- and short-circuit tests, for which the machine is driven at rated speed by an auxiliary motor, preferably calibrated.

Open circuit The stator is on open circuit and the field current is varied. The stator e.m.f. gives the *magnetisation characteristic* or open-circuit characteristic (o.c.c.) (*Figure 20.58*). The power input to the test machine comprises core and mechanical losses, which can be separated.

Short circuit The stator windings are short circuited. The field current is increased to give stator current up to slightly above full-load value. The power input to the test machine comprises the stator I^2R, mechanical and stray losses, the core loss usually being negligible. An approximation to the stray loss is obtained by subtracting from the input power to the calculated I^2R loss and the mechanical loss (from the o.c.c. test).

Excitation The excitation for given load conditions is derived as for a generator, with reversed current and load angle. The appropriate phasor diagrams are given in *Figure 20.58*. Typical V-curves, relating the stator and field currents for specified output powers, are plotted in *Figure 20.59.*

Loss summation The friction and windage and the core losses can be obtained from the open-circuit tests, and the stray loss from the short-circuit test, as already described. For any given load the field current can be determined and the field power obtained therefrom. To the summation must be added the loss in the exciter or excitation circuit, and the efficiency can then be calculated.

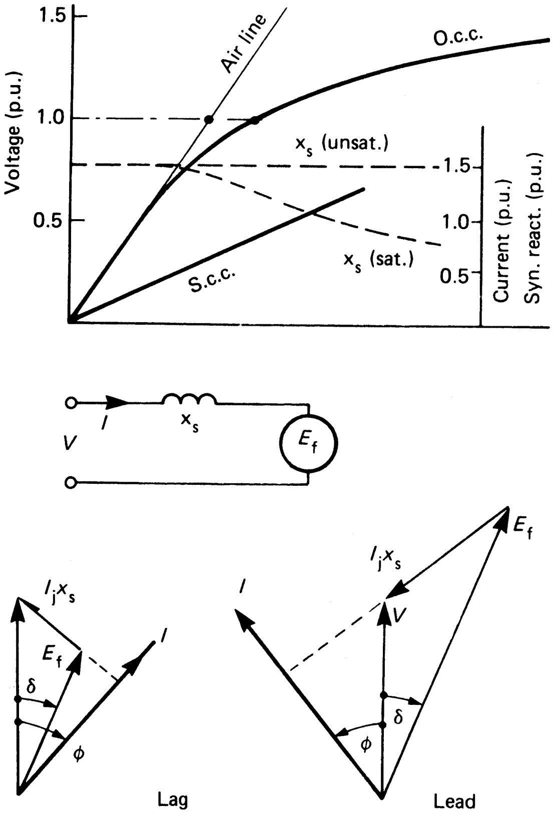


Figure 20.58 Synchronous motor: open- and short-circuit characteristics, equivalent circuit and phasor diagrams

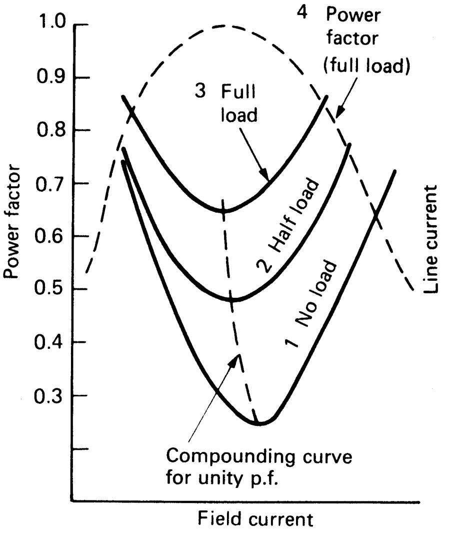


Figure 20.59 Synchronous motor: V-curves

Back to back If two similar machines are mechanically coupled so that their e.m.f.s are in phase, and they are driven mechanically at normal speed then, if they are connected in parallel electrically, a reactive current can be made to circulate between them by suitable adjustment of the field currents or by a booster transformer fed from an external source. The total electrical and mechanical input represents the total loss, but as the machines are not operating at their rated power factors the results are not valid for calculating full-load efficiencies. The test is, however, convenient for carrying out heat runs.

With the machines coupled so that their e.m.f.s are phase-displaced by an angle equal to twice the full-load load-angle, then full-load active and reactive powers can be circulated under conditions closely simulating normal load operation.

Zero power factor A pseudo full-load test can be carried out by loading the machine on inductors so that it operates at normal voltage and current but at a power factor near zero. The air-gap flux will be 4–5% above its normal full-load value and the field current will be up by 20–30%. If the machine is used for a heat run, temperatures will be higher than those that would obtain at rated load; however, appropriate adjustments can be made to simulate rated conditions more closely. Results from the zero-power-factor test can be used as described in BS 4296 to predetermine the field current on load.

20.3.10 Linear motors

Linear machines have translational instead of rotary motion. They can be applied to the drive of a conveyor belt, of a traversing crane on a limited track, of liquid metal in the heat-exchanger of a nuclear reactor plant, or of trains on a high-speed railway system. Short-stroke linear machines are suitable for powerful thrusting action.

20.3.10.1 *Forms*

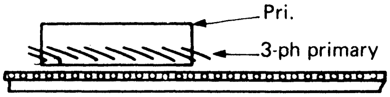
A linear machine can be regarded initially as resembling a normal rotating machine that has been cut and opened out flat. Of the two elements derived respectively from the stator and rotor, either may move. The member connected to the supply is called the *primary*, the other the *secondary*. In use, either member is fixed as the *stator*, while the other becomes the movable *runner*.

Two forms of linear machine (*Figure 20.60*) are the *flat* machine (geometrically an 'unrolled' rotary structure) and the *tubular* machine, which is equivalent to a flat machine 'rerolled' around the longitudinal axis. For electromagnetic force to be developed, it is necessary to ensure that interaction between the working flux and the working currents should be achieved: the directions of the two components in and around the air gap must be at right angles. In the flat machine (*Figure 20.60*(*a*)) the secondary interaction currents are arranged to flow across the element from front to back, with suitable end-connectors to complete the secondary circuits. In the tubular machine the flux enters the pencil-shaped secondary in a radial direction, so that the secondary interaction currents must flow circumferentially.

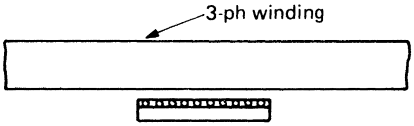
In both of the shapes in *Figure 20.60* there are certain practical difficulties: (i) how to deal with a linear movement which, if continued, must eventually cause a runner of finite length to part company with its stator; (ii) how to support a very strong magnetic pull tending to cause runner and stator to adhere; (iii) how to supply electrical energy to a linear-moving runner. In the flat machine, and in the tubular machine if it is not precisely centred, the otherwise unbalanced attraction force must be sustained by some suitable mechanical arrangement. The linear machine is inherently a device lacking some of the symmetry and balance of the normal rotary form.

20.3.10.2 D.c. and a.c. machines

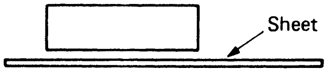
Any arrangement possible in a rotary machine can be realised in the flat form. The outlines in *Figure 20.60* show that the most convenient type is likely to be that corresponding to the cage induction motor, for then the stator can often be made the primary because of the convenience of its supply, while the moving secondary member will correspond to the cage rotor and will require a very simple winding with no supply connections. However, this arrangement is by no means the rule, and much depends on the particular conditions of a given application.



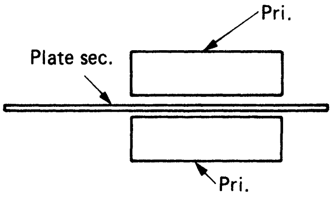
(a) Short primary



(b) Short secondary



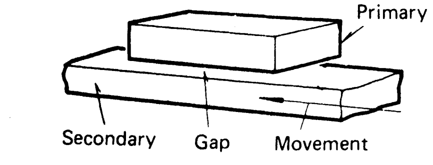
(c) Sheet secondary



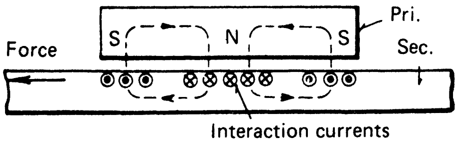
(d) Double primary

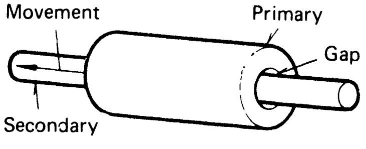
Figure 20.61 Polyphase linear-induction motors

The d.c. and induction forms are the most common. Because of the essential secondary current supply, the d.c. linear motor is usually found as an electromagnetic pump for liquid metal. There is more freedom of shape in the induction machine, as indicated in *Figure 20.61*. The short-primary arrangement (*a*) suits cases in which the total distance to be travelled is great, for it would be uneconomic to wind a long three-phase primary, with only that part in the neighbourhood of the secondary being effective at any one time. The short-secondary form (*b*) is useful if the total excursion is limited and the moving secondary must be comparatively light. In both (*a*) and (*b*) the secondary is shown as a flattened 'cage'. A conducting sheet or plate (*c*) is often as effective, and it obviates magnetic attraction. The now



(a) Flat





(b) Tubular

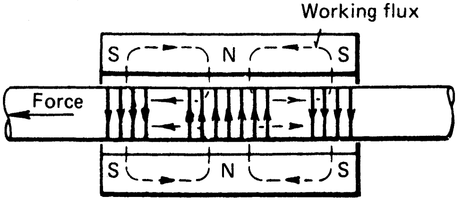


Figure 20.60 Linear motors: (a) flat; (b) tubular

indefinite 'gap', which increases the non-useful leakage flux, can be shortened by the use of a double primary (*d*).

Long-established methods for the design of orthodox rotary machines are not applicable to the radically changed geometry of linear flat or tubular machines. Lacking cylindrical symmetry, the magnetic flux is heavily distorted near the ends of the short element, and as the primary and secondary move relatively to one another, 'dead' regions of the secondary abruptly enter a magnetised gap, and 'live' regions abruptly leave it at the other side. As a result there are important transient effects, the elimination or mitigation of which will impose quite unusual restrictions on the design, affecting not only the length and number of poles of the primary, but also its optimum working frequency.

20.3.10.3 Duty

There are, in general, three operating duties, which influence the design and construction.

Power (drive) This is concerned with the transport of loads in conveyors, haulers, electromagnetic pumps, travelling cranes and railway traction with acceptable overall *power* efficiency.

Energy (accelerator) Here the duty is to accelerate a mass from rest within a specified time and distance, as in rope-break and car-crash test rigs and the launching of aircraft. The criterion is the *energy* efficiency, i.e. the ratio of the energy imparted to the load and the total primary electrical energy input, a figure that cannot exceed 0.5.

Force (actuator) This develops a thrust at rest, or over a short stroke, as in the operation of stop-valves, impact metal-forming, door closers and small thrustors. The criterion is the *force* per unit electrical power.

20.3.10.4 General principles

The following analysis, applying to power (drive) types of linear motor, outlines in simplified terms the basic principles.

Speed The speed of a d.c. linear machine is associated with the secondary applied voltage and the flux per pole in a way comparable to the relationship between these variables in a rotary machine.

The speed of an induction-type linear machine is associated with the synchronous speed u, which is given by $u_1 = f\lambda$ for a supply frequency f, where λ is the wavelength (i.e. the length of a double pole-pitch). The actual speed u differs from u_1 because of the slip. If $\lambda = 1$ m and the supply frequency is 50 Hz, then $u_1 = 50$ m/s = 180 km/h. For lower translational speeds on a 50 Hz supply it is necessary to shorten the wavelength. However, if λ is less than about 0.2 m, corresponding to $u_1 = 10$ m/s = 36 km/h, the performance is impaired because the pole-pitch is short compared with the gap length. The effect is most significant in machines with an 'open' magnetic circuit (*Figure 20.61*(*c*)). Much better low-speed performance can be achieved if a low-frequency supply is available. In some cases in which starting from rest at high translational force is sought, a primary with a graded pole-pitch may be of advantage; alternatively the frequency can be raised during the starting period if the method can be economically justified.

Power If in a secondary member the surface current density is J and the gap flux density is B, the force developed per unit area of gap surface is BJ, and the motional e.m.f. developed at a translational speed u is Bu per unit width. The working voltage to be supplied to maintain motion is $v = Bu + J\rho$ per unit width, where ρ is the resistivity of the secondary surface current conducting path. Where the secondary is a conducting sheet (solid or liquid) the secondary current paths are ill-defined, so that estimation of performance is not straightforward.

In an induction linear machine with a two- or three-phase winding fed at frequency f and providing a travelling wave of gap flux at synchronous speed $u_1 = f\lambda$, the runner moves in the same direction at a lower speed u, i.e. with a slip given by $s = (u_1 - u)/u_1$. At a point in the gap where the instantaneous gap flux density is B_x, the e.m.f. induced in the secondary is $e_x = B_x s u_1$ per unit width, producing a secondary current of linear density $J_x = e_x/\rho$ (ignoring inductive effects in the current path). The interaction force per unit width is therefore $B_x J_x = B_x^2 s u_1/\rho$. Summed over a wavelength (one double pole-pitch of length λ) the force is $\frac{1}{2}\lambda s u_1 B_m^2/\rho$, where B_m is the peak density of the travelling wave of magnetic field. As the runner moves at speed $u_1(1-s)$, the mechanical power produced per wavelength and per unit width is

$$P_m = \frac{1}{2}s(1-s)u_1^2 B_m^2/\rho$$

This is a maximum for $s = 0.5$. The simple analysis applies only to wavelengths remote from the ends of the shorter member of the machine. In general, the precise current circuit of the secondary is somewhat indefinite, there are effects of leakage inductance, and near the ends of the primary block the 'dead' regions of the secondary abruptly enter the magnetised gap while 'live' regions as abruptly leave it. As a result there are transient 'entry and exit' effects, the mitigation of which imposes restrictions on the design that are absent in rotary types.

20.3.10.5 Applications

Some typical applications are described.

Electromagnetic pump The electromagnetic pump utilises the good electrical conductivity of the liquid metal being pumped to establish electromagnetic forces directly within the liquid itself. The liquid is contained in a simple pipe, which can be welded to the rest of the circuit, so that valves, seals and glands are avoided; this is desirable since the low-melting-point metals are highly reactive chemically. Because of the absence of glands and moving parts, the pump reliability should be high and maintenance costs low. In addition, the pump itself is often smaller and the amount of liquid contained sometimes less, which is important when expensive liquids are being handled. Moreover, in favourable cases, particularly with high-conductivity liquid metals, the pump is likely to be cheaper.

The most important applications for circulating liquid metal are in nuclear energy projects where the metal is used as a coolant: sodium, sodium–potassium and lithium are the metals chiefly concerned. Most other industrial applications are restricted to low-melting-point metals, as in the die-casting industry or in chlorine plants for pumping mercury, but a form of liquid-metal pump has been found useful for stirring molten steel in arc furnaces.

The liquids mentioned divide into two distinct classes: (1) sodium, sodium–potassium and lithium; and (2) bismuth, mercury, lead and lead–bismuth, which have much poorer pumping properties. For example, the resistivity of bismuth is nearly eight times that of sodium, its viscosity is five times higher and its density is 11 times higher. A high resistivity lowers the efficiency, while high viscosity and density make the pump appreciably larger, for, to reduce hydraulic losses, the liquid velocity must be low, and this increases the size of the pump duct.

The wide difference in liquid pumping properties influences the type of pump and the efficiency obtainable. Liquids like bismuth usually necessitate conduction pumps, in which current is supplied directly to the liquid through the tube walls. In contrast, sodium can be pumped by conduction or induction.

D.c. conduction pump The general arrangement is shown in *Figure 20.62*. The magnetic field can be produced by a permanent-magnet system or by an electromagnet. A series-excited electromagnet is preferred for large pumps as it is usually smaller and cheaper and involves only the provision of a few turns around the pole, as close as possible to the gap, to carry the main operating current. The supply requirements vary from 5 kA at 0.5 V for a pump of 0.05–0.1 m^3/min capacity at a pressure differential of 350 kN/m^2, up to 250 kA at 3–4 V for a nuclear reactor pump delivering 25 m^3/min at 500 kN/m^2. These inconvenient requirements form the major disadvantage of the d.c. pump; the compensating advantages are the minimal insulation levels (desirable if the liquid is hot or radioactive), the accommodating performance which can deal with a wide range of flow rates and pressures with good efficiency, and the adaptability to a variety of metals, including bismuth.

In d.c. pump design it is important to assess the field distortion due to the current in the liquid metal and to apply compensation if necessary, by returning the current to the electrode between the pole and the pump duct, usually as a pole face winding. It is also important for good efficiency to restrict the useless components of electrode current, which flows in the tube wall and in the liquid outside the pole region. The tube-wall current can usually be limited only by choosing a thin-walled tube of high intrinsic resistivity. The useless current can be substantially reduced by fringing the field to match the natural fringing of the current density between the electrodes. It is also important in design to limit magnetic leakage by appropriately positioning the magnetising turns and by shaping the iron.

The operation can be approximately described in terms of the flow of a quantity q of liquid at velocity u in a tube of width b in the current direction and effective length c in the direction of the liquid movement, the liquid having a resistivity ρ:

Electrodes: voltage $V_0 = (Bu + J\rho)b$;
current $I_0 = I = I_t = I_s$

Pump: gross pressure $p = (B_m I)c$;
gross power $P_0 = pq$

Here B is the flux density, of mean value B_m, and J is the current density. The electrode current I_0 includes the useful current I, and non-useful shunt currents I_t in the tube walls and I_s in the liquid outside the magnetised region. The gross pressure p includes the hydraulic pressure, drop in skin friction, etc.

A.c. induction pump The two basic forms are the flat and the annular. The flat type (*Figure 20.63*) has a straight duct 10–25 mm wide and up to 1 m across. Copper bars are attached to each side of the channel to form the equivalent of the 'end-rings' of a normal cage winding. A flat polyphase winding is placed on each side of the duct. A heat-insulating blanket is usually necessary between duct and winding, which is force cooled. The annular form is similar except that the cross-section of the duct is an annulus and embraces a radially laminated magnetic core. The polyphase winding comprises a number of pancake coils surrounding the duct and set in comb-like stacks of radial punchings. If the duct is made re-entrant, with inlet and outlet at the same end, it is possible to remove the primary windings for repair and maintenance, should this become necessary, without disturbing the pipe-work.

Cranes Two linear induction motors can be applied to the traversing of an overhead crane, replacing the conventional rotary cage motor, gearing, drive shaft and control equipment. Maintenance is simplified and, as the linear motors are unaffected by atmospheric conditions, the likelihood of breakdown is reduced. The arrangement for one of the motors is shown schematically in *Figure 20.64*. Each motor comprises a horizonal pack of laminations carrying a three-phase winding. Motors are located at each end of the crane gantry and directly below the tracks from which the crane is suspended. The track functions as the secondary, the arrangement being basically that of *Figure 20.61*(*c*).

Traction Several schemes have been proposed for railway traction. One uses a short-primary form of motor on the locomotive and a fixed 'plate' secondary secured to the track. The moving primary demands either that power be supplied to the moving train or that a prime mover and three-phase generator be carried on board. Some recent

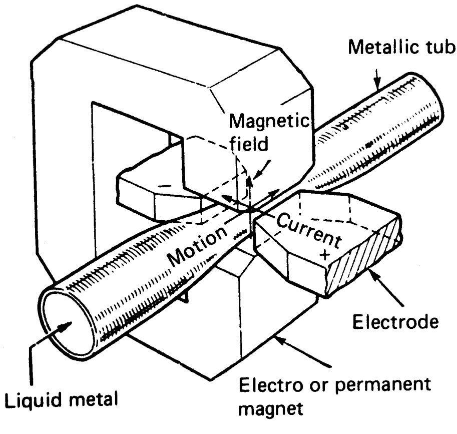


Figure 20.62 D.c. conduction pump

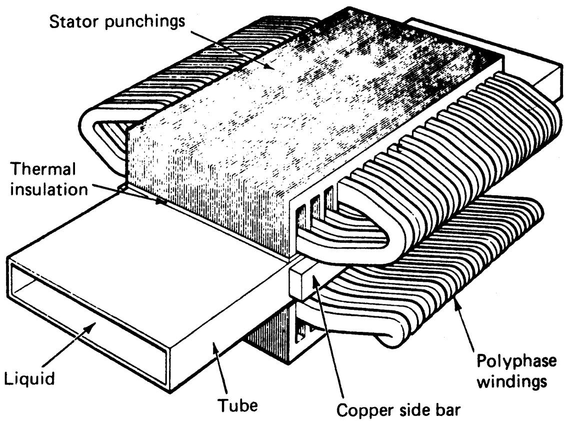


Figure 20.63 A.c. induction pump

developments exploit magnetic levitation and the elimination of running wheels.

Stirring The stirring of molten metal in furnaces can be done by external application of short linear induction motors. The process has considerable metallurgical advantage in improving the casting properties of aluminium and copper, and in accelerating the deoxidation processes in steel. The primary of the motor does not come into contact with the melt, is readily controlled and can be arranged for vertical, rotary or horizontal stirring.

Short-stroke devices The linear induction motor offers a compact and readily controllable means for the automation of punching, stamping and impact extrusion. With the wavelength graded on the fixed primaries to raise the 'synchronous' speed as the short secondary runner approaches the workpiece, end speeds of 30 m/s and considerable kinetic energy can be attained. Other industrial applications include the following:

Sliding doors: here an advantage is that, should the supply fail, the motor does not impede movement of the door by hand.

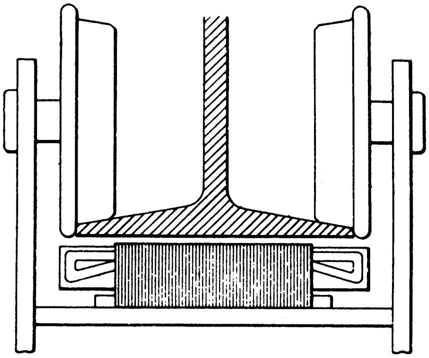


Figure 20.64 Linear traversing crane motor

Goods lifts: the arrangement is that of *Figure 20.64* turned vertically, the linear motor stator providing some of the counterweight.

Tensioning machines: the linear motor gives readily controllable tension for testing ropes, aluminium strip, etc.

Section E
Environment

21

Lighting

N A Smith

Contents

21.1 Light and vision 21/3

21.2 Quantities and units 21/3

21.3 Photometric concepts 21/4

21.4 Lighting design technology 21/6

21.5 Lamps 21/8
  21.5.1 Incandescent filament lamps 21/8
  21.5.2 Discharge lamps 21/10
  21.5.3 Mercury lamps 21/12
  21.5.4 Sodium lamps 21/17
  21.5.5 Control gear 21/19
  21.5.6 Electroluminescent devices 21/19
  21.5.7 Lamp life 21/20

21.6 Lighting design 21/20
  21.6.1 Objectives and criteria 21/20
  21.6.2 Luminaires 21/24

21.7 Design techniques 21/27
  21.7.1 Lighting systems 21/28
  21.7.2 Lighting surveys 21/28

21.8 Lighting applications 21/29
  21.8.1 Office and interior lighting 21/29
  21.8.2 Factory lighting 21/30
  21.8.3 Security lighting 21/30
  21.8.4 Floodlighting 21/30
  21.8.5 Public lighting 21/30
  21.8.6 Light pollution 21/31

21.1 Light and vision

Light is electromagnetic radiation, i.e., it has electric and magnetic fields, mutually at right angles and varying sinusoidally as shown in *Figure 21.1*. It is capable of causing a visual sensation in the eye of an observer. It is measured in terms of its ability to produce such a sensation.

The *spectral range* of visible radiation is not well defined, and can vary with the observer and with conditions. The lower limit is generally taken to be 380–400 nm (deep-blue radiation) and the upper limit 760–780 nm (deep-red radiation).

The human eye is not equally sensitive to all wavelengths, as shown in *Figure 21.2*. For normal daylight vision, referred to as *photopic vision*, the eye has a peak sensitivity at 555 nanometres (nm). The eye contains two distinct types of light-sensitive receptors referred to as *rods* and *cones*. The cones are responsible for colour vision whilst the rods operate in dark conditions.

At low levels of illumination the more sensitive rods begin to take over, and the resultant image appears less brightly coloured. Furthermore, the peak sensitivity shifts towards the blue/green region of the spectrum. This condition is known as *mesopic* vision.

At still lower levels vision is almost entirely by rod receptors and the eye is said to be dark-adapted. In this state, known as *scotopic* vision, the sensation is entirely in black and white and the peak sensitivity has moved to 505 nm.

The Commission Internationale de l'Eclairage (CIE) has defined an agreed response curve for the photopically adapted eye, known as the *spectral luminous efficacy* or $V(\lambda)$ function. Luminous flux, which is the rate of flow of light, is radiant power weighted according to its ability to produce a visual sensation by the $V(\lambda)$ function.

The luminous flux emitted by a source of light will vary with direction of emission. The rate of change of luminous flux with solid angle is termed the *luminous intensity*.

Illumination is the process whereby luminous flux is incident upon a solid surface and the corresponding quantity (flux density per unit area) is the *illuminance*.

Light striking a surface can be reflected, transmitted or absorbed according to the nature of the surface, and the fractions of the incident luminous flux thus affected are termed the reflectance, transmittance or absorptance, respectively.

21.2 Quantities and units

Each quantity has a quantity symbol (e.g. I for luminous intensity) and a unit symbol (e.g. cd for candela) to indicate its unit of measurement.

Luminous flux, Φ; lumen (lm) The rate of flow of luminous energy. A quantity derived from radiant flux by evaluating it according to its ability to produce visual sensation. Unless otherwise stated, luminous flux relates to photopic vision as defined by the $V(\lambda)$ function of spectral luminous efficacy.

If K_m is the maximum spectral luminous efficacy (about 680 lm . W^{-1} at a wavelength of 555 nm), then the luminous flux Φ (in lm) is related to the spectral power distribution $P(\lambda)$ at wavelength λ by

$$\Phi = K_m \int P(\lambda) \,.\, V(\lambda) \,.\, d\lambda$$

Luminous efficacy (of a source), η; lumens per watt (lm . W^{-1}) The quotient of the luminous flux emitted by a source to the input power. It should be noted that for discharge lamps the luminous efficacy may be quoted either for the lamp itself or for the lamp with appropriate control gear. The latter figure will be lower.

Luminous intensity, I; candela (cd) The quotient of the luminous flux $\delta\Phi$ leaving the source, propagated in an element of solid angle containing the given direction, by the element of solid angle $\delta\omega$ (see *Figure 21.3*).

$$I = d\Phi / d\omega$$

Illuminance, E; lux (lx) or lumens per metre^2 (lm . m^{-2}) The incident luminous flux density at a point on a surface. The quotient of the luminous flux incident on an element of surface, by the area of that element. Referring to *Figure 21.3*,

$$E = \delta\Phi / \delta A$$

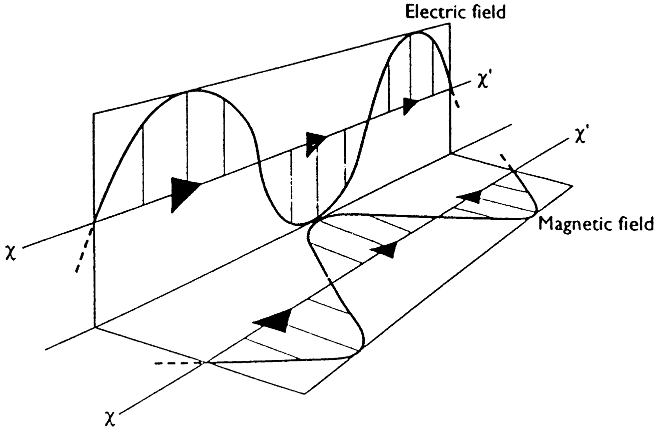


Figure 21.1 Electric field and magnetic field mutually at right angles. Electric field and magnetic field have same axis x–x′ but are shown separately for clarity only

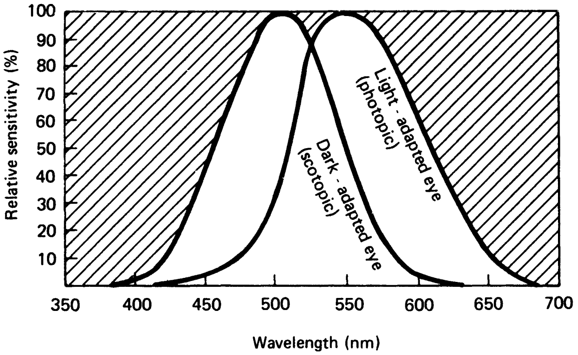


Figure 21.2 The relative spectral sensitivity of the human eye

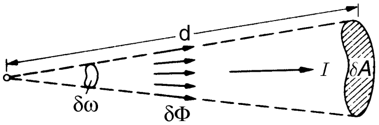


Figure 21.3 Luminous intensity and illumination

Note: the term *illuminance* is used for the quantity, while the term *illumination* describes the physical process.

Luminance, *L*; candela per metre^2 (cd m^{-2}) The luminous intensity in a given direction of a surface element, per unit projected area of that element.

Luminance is a physical measure of brightness, but it should be noted that an observer's assessment of the brightness of an object is subjective, unlike luminance which is the objective physical measure. It will depend upon the level of adaptation and other factors. For example, the luminance of a car headlight during the day and at night would be approximately the same, but the apparent brightness during the day would be significantly less.

21.3 Photometric concepts

Inverse square law The illuminance *E* at a point on a surface produced by light from a *point* source varies inversely with the square of the distance *d* from the source, and is proportional to the luminous intensity *I* towards that point. Referring to *Figure 21.3*, the illuminance is given by

$$E = I/d^2.$$

Cosine law The illuminance on a surface is proportional to the cosine of the angle θ between the directions of the incident light and the normal to the surface. This is due to the reduction of projected area as the angle of incidence increases from zero (normal incidence) to 90°. For a point source at distance *d*, the illuminance for angle θ is

$$E = E_0 \cos\theta = I \cos\theta / d^2$$

where E_0 is the illuminance for normal incidence, $\theta = 0$. With the working surface horizontal and the source mounted a distance *h* above the surface, the illuminance on the working surface is

$$E = I \cos^3\theta / h^2$$

involving θ as the only variable.

Reflection Light falling on a surface may undergo *direct* or *diffuse* reflection. Direct reflection is specular, as by a mirror. Diffuse reflection may be *uniform* or *preferential*: in the former the luminance is the same in all available directions; in the latter there are maxima in certain directions (see *Figure 21.4*). Direct and diffuse reflection may occur together as *mixed* or *spread* reflection.

Examples of reflecting surfaces are: direct (mirror glass, chromium plate); uniform diffuse (blotting paper); preferential diffuse (anodised aluminium, metallic paint).

Reflectance, *R* The ratio between the reflected luminous flux and the incident luminous flux.

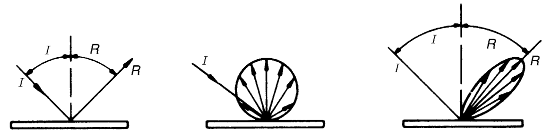
Transmission Light falling on a translucent surface undergoes partial transmission (*Figure 21.5*). The transmission may be *direct*, as through clear plate glass; *diffuse*, as through flashed opal glass; or *preferential*, as through frosted glass.

Transmittance, *T* The ratio between the transmitted luminous flux and the incident luminous flux.

Absorption That proportion of light flux falling on a surface which is neither reflected nor transmitted is *absorbed* and, normally, converted into heat.

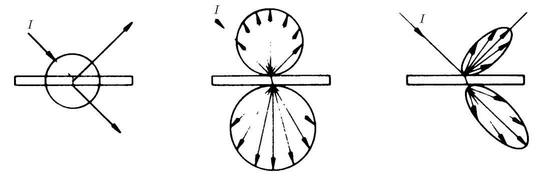
Refraction While a light ray is travelling through air, its path is a straight line. When the ray passes from air to glass (or any transparent material, e.g. clear plastics, diamonds, etc.), the ray is, in general, bent at the surface of separation. The path of the ray after bending or refraction from air to glass is always more nearly perpendicular to the bounding surface than is that of the incident ray. The degree to which the ray is bent depends on the type of glass or transparent material, the angle of incidence of the ray and also the colour of the light.

Should the ray, while in the glass, strike another bounding surface, it may again be refracted. In this case the



DIRECT REFLECTION DIFFUSED REFLECTION SPREAD REFLECTION

Figure 21.4 Reflection



DIRECT TRANSMISSION DIFFUSED TRANSMISSION SPREAD TRANSMISSION

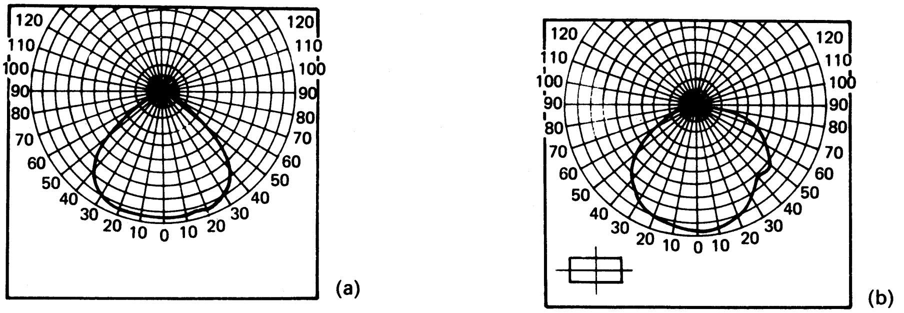
Figure 21.5 Transmission with partial reflection

refracted ray may be more nearly parallel to the bounding surface than is the incident ray. If the light ray strikes the bounding surface at any angle above a certain limit (the *critical* angle), it will not be refracted but will be totally reflected.

Both refraction and total internal reflection are used in the design of lighting units, the prismatic types of reflector being typical examples.

Polar intensity distributions Utilising the inverse-square and cosine laws, it is possible to calculate the direct illuminance at a point from a single luminaire, or an installation, using the 'point-by-point' method. The effect of inter-reflected light is not included, as the calculations are too complex to warrant it.

Figure 21.6 shows the intensity distributions for two different interior luminaires: (*a*) is for a luminaire with a luminous intensity distribution symmetrical about a vertical axis, and (*b*) is for a non-symmetrical luminaire with a luminous intensity distribution symmetrical about two orthogonal vertical planes. These are typical of discharge and fluorescent luminaires, respectively.



(a)

Luminous intensity (cd per 1000 lm)	
Angle (deg)	Mean vertical intensity (cd)
0	234
5	234
10	235
15	236
20	234
25	232
30	230
35	222
40	205
45	180
50	137
55	95
60	66
65	46
70	30
75	19
80	13
85	11
90	9
95	9
100	9
105	10
110	11
115	13
120	15
125	18
130	22
135	28
140	32
145	35
150	34
155	30
160	26
165	20
170	14
175	7
180	4

(b)

	Luminous intensity (cd per 1000 lm)	
Angle (deg)	Transverse plane (T)	Axial plane (A)
0	232	232
5	232	230
10	231	228
15	231	224
20	228	217
25	224	208
30	220	199
35	211	187
40	192	174
45	168	199
50	141	187
55	113	174
60	86	159
65	58	142
70	33	124
75	19	106
80	12	85
85	7	67
90	1	46
95	2	28
100	5	11
105	7	1
110	8	2
115	9	2
120	11	2
125	12	3
130	13	4
135	14	5
140	14	6
145	15	7
150	16	8
155	16	9
160	16	10
165	15	10
170	13	11
175	12	11
180	12	12

Figure 21.6 Luminous intensity distributions for: (a) symmetric luminaires and (b) non-symmetric luminaires

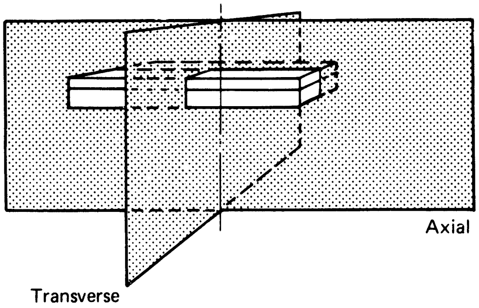


Figure 21.7 The transverse and axial planes in which the transverse and axial polar curves are measured

For symmetric luminaires, only one average intensity distribution is normally given, and this can be presented graphically on polar co-ordinates or in tabular form (which is easier to use). For non-symmetrical luminaires two or more distributions are given. The principal ones are the *axial* and *transverse* distributions, which lie in vertical planes down the axis of the luminaire, and at right angles thereto, respectively (*Figure 21.7*).

Many luminaires can accommodate various lamp types without affecting the shape of the intensity distribution. For this reason it is a common practice to quote intensities in candelas per 1000 lamp lumens ($\text{cd}\,.\,\text{klm}^{-1}$) rather than in candelas. This permits easy scaling of the data according to the luminous output of the lamps.

For streetlighting and floodlighting luminaires, the main distributions are usually insufficient, and contours of equal intensity (isocandela) are normally published on a convenient Cartesian grid system, as shown in *Figure 21.8*.

Isolux diagrams A convenient way of plotting the illuminance produced by single luminaires or complete installations is by contours of equal illuminance, or isolux contours. Isolux diagrams are frequently used to depict the performance of non-symmetrical luminaires such as 'wall-washers', and are now often used when the calculations are done by computer. *Figure 21.9* shows a typical isolux diagram for a reflector luminaire at a particular mounting height.

21.4 Lighting design terminology

Light output ratio (LOR) The ratio between the light output of the luminaire measured under specified practical conditions and the sum of the light outputs of individual lamps operating outside the luminaire under reference conditions.

Photometric centre The point in a luminaire or lamp from which the inverse square law operates most closely in the direction of maximum intensity.

Upward (downward) flux fraction (UFF (DFF)) The fraction of the total luminous flux of a luminaire emitted above (below) the horizontal plane containing the

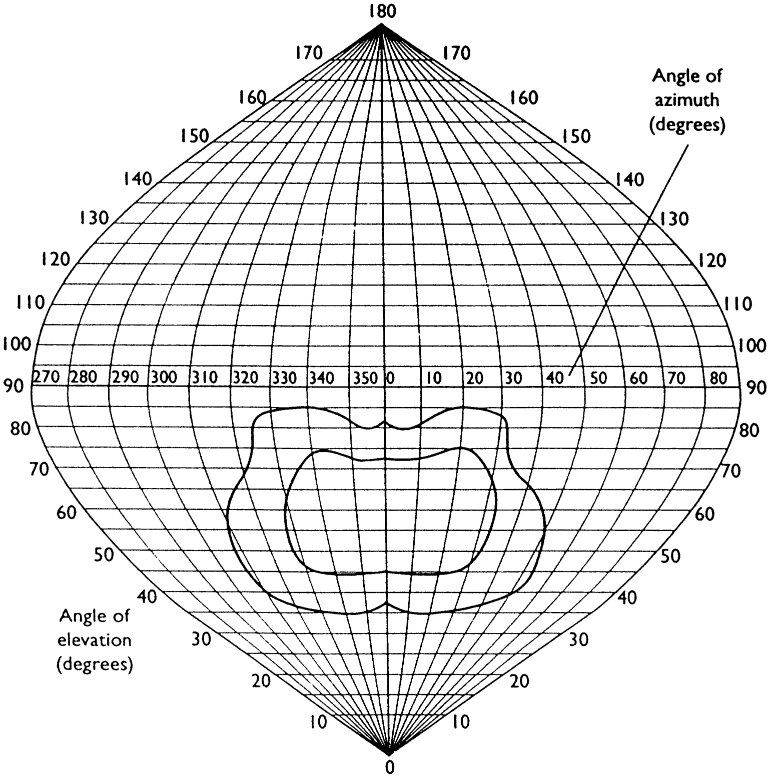


Figure 21.8 Typical isocandela diagram. Figures on contour lines represent luminous intensities in candela per 1000 lamp lumens

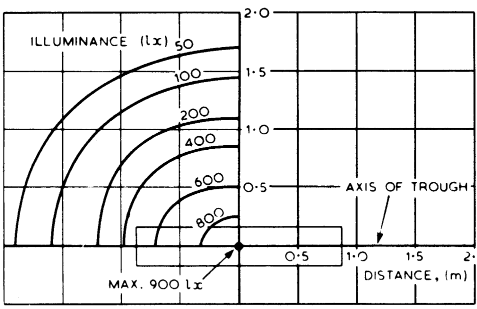


Figure 21.9 Isolux diagram for a 1.8 m long trough reflector luminaire

photometric centre of the luminaire. Also known as upper (lower) flux fraction.

Upward (downward) light output ratio (ULOR (DLOR)) The product of the light output ratio of a luminaire and the upward (downward) flux fraction.

Symmetric luminaire A luminaire with a light distribution nominally rotationally symmetrical about the vertical axis passing through the photometric centre.

Non-symmetric luminaire A luminaire with a light distribution nominally symmetrical only about two mutually perpendicular planes passing through the photometric centre. Where such a luminaire is linear, the vertical plane of symmetry normal to the long axis is designated the *transverse* plane, and the vertical plane passing through the long axis is designated the *axial* plane (see *Figure 21.7*). The vertical distributions taken in these planes are the transverse and axial distributions, respectively.

Working plane The horizontal, vertical or inclined plane in which the visual task lies.

Reference surface The surface of interest over which the illuminance is to be calculated. A reference surface need not contain the visual task.

Horizontal reference plane A horizontal reference surface. This is usually assumed to be 0.85 m above the floor and to correspond with the horizontal working plane. The horizontal reference plane is also the mouth of the floor cavity.

Plane of luminaires The horizontal plane which passes through the photometric centres of the luminaires in an installation. This is also the mouth of the ceiling cavity.

Floor cavity The cavity below the horizontal reference plane in a room (see *Figure 21.10*). The horizontal reference plane and the floor cavity may be designated by the reference letter F.

Walls The vertical surfaces of a room between the plane of the luminaires and the horizontal reference plane (see *Figure 21.10*). The walls may be designated by the reference letter W.

Ceiling cavity The cavity above the plane of the luminaires in a room (see *Figure 21.10*). The luminaire plane and the ceiling cavity may be designated by the reference letter C.

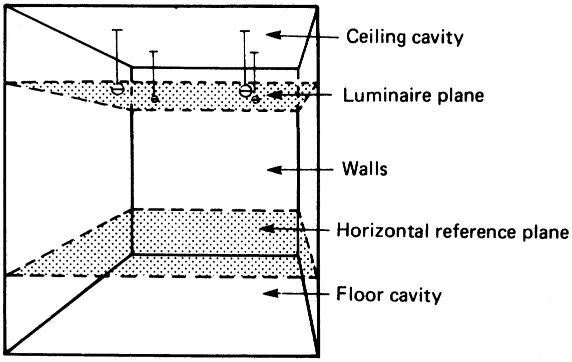


Figure 21.10 Ceiling cavity, walls and floor cavity

Distribution factor, DF(S) The distribution factor for a surface S is the ratio between the direct flux received by the surface S and the total lamp flux of the installation. DF(F), DF(W) and DF(C) are the distribution factors for the floor cavity, walls and ceiling cavity, respectively, treated as notional surfaces.

Utilisation factor, UF(S) The utilisation factor for a surface S is the ratio between the total flux received by the surface S (directly and by inter-reflection) and the total lamp flux of the installation. UF(F), UF(W) and UF(C) are the utilisation factors for the floor cavity, walls and ceiling cavity, respectively, treated as notional surfaces.

Direct ratio, DR The proportion of the total downward flux from a conventional installation of luminaires that is directly incident on the horizontal reference plane. The direct ratio is equal to DF(F) divided by the DLOR of the luminaires.

Zone factor The solid angle subtended at the photometric centre of a lamp or luminaire by the boundary of a zone. The zonal flux is obtained by multiplying the intensity of the lamp or luminaire, averaged over the zone, by the zone factor.

Room index, RI Twice the plan area of a room divided by the wall area (as defined above). The room is taken to have parallel floor and ceiling, and walls at right angles to these surfaces. From any point in the room all of the surfaces in the room should be visible.

Spacing/height ratio, SHR The ratio between the spacing in a stated direction between photometric centres of adjacent luminaires and their height above the horizontal reference plane. It is assumed that the luminaires are in a regular square array unless stated otherwise.

Maximum spacing/height ratio, SHR MAX The SHR for a square array of luminaires that gives a ratio between minimum and maximum direct illuminance of 0.7 over the central region between the four innermost luminaires.

Maximum transverse spacing/height ratio, SHR MAX TR The SHR in the transverse plane for continuous lines of luminaires that gives a ratio between minimum and maximum direct illuminance of 0.7 over the central region between the two inner rows.

Nominal spacing/height ratio, SHR NOM The highest value of SHR in the series 0.5, 0.75, 1.0, etc., that is not greater than SHR MAX. Utilisation factor tables are normally calculated at a spacing/height ratio of SHR NOM.

Maintenance factor, *MF* The ratio between the illuminance provided by an installation in the average condition of dirtiness expected in service and the illuminance from the same installation when clean. It is always less than unity.

Uniformity The ratio between the minimum and average illuminance over a given area. For interior lighting it should be not less than 0.8 over the task area. This requirement can be satisfied by ensuring that the spacing/height ratio of an installation does not exceed SHR MAX.

Daylight factor, *DF* The ratio between the illumination measured on a horizontal plane at a given point inside a building and that due to an unobstructed hemisphere of sky. Light reflected from interior and exterior surfaces is included in the illumination at the point, but direct sunlight is excluded.

21.5 Lamps

Light can be produced from electrical energy in a number of ways, of which the following are the most important.

(1) *Thermoluminescence*, or the production of light from heat. This is the way light is produced from a filament lamp, in which the filament is incandescent.
(2) *Electric discharge*, or the production of light from the passage of electricity through a gas or vapour. The atoms of the gas are excited by the passage of an electric current to produce light and/or ultraviolet energy.
(3) *Fluorescence*, a two-step production of light which starts with ultraviolet radiation emitted from a discharge; the energy is then converted to visible light by a phosphor coating within the lamp.

21.5.1 Incandescent filament lamps

Thermoluminescence is the emission of light by means of a heated filament. The term is normally synonymous with the tungsten filament lamp in its various forms. The most general form is the general lighting service (GLS) lamp (*Figure 21.11*). Light produced from a hot wire increases as the temperature of the wire is raised. It also changes from a predominantly red colour at low temperature to a white which approaches daylight as the temperature is increased.

Of the electrical energy supplied to an incandescent lamp filament, by far the greatest proportion is dissipated as heat, and only a small quantity as visible light (about 95% heat and 5% light) (see *Figure 21.12*). Because the quantity of visible light emitted depends upon the filament temperature, the higher the filament temperature the greater will be the visible light output in lumens per watt of electric power input. Thus, for an incandescent filament, a material is needed that not only has a high melting point, but is also strong and ductile so that it can be formed into wire. At present, tungsten metal is the material nearest to this ideal.

The colour temperature of a normal GLS filament lamp is typically between 2800 K and 3000 K. At the extremely high temperature of the filament, tungsten tends to evaporate. This leads to the familiar blackening of an incandescent lamp envelope. The evaporation of the tungsten filament can be reduced by filling the lamp envelope with a suitable gas that does not chemically attack the filament. Suitable gases are hydrogen, nitrogen, and the inert gases argon, neon, helium, krypton and xenon. However, gases also cool the filament by conducting heat away from it, and they decrease lamp efficiency. The gas used must therefore be carefully chosen. It should adequately suppress tungsten evaporation without overcooling the filament. In addition, it should not readily pass an electric current, for otherwise arcing may occur which would destroy the lamp.

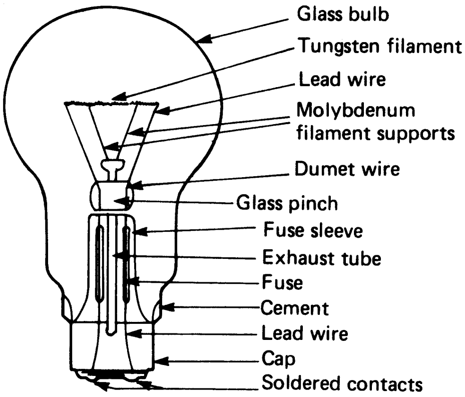


Figure 21.11 Construction of a GLS lamp

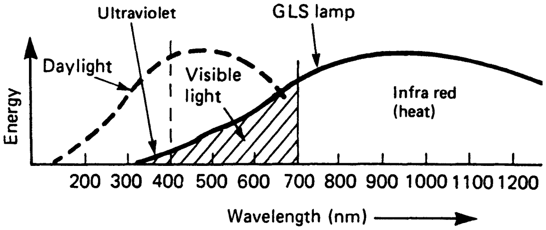


Figure 21.12 Spectral power distribution of daylight and a GLS lamp

Argon and nitrogen are the gases most commonly used. Nitrogen will minimise the risk of arcing, but will absorb more heat than argon. Argon is used by itself in general service lamps. A mixture of the two gases is used in incandescent lamps where the tendency for arcing is more likely, such as in projector lamps. In this case the amount of nitrogen present is kept very small—as little as 5%—in order to obtain optimum lamp efficiency.

Not all incandescent lamps benefit from gas filling. Mains voltage 15 W and 25 W lamps are mainly of the vacuum type, whereas lamps of 40 W and above are normally gas filled.

In general service lamps at least one lead is fused to prevent the envelope shattering should an arc occur. Modern fuses are encapsulated in a glass sleeve filled with small glass balls.

21.5.1.1 *Coiled and coiled-coil filaments*

If a filament is in the form of an isolated straight wire, gas can circulate freely round it. Filament temperature is thus decreased by convection currents, and has to be raised by increasing the electrical power input.

Coiling the wire reduces the cooling effect, the outer surface of the helix alone being cooled by the gas. Further coiling (coiled-coil filament) again reduces the effect of the gas cooling and results in further increase in lamp efficiency of up to 15%.

21.5.1.2 Glass envelopes

Clear-glass lamp envelopes have smooth surfaces and absorb the smallest possible amount of the light passing through them. The high temperature of the filament results in a high brightness which the envelope does not modify.

Early attempts to reduce glare from an unobscured filament used envelopes externally frosted. These were difficult to keep clean. The drawback is obviated in the modern *pearl* envelope by etching the inside surface instead. The light source appears to be increased in size and to have a larger surface area. The loss of light is negligible.

With the greatly increased illumination levels of modern lighting techniques, a further degree of diffusion is called for. This has been achieved by coating the inside of the envelope with a very finely divided white powder, such as silica or titania. In such lamps the lighted filament is not apparent. The luminous efficiency of the silica coated lamp is about 90% of that of a corresponding clear lamp of equal power rating. Silica coated lamps have a more attractive unlit appearance than either clear or pearl lamps.

In a coloured incandescent lamp the envelope is coated either internally or externally with a filter. All coloured incandescent lamps operate at reduced efficiency. In view of the low proportion of blue light in the spectrum, the efficiency of lamps of this colour is particularly low, as more than 90% of the light is filtered out. It is not possible to obtain a bright saturated blue colour.

21.5.1.3 Decorative and special-purpose lamps

The incandescent filament lamp in its simplest form is purely a functional light source, but the fact that an integral part of the lamp has a glass envelope enables the manufacturer to adapt this envelope to give some aesthetic appeal. The commonest form is the candle lamp, with glass clear, white or frosted. Other lamps have been marketed which combine the role of light source and decorative luminaire by virtue of their envelope shape. They are usually of larger dimensions than conventional lamps. Apart from their attractive shapes, they are made with silica coatings, coloured lacquer coatings and crown silvered tops, and are therefore rather more efficient as light-producing units than a combination of lamp and separate diffuser.

To cater for locations where vibration and shock are unavoidable, special *rough service* lamps are produced which combine filament wire modifications with the inclusion of an increased number of intermediate filament supports.

To provide directional beam control a further range of special-purpose lamps is made with blown or pressed paraboloidal envelope shapes coated with an aluminium reflector film. The filament is accurately placed at the focus of the reflector to provide the directional beam. More accurate beam control is provided by the pressed glass versions (PAR lamps).

In some lamps dichroic reflectors are employed. These reflect visible light but transmit infrared radiation. This permits the lamp to have a cooler beam, since the heat radiation is not focused. However, these lamps can be used only in luminaires able to dissipate the extra heat transmitted by the reflector. Dichroic coatings are widely used in film projector lamps with integral reflectors, to prevent excessive temperature at the film gate. *Figure 21.13* shows the concept of the dichroic lamp

The efficacy of an incandescent lamp is related to the quantity of visible light emitted per unit of electrical power input. Thus, a 100 W incandescent lamp having a total light output of 1200 lm has an efficacy of 1200/100 = 12 lm . W^{-1}.

A higher filament temperature increases lamp efficacy, but the temperature of a tungsten filament cannot be increased indefinitely, as it will melt catastrophically if the lamp efficacy approaches 40 lm . W^{-1}.

At high filament temperatures tungsten evaporation—even though it is reduced by gas filling—is more rapid and leads to a shorter lamp life. Thus, the more efficient an incandescent lamp is the shorter is its life.

Variations in supply voltage vary filament temperature, which, in turn, increases or decreases lamp life. *Figure 21.14* shows how the lamp efficiency, life, light output and input power vary with supply voltage. For example, if a lamp is under-run by 5% below its rated voltage, its life will be nearly doubled (190% of rated 1000-h life) but the lamp power would be reduced to around 92% of the rating and the light output to less than 85% of the normal lumen output.

21.5.1.4 Tungsten halogen lamps

If the envelope of a tungsten lamp is made of quartz instead of glass, it can be much smaller, because quartz can operate safely at a higher temperature. As with a glass lamp, tungsten evaporated from the filament will deposit on the quartz envelope, causing it to blacken. However, if a small quantity of one of the halogen elements e.g. iodine is introduced into the lamp, and if the temperature of the quartz envelope is above 250°C, the iodine combines with the tungsten on the inner face of the quartz to form tungsten iodide, a vapour.

When the tungsten iodide approaches the much hotter filament, it decomposes; the tungsten is deposited on the filament and the iodine is released, to perform its cleaning cycle again (see *Figure 21.15*).

Unfortunately, the tungsten is not necessarily redeposited on those parts of the filament from which it originally

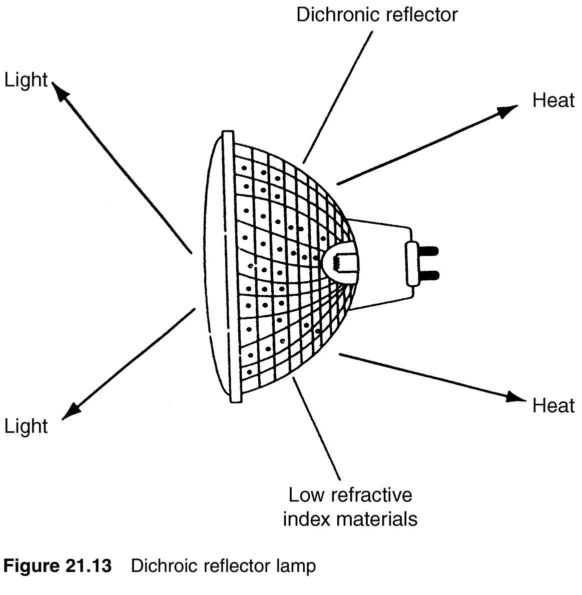


Figure 21.13 Dichroic reflector lamp

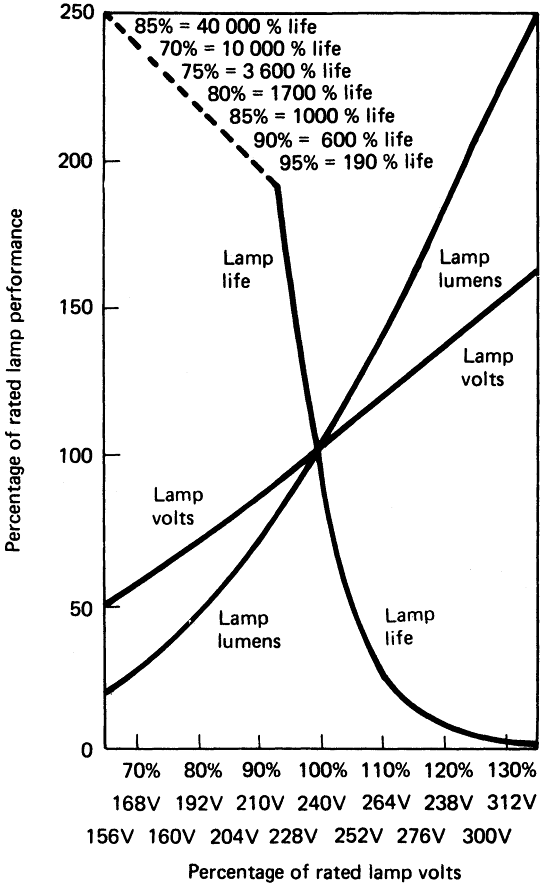


Figure 21.14 Typical effect of voltage on incandescent filament lamps

evaporated. Even so, substantial improvements in life and/or higher filament operating temperatures can be achieved, giving higher lumen outputs compared with the equivalent GLS lamp.

The increase in life is mainly due to the increased gas pressure, which can be employed in a tungsten halogen lamp to reduce filament evaporation. This, in turn, is only possible because a small outer envelope can be used without risk of lamp envelope blackening.

Tungsten halogen lamps give greater life and greater efficiency than their incandescent counterparts. For this reason they are widely used in floodlighting, photographic lighting, display lighting and automobile lighting. A typical tungsten halogen lamp will provide about 50% greater light output and about twice the life of a conventional tungsten lamp of an equivalent wattage.

Tungsten halogen lamps come in several common forms. Small capsule lamps with bi-pin lamp holders are used in spotlights and for similar applications needing good optical control and a small optical light source. The same type of lamp, but optimised to give shorter life and much higher light output, is used in photographic projectors and similar applications.

Small capsule lamps can also be built into glass or metal reflectors which form part of the lamp. These small diameter, high performance lamps are now widely used for display lighting. With such lamps the wattage, lamp diameter (typically 20 mm or 50 mm) and beam angle are normally specified. One advantage of this type of mirror lamp is that the reflector can be manufactured from glass with dichroic coatings. Such reflectors are designed to reflect light into the beam but not the heat. Lamps of this type are popular for display lighting because they direct very little heat onto the merchandise being displayed.

The third type of tungsten halogen lamp is the linear lamp. These are used for applications where a wide beam angle is required together with a high wattage. Such lamps are used for floodlights where low capital cost or instant white light is required. The lamps have an electrical contact at each end and vary in length according to wattage.

Tungsten halogen lamps are designed to operate with an envelope temperature of 250°C to 350°C. The design of the luminaire and the location of the equipment should ensure that people cannot touch the lamp and burn themselves. Similarly, flammable objects should be kept away from the lamp.

In a conventional tungsten lamp, although some ultraviolet radiation is produced, most of it is severely attenuated by the glass outer envelope. In a tungsten halogen lamp, the higher operating temperature and quartz envelope produces a greater ultraviolet content. Although ultraviolet radiation is present in sunlight and daylight, steps should be taken to limit human exposure where high lighting levels are involved.

Both of the above problems can be eliminated if either the lamp or the luminaire is fitted with an ultraviolet radiation absorbing front glass.

The quartz envelope of a tungsten halogen lamp should not come into contact with the human skin e.g. fingers. The greases and acids, which are present on the surface of the skin, will attack the quartz. This will subsequently produce blistering of the envelope leading to premature lamp failure.

21.5.2 Discharge lamps

21.5.2.1 Principle

A discharge lamp consists essentially of a tube of glass, quartz or other suitable material, containing a gas and, in most cases, a metal vapour. The passage of an electric current through this gas/vapour produces light or ultraviolet radiation. Most practical discharge lamps (excluding those used for coloured signs) rely upon discharges in metallic vapours of either sodium or mercury, with an inert gas filling. The nature of the filling, the pressure developed and the current density determine the characteristic radiation produced by the arc. In most lamps the arc tube is enclosed within an outer glass or quartz jacket. This affords protection, can be used for phosphor or diffusing coatings, control of ultraviolet radiation emission and, by suitable gas filling, can control the thermal characteristics of the lamp.

All discharge lamps include some mechanism for the production of electrons from the electrodes within the lamp. The commonly used methods are thermionic emission and field emission, and in both cases emissive material such as barium oxide is often contained within the electrode to lower its work function and, hence, reduce energy loss.

When the lamp is put into circuit and an electric field is applied, the electrons begin to accelerate towards the positive electrode, and may collide with gas or metal atoms.

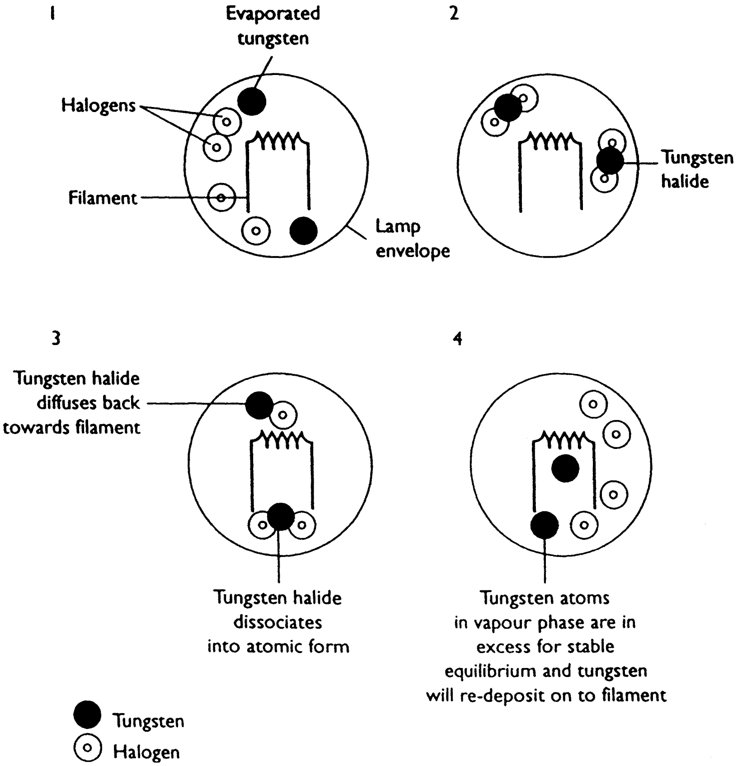

1
Evaporated tungsten
Halogens
Filament
Lamp envelope
2
Tungsten halide
3
Tungsten halide diffuses back towards filament
Tungsten halide dissociates into atomic form
4
Tungsten atoms in vapour phase are in excess for stable equilibrium and tungsten will re-deposit on to filament
● Tungsten
⊙ Halogen

Figure 21.15 Tungsten halogen cycle

These collisions may be elastic, in which case the atom and electron only change their velocities, or inelastic, in which case the atom changes its state. In the latter case, if the kinetic energy of the electron is sufficient, the atom may become excited or ionised. Ionisation produces a second electron and a positive ion, which contribute to the lamp current and which may cause further collisions. Left unchecked, this process would avalanche, destroying the lamp. To prevent this catastrophe some form of electrical control device (such as an inductor) is used to limit the current. Excitation occurs when the electrons within the atom are raised to an energy state higher than normal (but not high enough to cause ionisation). This is not a stable condition, and the electrons fall back to their previous energy level, with a corresponding emission of electromagnetic radiation (which may be visible, ultraviolet or infrared).

In some lamp types, an inert gas is used to maintain the ionisation process, while it is the metal vapour which becomes excited. The vapour pressure in the lamp affects the starting and running characteristics, and the spectral composition of the emitted radiation.

In most lamps there is a run-up period, during which the metal is vaporised and the pressure increases to its operating condition. In some lamp types this may take 10–15 min. If, once the lamp is run-up, the supply is interrupted, then it will extinguish; and unless special circuits and suitable lamp construction are used, the pressure will be too high for the arc to restrike until the lamp has cooled.

Broadly, practical discharge lamps for lighting are either mercury vapour or sodium vapour lamps, at either high or low pressure.

21.5.2.2 Run-up efficiency

Smith devised a method of calculating the 'Run-up efficiency' of a discharge lamp (see *Lighting for Health and Safety*, Butterworth-Heinemann, ISBN 0-7506-4566-0).

Figure 21.16 shows the concept of 'run-up efficiency', which describes the efficiency with which a discharge lamp attains steady-state luminous output from a cold start. The diagram shows a typical locus of the instantaneous values of light output of a discharge lamp, with increasing time from switch-on from a cold start. Area 'A' represents the mathematical product of light output and elapsed time during the lamp 'run-up' period (measured in percentage-minutes). Area 'B' represents the mathematical product of the steady-state light output and time (also measured in percentage-minutes) over the same time duration as that taken for the lamp to run-up, i.e. the same time duration as that applying for area 'A'.

The run-up efficiency is then calculated from:

$$\text{Run-up efficiency} = \frac{\text{Area A}}{\text{Area B}} \times 100\%$$

21.5.2.3 Discharge lamp types

Discharge lamp types include low pressure mercury (fluorescent), induction, high pressure mercury vapour, mercury-blended, metal halide, low pressure sodium and high pressure sodium.

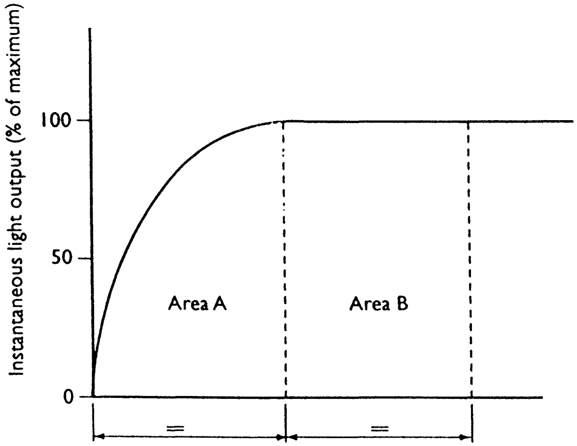


Figure 21.16 Concept of run-up efficiency

21.5.3 Mercury lamps

21.5.3.1 Low pressure mercury fluorescent lamps

Construction A typical mercury fluorescent tube consists of a glass tube, and up to 2400 mm long, filled with argon or krypton gas and containing a drop of liquid mercury. A diagram of the tube is shown in *Figure 21.17*. The interior surface of the tube is coated with a fluorescent powder, the phosphor, which converts the ultraviolet light produced by the discharge into visible light. At each end of the tube are electrodes which serve the dual purpose of cathode and anode, for generally these tubes are used on a.c. circuits.

The cathodes of a hot cathode fluorescent lamp consist of coiled-coil, triple-coiled or braided tungsten filaments, coated with a barium oxide thermionic emitter and held by nickel support wires. Cathode shields in the form of metal strips bent into an oval shape surround the cathodes in certain sizes of tube and are supported on a separate wire lead. These shields trap material given off by the cathodes during life and prevent black marks forming at the ends of the tube. The shields also reduce flicker which is sometimes noticeable at the ends of the tubes, and protect the more delicate cathodes by acting as anodes on alternate half periods. The bases of the electrode support wires are gripped in a glass pinch through which the lead wires pass, forming a vacuum-tight glass-to-metal seal.

The lead-in wires are welded to the pins of the bi-pin cap, which is itself sealed to the ends of the glass tube. Some tubes are still available with a BC cap, but the bi-pin cap is now British standard.

Principle An external control circuit is required, which causes a preheating current to flow through the electrodes. This causes electrons to be emitted by the emitter coating. Once these have been produced, the control circuit must apply an electric field across the length of the lamp to accelerate the electrons and strike the arc. Once struck, the arc must be stabilised by the control circuit.

Colliding electrons excite mercury atoms, and produce ultraviolet and visible radiation (about 60% of the energy consumed is converted to ultraviolet radiation). This radiation, when absorbed by the phosphor on the inside of the glass wall, is converted to visible light.

The colour and spectral composition of radiated light will depend upon the phosphors used. Lamps can be made with a 'white' appearance but with widely different efficacies or colour rendering properties.

Basic starter-switch circuit The basic starter-switch circuit is shown in *Figure 21.18*(*a*). The principle of operation is as follows:

(a) When the mains voltage is applied, a glow discharge is created across the bi-metal contacts inside the glow starter (enclosed in a small plastic canister). The contacts warm up and close, completing the starting circuit and allowing a current to flow from the 'L' terminal, through the current limiting inductor through the two tube cathode filaments and back to the 'N' mains terminal.

(b) Within a second or two, the cathode filaments are warm enough to emit electrons; a glow is seen from each end of the tube. At this stage, the starter-switch bi-metal contacts open (because the glow discharge, which caused them to heat and close, ceases when they touch, and they cool and open), and interrupt the preheat current flow. If an inductor (*choke*) ballast (coils of copper wire around a laminated iron core) is used the magnetic energy stored in the core collapses to produce a high-voltage pulse (600–1000 V) across the fluorescent tube sufficient to strike the arc and set up the electric discharge through the tube.

(c) Once the tube arc has been struck, the current through the tube gradually builds up. This means that the current through the inductor also increases. As this happens, the voltage across the inductor increases and the tube voltage falls. The inductor is so designed that when the tube and inductor current rise to a value determined by the inductor design setting, the circuit stabilises.

Electronic start circuit An improvement of the basic starter-switch circuit (*Figure 21.18*(*b*)) is the electronic start circuit. It is identical with the starter-switch circuits in all respects but one; the glow starter is replaced by an all electronic starter. In some cases it may be fitted into a conventional glow start canister, as a direct replacement, and in other more sophisticated luminaires it is a small encapsulated box.

The main advantages of this circuit are that it affords reliable starting, does not shorten tube life (a problem with

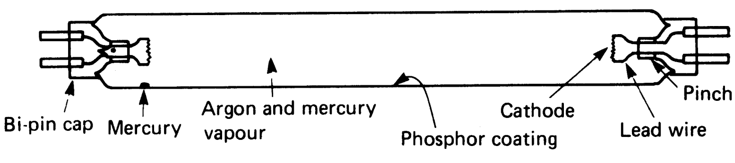


Figure 21.17 Low pressure mercury vapour fluorescent tube

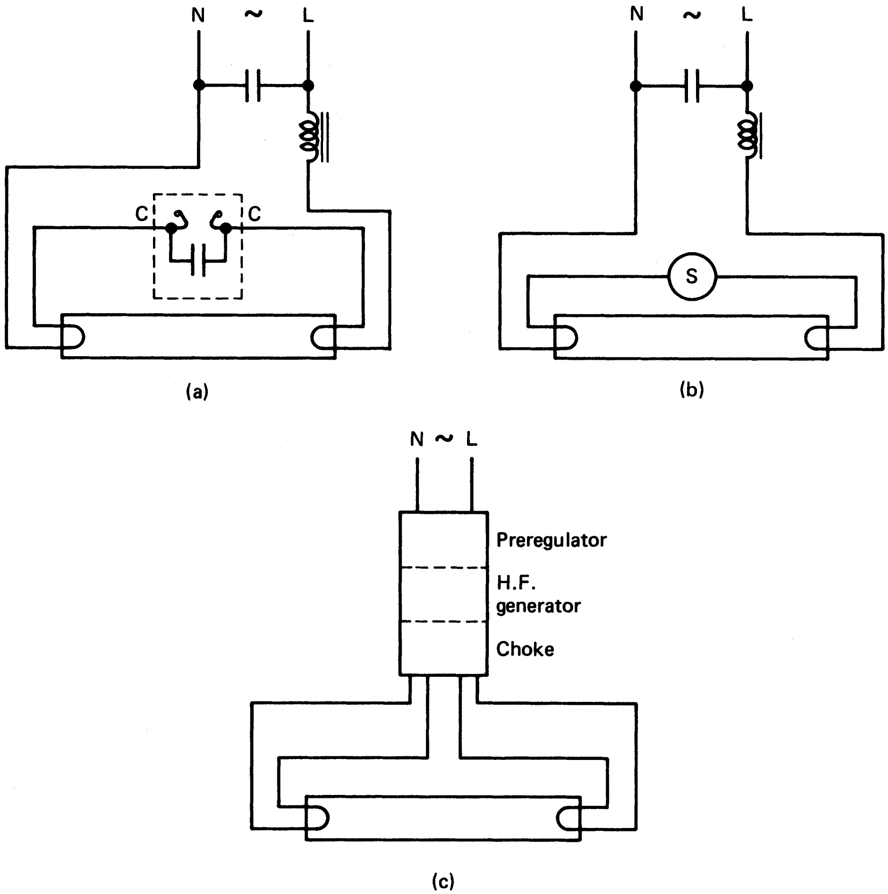


Figure 21.18 Starting methods: (a) glow starter; (b) electronic starter; (c) electronic control gear

glow switches at the end of their life) and will last the life of the luminaire without replacement. Some circuits also inhibit the start of faulty lamps after a number of unsuccessful attempts.

Low loss control gear The main power losses in the control gear are the result of heating and eddy current losses. In order to improve energy efficiency, thicker wire with lower resistance can be used in the ballast and the ballast can be made more cubic in shape. When this happens the power losses in the ballast are reduced but the ballast is more costly. Such ballasts are commonly referred to as 'low loss' and 'super low loss' according to their power dissipation. There is no definition of standard, low loss or super low loss; therefore comparisons must be based on measured losses.

It should be noted that, although they improve efficiency, such ballasts are more expensive, heavier and occupy a much larger and more awkward volume than conventional ballasts. This may cause problems by increasing the weight and bulk of the luminaires.

Electronic ballast Electronic ballasts offer several major advantages. They are lighter and replace several discrete components by one unit. They dramatically reduce the losses in the control gear, saving energy. Most designs provide a near-unity power factor, reducing the current drawn from the supply. The ballast also operates the lamp more efficiently, reducing the power losses within the lamp itself. Better starting will normally prolong lamp life. Better designs will operate over a wide range of supply voltage fluctuation.

Some designs can only be used with special lamps, but most will work with any standard lamp type from any manufacturer.

Electronic ballasts generate a high-frequency supply to the lamp (typically 30 kHz). At this frequency the normally bulky wire wound ballast can be made small and light. The ballast can be very efficient, but the generation of high frequency can give rise to conducted and radiated interference. Therefore, most circuits are in three parts. The first part of the circuit is designed to ensure that the supply form is not corrupted and that interference is not radiated. The second part of the circuit generates a high-frequency supply. The third part of the circuit uses the high-frequency supply through very small chokes to control the current and voltage to the lamp (*Figure 21.18*(*c*)).

Older 'hybrid' designs of electronic ballast use large iron core chokes in order to filter the supply waveform and prevent supply corruption. Such ballasts are normally heavier and bulkier than their conventional counterparts and do not offer the advantages of the fully electronic designs.

Not all ballasts generate true high frequency. Some simply chop the mains waveform at high frequency. Those which generate true high frequency have two other advantages. Firstly, at 30 kHz fluorescent lamps operate more efficiently. Typically the lamp efficiency improves by 5–10% according to type. Secondly, it has been discovered recently that the light fluctuation at 100 Hz which occurs with conventional mains operation of fluorescent lamps, although not visible, is detectable in the human visual system. It is suggested that some individuals suffer a higher incidence of reported headaches and eyestrain as a result of this invisible fluctuation. Higher frequency lighting has been shown to minimise this problem and, as a result, improve productivity.

Variable light output electronic ballasts Now that electronic ballasts are widely used it has become possible to build in extra circuitry at modest cost in order to permit the light output of the lamp to be controlled. These controllable ballasts will respond to simple control signals and vary the light output of the lamp.

A good design should be able to vary the light output from 1% to 100% of full output. This is normally achieved by an extra-low-voltage control circuit connected to a simple variable resistor. Several ballasts would normally be connected to one resistor to control the lighting in a room. Such a set up is less expensive than conventional dimming circuits.

In some designs an interface can be provided which will link the ballasts to the output of a conventional dimmer without adding load. In this way a conventional tungsten load can be controlled along with the fluorescent lighting. The dimming range is normally restricted to 10–100% because of the dimmer circuit.

Not all ballasts can provide full control. Some will only operate reliably with special lamps. Some ballasts can only control the light output from 25% to 100%. Although this may seem a large variation (4:1), because of the way in which we see, it is only perceived to be about a 2:1 change in brightness at full light output. This degree of control is of little use except for minor energy management where the lighting levels can be raised and lowered to top up daylight over a limited range.

Controllable electronic ballasts can be linked together and controlled in a number of different ways. The ballasts can be linked to a full lighting management system connected to local controls, programmable clocks, occupancy sensors and daylight photocells.

Fluorescent tube replacement It is not easy to determine the end of the useful life of a fluorescent tube. Although failure to start will eventually occur due to exhaustion of the oxide coating on the electrode filaments, it is normally possible to justify replacement of the tube before this stage. As most installations of fluorescent luminaires are designed to give a planned illuminance, random replacement of tubes at end of life will result in a non-uniform illuminance, which is uneconomic when related to energy costs and labour costs for replacement.

Figure 21.19 indicates typically the inherent deterioration of the illuminance from the luminaires of a fluorescent tube installation, and the gains that result from regular cleaning and lamp replacement. It is assumed that the use of the installation is 3000 h/year.

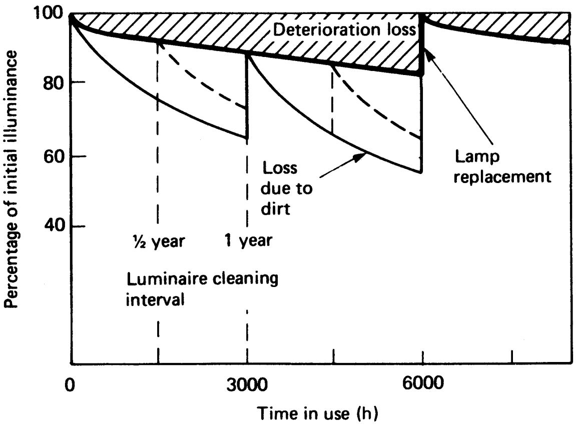


Figure 21.19 Effects of deterioration, cleaning and lamp replacement on the illuminance of a fluorescent-tube installation

Tube colours *Table 21.1* shows how the tube colours are graded in terms of lumen output efficacy and colour rendering quality. In choosing a tube colour a choice must be made between light output and colour, since tubes with high lumen output have only modest amounts of blue and red energy, whereas good colour rendering lamps have reduced yellow/green content and the tube lumen output is subsequently reduced.

A development in fluorescent tube phosphors is the multi-phosphor. Based on the principle that a mixture of red, green and blue light produces a white light, efficient red, green and blue phosphors are mixed in appropriate proportions to produce a white light when irradiated by ultraviolet light in a fluorescent tube. This produces a high-efficiency tube with good colour rendering for general applications. As the phosphors are costly, the tubes using them are more expensive and are frequently made in a diameter of 25 mm instead of 38 mm.

The characteristics of different fluorescent tubes are described and shown in *Table 21.1.*

Standard halophosphate lamps At one time these lamps were the best choice for general lighting because they combined acceptable colour with high efficiency. They have now been replaced by the superior multiphosphor lamps which give higher output, better colour and superior economy through life.

Special lamps These lamps have specific colour rendering characteristics.

Multiphosphor/triphosphor lamps These lamps are superior in colour quality and efficiency to standard halophosphate lamps. They therefore reduce installation costs because fewer luminaires are needed and also improve lighting quality. They are the primary choice for general lighting.

Deluxe multiphosphor lamps These lamps are superior in colour quality to normal multiphosphor/triphosphor lamps and are designed to replace the special lamps referred to previously, but with significantly better efficiency.

21.5.3.2 Compact fluorescent lamps

In recent years compact fluorescent lamps have been developed. For a given light output, compact fluorescent

Table 21.1 Fluorescent tube colours

Tube colour	*Light output relative to white* (%)	*Luminous efficacy** (lm . W^{-1})	*Colour rendering*	*Colour appearance*	*Application and remarks*
Standard halophosphate lamps					
White	100	45–65	Fair	Intermediate	A general-purpose tube which combines good lumen output with a white appearance. Warmer than daylight, but cooler than incandescence
Warm white	98	45–65	Fair	Warm	For general lighting where a warmer appearance than white is required. Incandescence effect, but without good red
Daylight cool white	94	45–65	Fair	Cool	For general lighting where a cooler appearance than white is required. Daylight effect, but lacking in red
Special lamps					
Northlight, colour matching	59	2–40	Excellent	Cool	For displays in lighting where a cool north skylight (winter light) effect is required, with normal red rendering. For colour matching
Artificial daylight	41	20–40	Excellent	Cool	Special tube with added ultraviolet to give a very close match to natural daylight. For colour matching cubicles
Natural	70	30–50	Good	Intermediate	For office and shop lighting to give a cool effect. Close to natural daylight but with a flattering red content
De luxe natural	49	15–25	Very good	Intermediate	For food and supermarket displays with meat or highly coloured merchandise. Combination of good blue and red rendering

*Based on total circuit power.

lamps have small dimensions compared with linear fluorescent lamps. This reduction in size is normally achieved by folding the discharge path.

These lamps have two major application advantages. Firstly, because they are available in similar sizes and light outputs to conventional filament lamps they can be used as replacements, either in existing luminaires or in new designs. They will use about 25% of the power of the tungsten lamp equivalent of similar light output and will typically last 10 times longer. Secondly, compact fluorescent lamps can be made smaller than their linear fluorescent counterparts. This means that smaller more attractive luminaires can be designed with similar light output to conventional fluorescent lamps, but without the need to be long and awkward in shape.

Therefore compact fluorescent lamps are widely used either for applications where, previously, tungsten lamps would have been used, such as decorative lighting in hotels, or for applications where fluorescent lamps would have been used but a more acceptable luminaire style or shape can be designed (e.g. a 500 mm luminaire can replace a 1200 mm × 300 mm luminaire).

Theoretically, there are many good reasons why tungsten lighting in the home should be replaced by compact fluorescent lighting. Despite the superior life and economy (the major cost of a lamp is the electricity it consumes), these lamps have not been widely used in the home. However, they are extensively used in industrial and commercial situations.

Compact fluorescent lamps can be divided into two broad categories: replacements for GLS lamps and light sources for new luminaires.

Replacements for GLS lamps Some lamps have integral control gear. They can be used as direct plug-in energy saving replacements for conventional lamps. The only factor to note is that some types are much heavier and rather more bulky than the lamps that they replace. Some lamps have integral electronic control gear. The main disadvantage of this type of approach is that when the lamp is thrown away then the expensive integral control gear is thrown away with it. An alternative approach is to provide adapters which can be plugged in to lamp sockets and convert them to compact fluorescent use.

Light sources for new luminaires Some lamps were developed to make possible the design of compact luminaires with all the advantages of fluorescent lamps. The use of these lamps permits plastic and other materials to be used close to the lamps. This increases the scope for attractive designs using novel materials. There is a wide range of lamps with light output suitable as substitutes for conventional lamps of 100 W and below. However, in commercial and industrial lighting, greater light outputs are required.

21.5.3.3 Cold cathode lamps

By comparison with conventional hot-cathode fluorescent lamps, cold-cathode lamps, which are often used in advertising signs, rely upon a relatively high lamp voltage for establishment of the arc. Electrical isolation of such lamps is typically achieved by means of a 'fireman's switch' usually located on the frontage of premises.

The electrodes used for cold-cathode lamps are typically plain nickel or iron cylinders whose size is substantial in order to limit the current density at their surface to an acceptably low value.

Lamp life values for such lamps are much higher than for hot-cathode fluorescent lamps.

21.5.3.4 Induction lamps

The induction lamp is also referred to as the 'electrodeless lamp'. It relies upon both magnetic and fluorescent principles for its operation. The constructional features of the lamp are shown in *Figure 21.20*.

Energy transfer using magnetism (following the electrical transformer principle) is employed with the low pressure mercury filling in the lamp acting as a secondary coil of the transformer. A high frequency alternating electrical current in the primary winding is supplied from an external source. The current induced in the mercury vapour gives rise to emission of ultraviolet radiation in a similar manner to that in a conventional fluorescent lamp and the phosphor coating on the inside of the lamp envelope converts this UV radiation into visible light. The lamp life is typically 60 000 h, which makes the use of such lamps in relatively inaccessible areas particularly beneficial.

21.5.3.5 High-pressure mercury lamps

Construction The lamp consists of a quartz glass (pure fused silica) arc tube, enclosed within a borosilicate outer envelope. The envelope can be tubular, but is normally ellipsoidal to ensure an even outer-envelope temperature. (Ellipsoidal outer envelopes operate at about 550 K, compared with 800 K for the hottest parts of tubular envelopes.)

The arc tube contains a controlled quantity of mercury sufficient to produce the desired pressure at operating temperature.

Electrodes, in the form of helices of tungsten wire, about a tungsten or molybdenum shank, are fitted at opposite ends of the arc tube. Emissive material is coated onto the electrodes (or may be held inside in pellet form). To assist in starting the lamp, an auxiliary electrode is mounted in close proximity to one of the main electrodes and is connected to the other electrode via a high resistance (10–30 kΩ). The electrodes are sealed into the quartz glass arc tube by means of molybdenum foil (as with tungsten halogen lamps).

The outer envelope (*Figure 21.21*) is filled to a pressure of 0.25–0.65 atm with either nitrogen or a nitrogen–argon mixture.

High-pressure mercury fluorescent lamp The quartz glass arc tube transmits ultraviolet light and enables a phosphor coating to be used on the inside of the outer envelope. The phosphor improves the colour rendering properties and luminous efficacy. Phosphors have upper limits of operating temperature and the use of ellipsoidal bulbs ensures the minimum size of envelopes. Improvements in phosphors to give operation at higher temperatures have resulted in smaller envelope sizes.

High pressure mercury vapour lamps (see *Figure 21.22*) have efficacies of about 58 lm . W^{-1} with acceptable colour rendering for factories, storage areas, offices, etc.

High-pressure mercury fluorescent reflector lamp This lamp (*Figure 21.23*) is identical with the high pressure mercury vapour lamp, except that the outer envelope is shaped to form a reflector. Titanium dioxide is deposited inside the conical surface of the outer envelope: this reflects about 95%

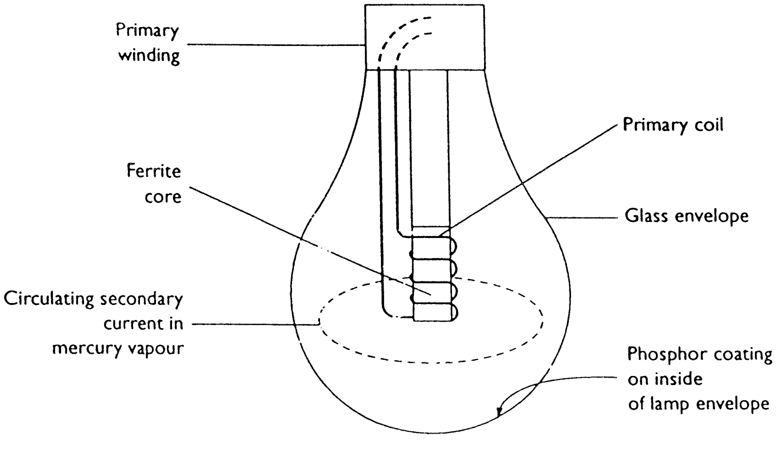

Primary winding
Primary coil
Ferrite core
Glass envelope
Circulating secondary current in mercury vapour
Phosphor coating on inside of lamp envelope

Figure 21.20 Induction lamp

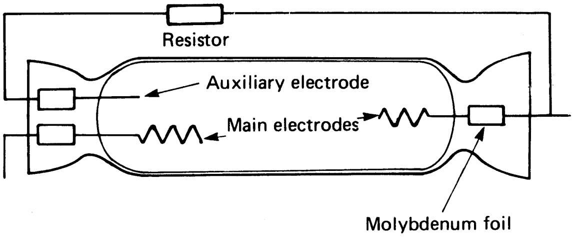


Figure 21.21 Arc tube construction of a high pressure mercury lamp

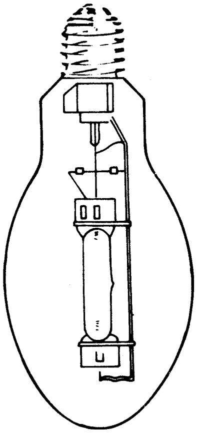


Figure 21.22 A high pressure mercury lamp with phosphor coating

of all the light in a diffuse manner. The phosphor is applied over this coating. The front face is left uncoated (although it may be etched or have a diffusing coating of silicon dioxide) and about 90% of the light is emitted through this opening. The lamp is less efficient but has a directional light output, suitable for installations where the luminaires have to be mounted high up (as in storage areas, hangars, etc.), and where dirty conditions obtain.

High-pressure mercury-blended lamp By including a filament within the outer envelope the need for special control gear can be eliminated, as the filament can act as a ballast. The efficacy of the lamp (12–25 lm . W^{-1}) is poor compared to a high pressure mercury vapour lamp and inductor, but the lamp has two important advantages: (1) it can be fitted into any standard lampholder as a direct replacement for a tungsten lamp of the same rating, and will produce more ultraviolet and visible light; and (2) unlike other mercury lamps, it will emit some light from the filament immediately upon switching on.

High-pressure mercury-metal halide lamps The efficacy of high pressure mercury lamps is lowered because regions of the lamp are at potentials too low for the excitation of mercury. By adding suitable metals with lower excitation potentials to the arc tube, it is possible to increase the light output and improve colour rendering. The only suitable metals are highly reactive. These would damage the quartz glass arc tube and seals. By adding the metals (sodium, thallium, gallium, scandium and others) in the form of halides (usually iodides), these problems can be eliminated.

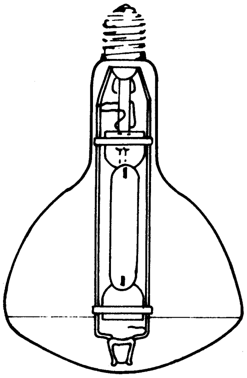


Figure 21.23 A high pressure mercury lamp with phosphor coating and integral reflector

The halides dissociate in the arc itself, but recombine at the arc tube wall.

The metal halide lamps are better than their counterparts, in both colour rendering and efficacy (85 lm . W^{-1}). *Figure 21.24* shows a high pressure mercury-metal halide lamp.

21.5.4 Sodium Lamps

21.5.4.1 Low-pressure sodium lamps

The low-pressure sodium lamp is characterised by its monochromatic yellow light, which consists of two radiation lines (resonant doublet) at 589.0 and 589.6 nm. The lines are close to the maximum spectral sensitivity of the eye (555 nm) and the lamp is therefore efficient. Efficacies of over 150 lm . W^{-1} are typical.

The lamp has poor colour rendering properties. Because the light is monochromatic, it is restricted to applications where the colour of the source and colour discrimination can be sacrificed for high efficacy. For example, the lamp is used for floodlighting and street lighting.

Construction and operation The lamp consists of a long arc tube of glass construction, known as 'ply tubing', filled with low-pressure gas (usually neon + 1% argon). The lamp also contains a small quantity of metallic sodium (solid at room temperature), which provides a pressure of about 0.66 Pa in the lamp when operating. Precautions must be taken to prevent the sodium from attacking the lamp seals. At each end of the arc tube is a tungsten electrode of coiled-coil,

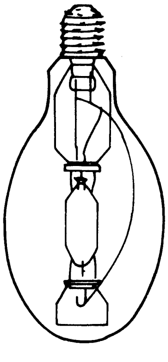


Figure 21.24 A high pressure metal halide lamp

triple-coil or braided construction (similar to those in tubular fluorescent lamps) with an emissive coating (barium oxide or similar material).

The temperature of the arc tube must be maintained at about 270°C in order to successfully vaporise the correct amount of sodium to give a vapour pressure of about 0.66 Pa. It is therefore essential to avoid excessive heat loss if this temperature is to be maintained. In early lamp designs a Dewar flask was placed around the arc tube.

Modern lamps have an outer glass envelope enclosing the arc tube, and the space between the two is fully evacuated. In addition, the outer envelope is coated with an infrared reflecting film (bismuth oxide, tin oxide or gold) which transmits light but reflects heat back onto the lamp.

The requirements for efficient operation of a sodium lamp are: (1) the arc voltage gradient must be low (long arc tubes for a given power), and (2) the current density must be low (arc tubes must have a large diameter). Sodium vapour readily absorbs light at the resonant doublet wavelengths, and therefore light generated in the centre of the arc is reabsorbed by the vapour; thus, only the outer surface of the arc emits light. This conflicts with the need for a large-diameter arc tube, and demands a compromise.

Sodium, especially hot sodium vapour, is highly reactive and attacks any glass with more than a small proportion of silica in it (i.e. almost all normal glasses and quartz glass). Special glasses have been developed with low silica content which resist the attack of sodium vapour, but they are expensive, are difficult to work and are attacked by moisture. Hence, ordinary soda-lime glass tubing has a coating of this resistant glass flashed onto its inside surface, and the resultant cheap, easily worked material is known as ply tubing.

Unless checked, sodium vapour readily migrates along the lamp to the cooler parts. To prevent this, small protrusions are moulded into the arc tube. They project out from the arc tube and are therefore slightly cooler than the surrounding wall. They act as reservoirs for the sodium metal and help to maintain the correct vapour pressure at all points along the lamp.

A long arc tube is folded into a tight U shape (*Figure 21.25*). Mutual heating is provided by the two arms of the arc tube, but also each arm absorbs any light from the other arm which may strike it. The two effects are almost self-cancelling, but do produce a slight improvement in efficacy. The resultant lamp is fairly compact, with the advantage that all of the lamp connections are at one end. A single-bayonet lamp cap can therefore be used. The lamp efficacy is around 100 lm . W^{-1}.

When a sodium lamp is first switched on, the sodium is all present as solid metal. An arc discharge cannot therefore occur unless the sodium is first vaporised. For this reason a mixture of neon and argon is used in the lamp. The initial discharge occurs through this gas; the sodium metal is vaporised by the heat from this discharge, and slowly takes over.

The mixture of neon and argon (1%), known as a 'Penning mixture', reduces the starting voltage required. The energy required to *excite* neon is slightly higher than that to *ionise* argon. Electrons passing through the gas collide with the main gas (neon) and excite the atoms, which may collide with argon atoms giving up their energy by ionising the argon and producing an extra electron. This mixture reduces the starting voltage by typically 30–50%.

All low pressure sodium lamps when first switched on produce a distinctive red neon glow. Should sodium migrate from any part of the lamp, that part will also exhibit the red neon glow.

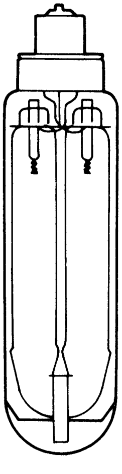


Figure 21.25 A low pressure sodium lamp

The effect of the Penning mixture is to make the starting voltage almost independent of ambient temperature. A low pressure sodium lamp can therefore be restruck when hot within about 1 min.

Although electrodes are fitted at each end of the arc tube, it is not normal to provide them with a heating current; and the two ends of an electrode are connected together.

21.5.4.2 High pressure sodium lamps

If the sodium vapour pressure in a low pressure discharge tube is raised by a factor of 10^5, the characteristic radiation is absorbed and broadened, greatly improving the colour rendering. However, hot sodium vapour is highly reactive, destroying or discolouring conventional arc tube materials; further, to achieve a high vapour pressure the coolest region of the arc tube must have a temperature of 750°C. These phenomena, known in the 1950s, were not exploited until the development of a translucent ceramic material—isostatically pressed doped alumina—capable of operating at temperatures up to 1500°C and of withstanding hot sodium vapour. The difficult process of sealing electrodes into the ends of the arc tube has also been solved. Typical methods are (a) brazing a niobium cap to the alumina tube, and (b) using hermetically sealed sintered ceramic plugs holding the electrodes.

Most high pressure sodium lamps have an arc tube containing metallic sodium doped with mercury and argon or xenon. Radiation of light is predominantly from the sodium. The mercury vapour adjusts the electrical characteristics, and acts to reduce thermal conductivity and power loss from the arc. Argon or xenon aids starting. The arc tube is sealed into an evacuated outer jacket to minimise power loss and to inhibit oxidation of the end caps, lead wires and sealing medium. Typical are tube operating temperatures are 700–1500°C, and efficacies of 100–200 lm . W^{-1} are achieved.

Starting is effected by high-voltage pulses (2–4.5 kV) to ionise the xenon or argon gas. The ionisation heats the lamp and the sodium vapour discharge takes over. The mercury vapour does not ionise, as its ionisation potential is higher than that of sodium, but its effect is to increase the lamp impedance, raising the arc voltage from about 55 V to 150–200 V. Electrodes are not heated during lamp operation. A hot lamp, after extinction, will not restart until cooled, unless 'hot-restart' ignitors (giving, e.g. 9 kV pulses) are provided.

The conventional high pressure sodium lamp (*Figure 21.26*(*a*)) has an outer envelope, which is elliptical and coated with a diffusing material. Tubular and double-ended versions are available as shown in *Figure 21.26*(*b*) and *Figure 21.26*(*c*).

It is interesting to note that, although the low pressure sodium lamp is more efficient than the high-pressure version, the large dimensions of the former make accurate optical control difficult. Furthermore, the luminaires are large, cumbersome and expensive. High pressure sodium lamp luminaires are therefore the better choice for many applications, and only in street lighting does low pressure sodium find a major application. High pressure sodium lamps are employed for street lighting, floodlighting and industrial and commercial interiors.

By increasing the xenon pressure in a high pressure sodium lamp and making some other changes, the output of the lamp can be improved without any loss of life or colour. The lamp efficacy improves by about 10–20% and the light output depreciation of the lamp is improved. The only penalty is that the lamp needs a much higher starting voltage (i.e. a different ignitor is needed).

If the internal pressure of a high pressure sodium lamp is increased then the colour quality and colour temperature of the light is improved at the expense of life and efficiency (life reduced by about 50% and efficacy by about 25%). These lamps are known as 'deluxe high pressure sodium lamps'. The colour rendering index is normally improved from 21 (class 4) to about 65–70 (class 2). This makes the lamp acceptable for merchandising applications and other areas where good colour is an important consideration. The lamps are widely used for lighting offices (via up-lighters) and for commercial lighting. The lamp is warm in appearance and for some applications this may be unacceptable. A recent development is the low wattage 'White' high pressure sodium lamp with very high colour rendering which has been designed for display lighting purposes. These lamps need special electronic control gear.

21.5.5 Control gear

The term is used affectionately to describe the equipment that is necessary for the safe and efficient operation of discharge lamps. The function of the control gear is essentially twofold: it assists in providing a high voltage pulse to enable the arc to be established within the discharge tube in the lamp and once the arc has been established the control gear takes on the role of a current limiting device in order to limit the current flowing through the lamp. Discharge lamps exhibit a negative resistance characteristic.

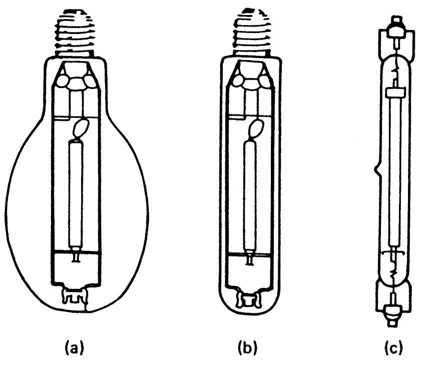


Figure 21.26 High pressure sodium lamps: (a) SON lamp; (b) SON/T tubular lamp; (c) SON/TD double-ended linear lamp

It has to be appreciated that heat is produced during normal operation of lamps and control gear. Nevertheless many items of control gear have recommended maximum temperatures and it is essential that such values be not exceeded.

Individual items of electrical and/or electronic equipment used include ballasts, transformers, ignitors and power factor correction capacitors. Some items of control gear will consume power from the supply and the resulting electrical energy costs will have to be borne by the consumer. It is therefore beneficial to be aware of the power ratings of control gear when considering the economics of lighting.

21.5.6 Electroluminescent devices

Electroluminescence is the emission of light from a semiconductor under the action of an electric field. The process involves heat only as a by-product of the mechanism, which is essentially a 'cold' one. The phenomenon is of commercial interest because of its relatively highly efficient production of visible light.

Characteristics Luminescence decays with time exponentially, so that is convenient to quote the half-life of an electroluminescent source. The half-life varies from hundreds to millions of hours, depending on the type and purity of the semiconductor materials used. The sources emit light in comparatively narrow spectral bands, producing colour without the use of filters. Sources can be of almost any size, down to areas of less than 0.05 mm^2. The surface brightness is, however, comparatively low.

In an electroluminescent material two processes must take place. First, an adequate supply of electrons must be available in the conduction band, made possible by using an electric field to raise the energy level of electrons in the valence band. Second, the electrons must give up their energy in the form of photons and so return to the valence band. The recombination process is dependent on the 'forbidden gap', the wavelength λςof the emitted radiation being defined by the energy jump E through the relation λς$=hc/E$, where h is Planck's constant and c is the free-space velocity of electromagnetic waves. Consequently, materials with a bandgap E of 1.65–3.2 eV are capable of producing visible light by electroluminescence, provided that the electron return from the conduction to the valence band is not made in two or more stages by reason of the presence of lattice impurities; but if this is the case, luminescence can be obtained with materials having an intermediate gap energy.

Materials The requirements for an electroluminescent material emitting in the visible spectrum are, generally, that it should have a bandgap of at least 2 eV, be susceptible to both p- and n-type 'doping', and should have either a direct bandgap or an activator system permitting 'steps'. Typical semiconductor materials are ZnS, GaP, GaAs and SiC.

The preparation of luminescent panels based on ZnS involves baking pure zinc sulphide with various dopants such as manganese, copper and chlorine. The sintered mass is ground to a particle size of the order of 10 μm, washed and dried, and mixed with a suitable resin. The suspension is coated on to a glass substrate supporting a very thin layer of gold or SnO, to form one electrode. The other electrode, an evaporated layer of aluminium on the rear surface of the

Table 21.2 Lamp characteristics

Lamp type	*ILCOS lamp coding*	*Typical lamp life* (hours)
Tungsten filament	I	1000–2000
Tungsten halogen	HS	2000–4000
Low pressure	FD (tubular)	6000–15 000
Mercury (fluorescent)	FS (compact)	8000–10 000
High pressure mercury	QE	14 000–25 000
Metal halide	M	6000–13 000
Low pressure sodium	LS	11 000–22 000
High pressure sodium	S	12 000–26 000
Induction	XF	60 000

cell, also acts as a reflector. The small cells are produced from GaP crystals, with areas selected for fabricating electroluminescent diodes. They are cut into 0.5 mm squares, and a p–n junction is formed by alloying a tin sphere into the GaP, which is doped with zinc to produce the junction. The diode structure is completed by alloying a Au–Zn wire into the GaP to make an ohmic contact, and then connecting a wire to the tin sphere.

21.5.7 Lamp life

Table 21.2 gives details of lamp life values.

21.6 Lighting design

21.6.1 Objectives and Criteria

To design a lighting scheme the basic objectives must be first established. What sort of tasks will be carried out in the area to be lit? What mood needs to be created? What type of lighting will create a comfortable, pleasant environment? The objectives having been established, they have to be expressed as a series of lighting criteria. For example, what level of illuminance is required? How much glare is acceptable?

The designer then plans a scheme that will best meet the criteria by selecting the appropriate luminaires and considering the practical problems.

21.6.1.1 CIBSE code for Interior Lighting

The Chartered Institution of Building Services Engineers publishes a code of recommendations for the lighting of buildings. It puts forward ideas and methods representing good modern practice and is concerned with both quantity and quality of light.

The CIBSE code for Interior Lighting deals with: how building design affects lighting; lighting criteria; lighting and energy consumption; design methods; lighting equipment; and methods of maintenance. In addition, there is a schedule giving specific recommendations for a wide variety of areas such as assembly areas, factories, foundries, schools, hospitals, shops and offices.

For each entry a maintained illuminance and position of measurement is quoted. The amount of discomfort glare that can be tolerated is given by a limiting glare index. A list of suitable lamp types and their colour appearance is given together with notes on special problems that may be encountered.

The recommended maintained illuminance—in other words, the recommended illuminance that should be provided for a particular application—is a useful guide for the designer. But it is only a recommendation: in some circumstances it should be increased. For example, if errors could have serious consequences in terms of cost or danger, or if unusually low reflectances or contrast are present in the particular task, or if tasks are carried out in windowless interiors where the recommended standard service illuminance is less than 50 lx, the illuminance level should be increased. On the other hand, there are circumstances (e.g. when the duration of the task is unusually short) when the designer would use his judgement to reduce the standard service illuminance. *Table 21.3* gives typical recommended maintained illuminance values.

21.6.1.2 Uniformity

In addition to providing the correct illuminance, the uniformity of the lighting level is also important. The uniformity is expressed as the ratio between the minimum and the average illuminance over the working area. It should not be less than 0.8 in areas where the tasks are performed.

21.6.1.3 Reflectance

Reflectances should also be considered. The effective reflectances of walls in a room should be between 0.3 and 0.7; the ceiling should have a reflectance of 0.6 or greater; and the floor should have a reflectance of between 0.2 and 0.3 (see *Figure 21.27*).

21.6.1.4 Illuminance ratio

The ratio between the illuminance on the wall and that on the task, and between the ceiling and that on the task, is also important. The wall illuminance should be between 0.5 and 0.8, and the ceiling illuminance should be between 0.3 and 0.9, of the task illuminance (see *Figure 21.27*).

For normal working environments, if the reflectances and the ratios between wall illuminance and the task, and between ceiling illuminance and the task, are outside the recommended levels, they will be unacceptable. However, there are exceptions if a particular mood or atmosphere is being created.

21.6.1.5 Modelling and shadow

The direction of light and the size of the luminaires in an interior affect highlights and shadows, influencing the appreciation of shape and texture. The term 'modelling' is used to describe how light reveals solid forms. It may be harsh (the contrast may be excessive and produce deep shadows) or flat (the light provides low contrast with little shadowing). Either extreme occurring in the lighting of a general working area makes vision difficult or unpleasant. Therefore, the aim of good lighting design must be to produce a suitable compromise between the two, although it may not be the same for different applications. Texture and surface details are best revealed by light with fairly strong directional characteristics. Reflections of lamps on polished surfaces may veil what one needs to see, reducing contrast and handicapping one's vision.

Table 21.3 Typical recommended maintained illuminance values

Workplace	*Typical recommended maintained illuminance values* (lux)	
Engineering workshops	rough e.g. bench work	300–400
	detailed work	700–800
	inspection	500–2000
Clothing and textile	preparation	300–400
	cutting	750–850
	inspection	1500
Construction sites	site roadways	5–15
	general areas	20–25
	crane loading	100–150
Electronics equipment manufacture	component assembly	300–1500
	printed circuit board work	500–750
	inspection	1000–1500
Food and drink	abattoirs	500–750
	bakeries	300–500
	breweries	300–750
	dairies	300–500
	flour mills	300–500
Metal working	iron and steel mills	300–500
	foundries	300–500
	inspection	500–1500
Furniture and timber	sawmills	300–750
	workshops	300–750
	furniture manufacture	300–750
	upholstery	500–1500
Glass production	general production	300–750
	inspection	1000
Health care premises	general wards	150–250 (day) 3–5 (night)
	maternity wards	200–1000
	laboratories	300–750
	operating theatres	500–50 000
Paper and printing	paper mills	300–750
	printing works	300–750
	inspection	1000

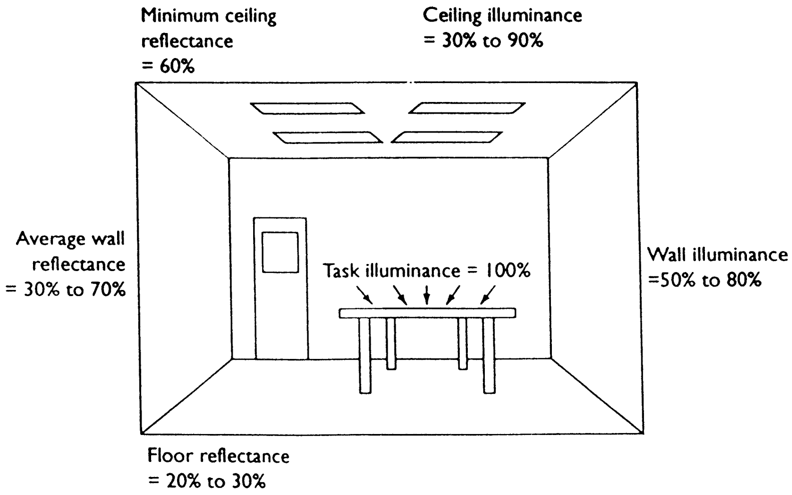

Minimum ceiling reflectance = 60%
Ceiling illuminance = 30% to 90%
Average wall reflectance = 30% to 70%
Task illuminance = 100%
Wall illuminance =50% to 80%
Floor reflectance = 20% to 30%

Figure 21.27 Typical room surface illuminance ratios. Illuminance values quoted are relative to a task illuminance of 100 per cent

21.6.1.6 Glare

The visibility of a working task is influenced to a large extent by any sources of glare within the visual field. One definition of glare is that it is any excessive variation in luminance within the visual field. Glare can be conveniently divided into two main groups i.e. disability glare and discomfort glare.

Glare can be thought of as 'direct' when it occurs as a consequence of bright sources directly in the line of vision, or alternatively as 'reflected' when light is reflected onto surfaces that have high reflectance values. Factors involved in the production of direct glare include the luminance of the light source and the location of the light source.

A form of glare that will disable an individual from carrying out a particular visual task is referred to as disability glare. An everyday example of disability glare occurs when an individual looks at the headlights of a stationary vehicle during darkness. Under these circumstances it is impossible to discern the scene at the sides of the vehicle immediately behind the headlights. The glare disables the individual from carrying out the visual task. The magnitude of the disabling effect experienced with disability glare is unlikely to occur with discomfort glare. An individual will experience a feeling of discomfort when the exposure time is prolonged. Disability glare and discomfort glare are the major forms of glare. Discomfort glare is more prevalent in interiors and often occurs as a consequence of either a badly designed lighting installation or a change of use of the interior from that which it was originally intended. In addition it is possible to assign numerical values to the degree of discomfort glare prevailing in a given interior. It is therefore possible, at the design stage, to eliminate any such adverse effects.

A method used for calculating the 'limiting glare index' (LGI) is shown in detail in the CIBSE Code for Interior Lighting. The method is a 'step-by-step' process where an initial and uncorrected glare index value is obtained from published tables. In these tables the major dimensions of the room interior (length and width) are given in multiples of the mounting height H which is taken as the distance between the horizontal lines passing through the eye level of a seated observer and the centre line through the luminaire(s). The eye level of a seated observer is taken, initially, as 1.2 m above floor level.

A second stage to the calculation involves applying two correction factors to the initial glare index value obtained. These factors involve variations in (a) the luminous flux of the luminaire and (b) any variation in mounting height from the normal seated eye level of 1.2 m.

Once the 'final glare index' value is calculated it is compared with reference values of limiting glare index (LGI), which are published in references e.g. CIBSE Code for Interior Lighting. Should the final glare index value calculated be lower than the LGI value the development of discomfort glare is unlikely.

It is also possible to categorise glare as either 'direct' or 'reflected'. Direct glare occurs when the origin is bright sources. Reflected glare occurs when light is reflected from specular or mirror-like surfaces. When considering glare it is essential to consider the luminance of the source, the position of the source, the luminance distribution and the time duration of exposure. The maximum value of luminance, which is tolerable by the eye as a result of direct viewing, is approximately 7500 $\mathrm{cd.m^{-2}}$.

The visual comfort of an individual is influenced by the distribution of luminance across the immediate field of vision and furthermore by the luminance of the environment when seen by individuals who glance away from their work.

Ideally the work surrounds should be less bright than the work itself. Optimum visual comfort occurs when the work is slightly brighter than the near surround. This in turn should be slightly brighter than the far surround.

The ideal ratio for distribution of luminance across a task is usually taken as 10 : 3 : 1. It is possible to obtain a 'trade-off' in luminance values, away from a visual target by using different materials. *Figure 21.28* shows the ideal luminance distribution across a visual task.

The luminance contrast of a visual task will be influenced by, inter alia, the reflectance values of the task and furthermore upon the manner in which the task is lit. Task materials with matt finishes will reflect the incident light equally in all directions and it follows that the direction of the incident light falling on the task material is unimportant. Conversely if the task material has a specular finish then the direction of the incident light is important.

When the image of a source of high luminance, e.g. a luminaire or the sun, is reflected from a surface being viewed by an individual then veiling reflections will be created. An example of this is experienced when looking at a display screen. If the geometry of the individual, screen and luminaire is not controlled, an out-of-focus image of the luminaire will be seen in the screen. This will throw a veil of light on the front of the screen, hence the term veiling reflections, and some of the text on the screen will become illegible, see *Figure 21.29*.

Downlighters with a strong downward component of light distribution will produce a significant loss of task visibility due to veiling reflections. A lighting installation should ideally provide a minimum directional component immediately above the task itself. Light should preferably reach the task from wider angles, which will reduce glossy reflections seen by the individual in the visual task.

21.6.1.7 Colour rendering of lamps

Generally the apparent colour of objects seen under an artificial lighting installation is of some importance, sometimes it is vital, as for instance in the buying and selling of foodstuffs, the preparation of paints and dyes and the matching of silks, cottons, etc., to fabrics and to one another.

A red article looks red because it reflects red light more strongly than other colours of light, but it can only do so if there is some red light present for it to reflect. Similarly, a green article can only look green if there is some green light present, and so on. It follows that if all colours are to be seen well, the light must contain a mixture of all possible colours of light of roughly equal strength. Such a light will look white, or nearly so. Unfortunately, it is possible for a

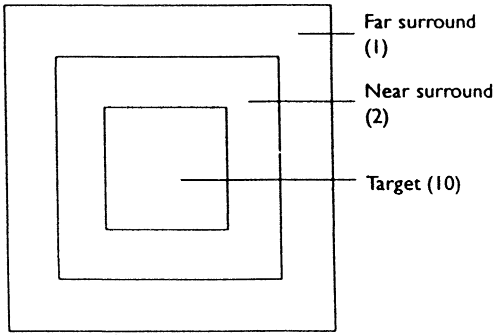


Figure 21.28 Ideal luminance distribution across a visual task

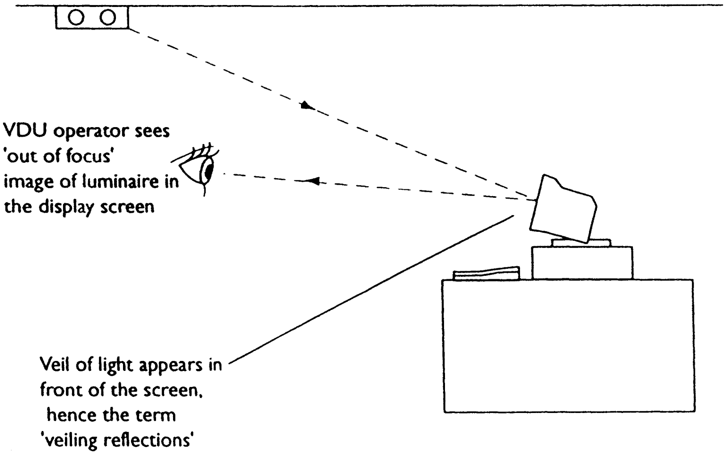

VDU operator sees 'out of focus' image of luminaire in the display screen
Veil of light appears in front of the screen, hence the term 'veiling reflections'

Figure 21.29 Veiling reflections

light to look white even though lacking some of the possible colours (e.g. the light of an ordinary mercury lamp can in certain circumstances look fairly white, though it lacks more colours than it contains, and the poor colour rendering of this type of lamp is well known).

Filament lamps of all types emit all possible colours of light, but in far greater strength at the red end of the spectrum than at the blue end. Thus, red and yellow objects appear in stronger colour than in natural daylight, whereas blues are weak and somewhat muddy in appearance. This, however, is a distortion familiar to most people, who have learned to make allowance for it in choosing decoration materials, etc. The light of a 'deluxe warm white' fluorescent lamp is very similar to that to a filament lamp.

For accurate colour judgement a light similar to north-sky daylight is necessary. An 'artificial daylight' fluorescent lamp has been developed which meets the requirements of BS 950:1967. The 'colour-matching' or 'north-light' type of fluorescent lamp gives a close approximation to north-sky light and is found entirely satisfactory in many industries critically concerned with colour, but this very white light is perhaps too 'cool' in appearance to be used alone in interiors of a more domestic nature, where a 'warm' light is traditional. The other colours of fluorescent lamps are mainly of higher efficiency than the previously-named two types, and the choice will lie between them according to the relative importance of their efficiency and their particular colour rendering properties.

21.6.1.8 Display screens

When a display screen equipment operator is viewing a screen, other parts of the interior will be within their visual field and this will lead to problems with adaptation. Light scattered in the eye will reduce the contrast of the image subsequently formed on the retina. The effect of this is to produce impaired vision. In addition if the display screen operator momentarily glances away from the screen, transient changes in adaptation will occur and again vision can be impaired.

There are several options available to the designer in an attempt to produce optimum lighting conditions. The most simple involves a repositioning of one or more of the light sources, the screen and the operator.

It will be evident that any re-positioning of the light source is the least practical since in many cases luminaires are ceiling-mounted in fixed locations. There are however much simpler remedies, for example the ergonomics of the operator is highly significant.

In interiors where the installation of uplighters is inappropriate then downlighters can be used. In such situations restrictions are placed upon the luminous output characteristics of the luminaires. In an attempt to avoid high-luminance reflections appearing on display screens it is important to use luminaires with an appropriate luminous intensity distribution, which subsequently limits the luminance seen by screen operators.

If the relationship between the terminal and operator could be established then it would be possible to determine the appropriate geometry, which in turn would allow calculation of luminaire characteristics necessary to prevent direct vision of the light sources. However such information is seldom available due to the wide range of tasks and screen types in use.

The work activities of those who use display screen equipment are covered by both EC and UK legislation, which specifies minimum standards for the visual environment, including the lighting conditions. In order to satisfy the legislation, display screen equipment tasks must be classified in accordance with the conditions specified therein. The tasks can influence the magnitude and severity of reflections likely to be encountered on the screen. The three categories relative to areas where display screen equipment is used are designated Category 1, Category 2 and Category 3. Luminaires for use in each of the three areas specified are known as Category 1, Category 2 and Category 3 luminaires. For each of these luminaire categories the luminance above a predetermined angle, referred to as the critical angle, is limited to 200 candelas per metre^2. The values of critical angles referred to, for categories 1, 2 and 3 luminaires, are 55°, 65° and 75° respectively. *Figure 21.30* shows the geometry of the critical angles described.

Category 1 luminaires are used typically where there is a high number of screens, in an area where extensive use is likely and where errors occurring as a consequence of misreading the data on the screen are unacceptable. Category 1 luminaires are typically used in air traffic control rooms and emergency services control rooms. Category 2 luminaires tend to be used far more than other Category

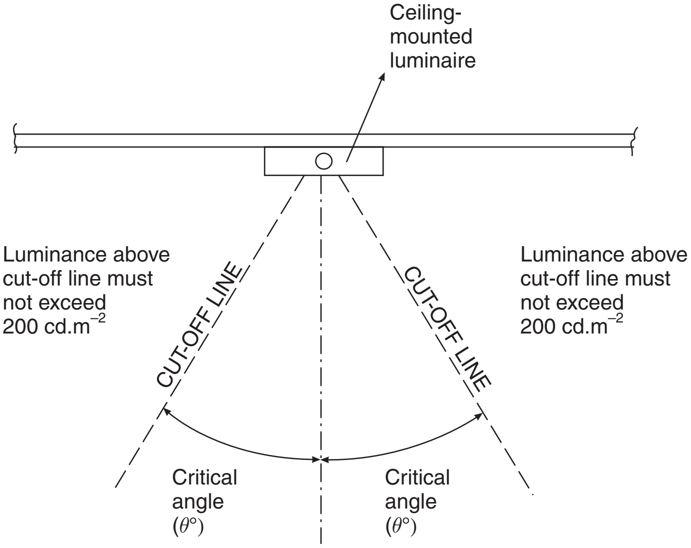

Ceiling-mounted luminaire
Luminance above cut-off line must not exceed 200 cd.m^{-2}
CUT-OFF LINE
CUT-OFF LINE
Luminance above cut-off line must not exceed 200 cd.m^{-2}
Critical angle ($\theta°$)
Critical angle ($\theta°$)

Figure 21.30 Critical angles for downlighter luminaires used in VDU areas

luminaires and their use is typically in areas where display screen equipment is located at each workstation. Category 3 luminaires are usually used where screen usage is random and where the number of screens is relatively low.

The Health & Safety (Display Screen Equipment) Regulations 1992, under the Health & Safety at Work Act 1974, refers to minimum health requirements in relation to work with display screen equipment. Employers must take steps to ensure that all workstations under their control comply with the regulations. In general terms the regulations require lighting conditions which are satisfactory and which ensure that an appropriate contrast between display screen equipment and the background environment be provided. Any glare or distracting screen reflections must be prevented.

21.6.2 Luminaires

The luminaire formerly referred to as the 'light fitting' provides support, protection and means of electrical connection to the lamp, which is contained within the luminaire. Additionally the luminaire has to be able to operate safely and to withstand the environmental conditions in which it is likely to be installed.

Luminaires may also be classified according to the following categories: type of protection provided against electric shock, the degree of protection provided against the ingress of dust or moisture, type protection provided against electrical explosions and according to the characteristics of the material of the surface to which the luminaire can be fixed.

Those luminaires that are designed for use internally, and those for use in some external applications, will normally operate efficiently in dry and well-ventilated atmospheres.

Some luminaires are installed in atmospheres that are far less acceptable. In order to be able to perform within the environments likely to be encountered, the equipment must therefore be manufactured so as to comply with strict specifications. Such environments are referred to as hostile or hazardous. Examples of hostile areas include: high humidity conditions (requiring drip-proof equipment), dusty and corrosive atmospheres, and food factories (where the interior walls, floors and ceilings are required to be hosed down).

Hazardous atmospheres are usually created by the presence of flammable or explosive dusts and gases in the atmosphere.

21.6.2.1 Optical control of light output from luminaires

When a bare lamp is used, i.e. one without any form of control of the directional qualities of the light emitted, then the distribution of light is likely to be completely unacceptable. Furthermore the bare lamp is also likely to create a source of disability glare to the occupants of the interior. It is likely that the lighting installation will be uneconomical and whilst some fraction of the light output from the bare lamp will reach the working plane either directly or indirectly, the efficiency of the installation will be relatively low.

It will be clear that some means of control of the light output from the bare lamp is essential and four of the most widely used methods are detailed:

Obstruction When a bare lamp is installed within an opaque enclosure, which has only one aperture from which the light can escape, then the light distribution from the basic luminaire will be severely limited (*Figure 21.31*).

Reflection This method of light control uses reflective surfaces, which may range from matt to specular. By comparison with the 'obstruction' method of light control, reflection is more efficient. With reflection, stray light is collected by the reflectors and then redirected (*Figure 21.32*).

Diffusion When a lamp enclosure, is constructed of a translucent material, two benefits will accrue. The apparent size of the light source is increased and simultaneously there is a reduction in perceived brightness. One disadvantage as a consequence of the use of diffusers is that they absorb some of the light emitted from the source itself and so therefore there will be a reduction in the overall

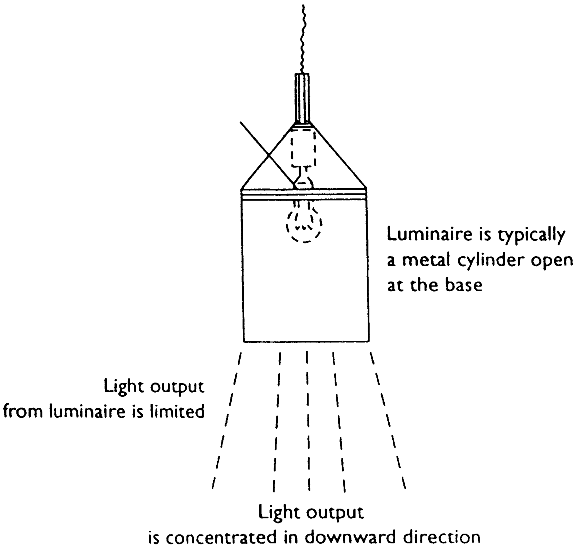

Luminaire is typically a metal cylinder open at the base
Light output from luminaire is limited
Light output is concentrated in downward direction

Figure 21.31 Light control by obstruction

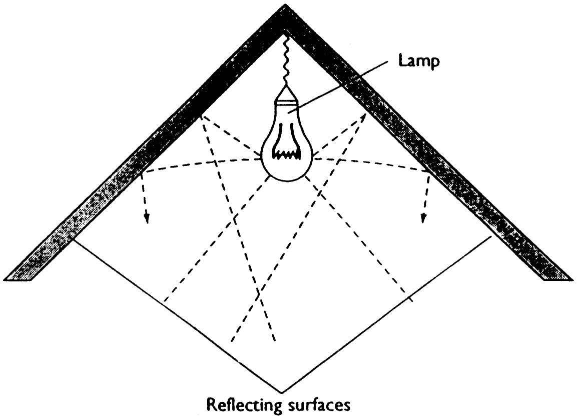


Figure 21.32 Light control by reflection

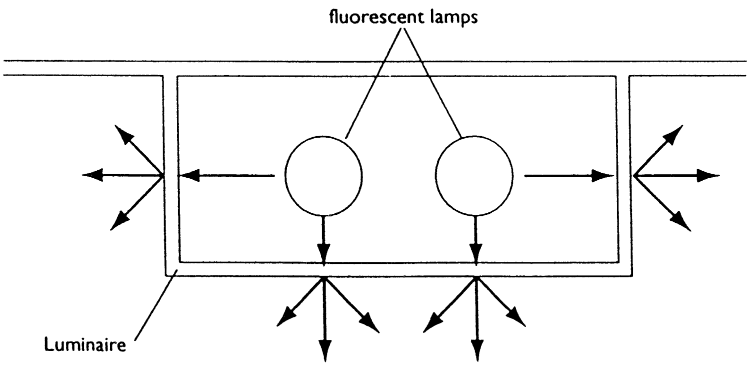


Figure 21.33 Light control by diffusion

luminaire efficiency. *Figure 21.33* shows the concept of diffusion.

Refraction This method utilizes the prism effect so as to bend the light output from the bare lamp in the required direction. This method of light control has the benefits of effective glare control and additionally of producing an acceptable level of luminaire efficiency (*Figure 21.34*).

21.6.2.2 Downlighter luminaires

Those luminaires classified as downlighters will emit the major proportion of their light output in the downward direction. They are typically ceiling-mounted or ceiling-suspended. When compared with uplighters, downlighters are far more efficient since they provide direct lighting of the working plane, and do not rely totally on light reflected from room surfaces.

21.6.2.3 Uplighter luminaires

Those luminaires that are classified as uplighters will emit the major proportion of their light output in the upward direction. *Figure 21.35* shows the general principle of lighting using uplighters. When compared to downlighters, uplighters are relatively inefficient since they rely on light being reflected from room fabrics. This places a greater importance on the room decor and reflectance values.

21.6.2.4 Luminaire materials

The materials from which luminaires are constructed typically include metals, glass and plastic. Glass had been used extensively in earlier luminaires but has largely been replaced by plastics, except for some specialist applications e.g. where the light source produces excesses of heat, in which case glass is preferred from a safety viewpoint. Furthermore glass still finds use in situations where there is the likelihood of physical damage to the luminaire, since glass will be more robust.

Two types of glass are in common use for the manufacture of the bowls of luminaires i.e. borosilicate glass which can withstand high working temperatures and soda-lime glass. Both of these types of glass have to be subjected to toughening processes in order to provide additional

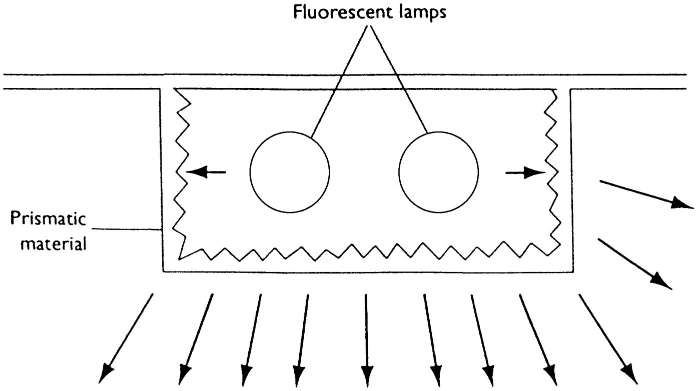

Fluorescent lamps
Prismatic material

Figure 21.34 Light control by refraction

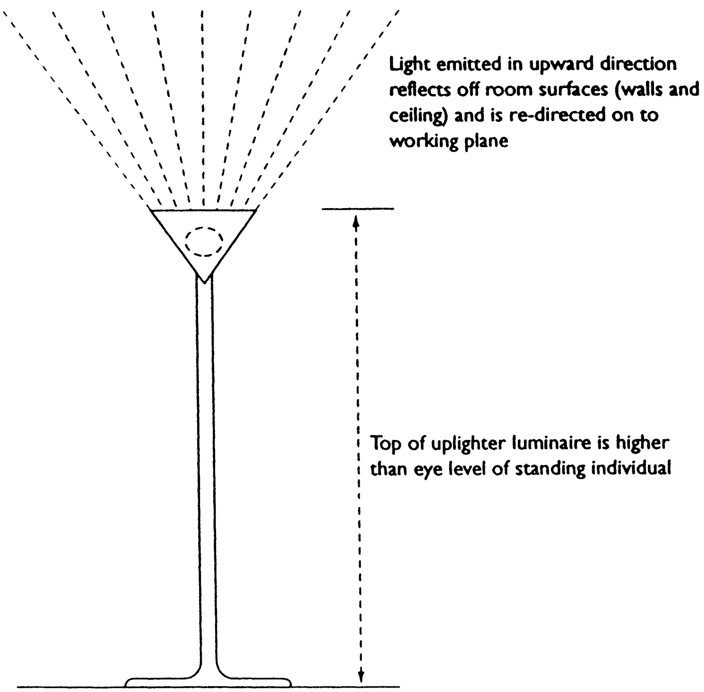

Light emitted in upward direction reflects off room surfaces (walls and ceiling) and is re-directed on to working plane
Top of uplighter luminaire is higher than eye level of standing individual

Figure 21.35 Principle of uplighter luminaire

protection if they are likely to be installed and operated in environments where there are rapid temperature variations.

Plastics, which are either thermosetting or thermoplastic, find their main use in the construction of diffusers. Polystyrene, acrylic and polycarbonate are used for diffusers. Polycarbonate, which is used in the manufacture of some types of safety spectacles, has a greater impact resistance than acrylic although both materials have been used for the construction of enclosures for use in areas where the luminaires are subjected to acts of vandalism. A form of plastic known as glass-reinforced plastic (GRP) is often used for luminaire canopies.

21.6.2.5 Mechanical strength of luminaires

Mechanical characteristics of luminaires include strength, windage resistance and resistance to vibration.

Luminaires must be capable of operating safely and efficiently within the environments in which they are expected to be installed. If they are operating in external environments they are likely to be subjected to meteorological factors such as wind, ice, snow all of which must be taken into account when considering the construction of the luminaires. Luminaires subjected to strong winds may also suffer the effects of vibrational oscillations that may lead to adverse operation.

21.7 Design techniques

The majority of industrial and commercial lighting problems are solved by the use of general lighting. One method of calculating the illuminance derived from a lighting system is known as the 'Lumen' method. Details of this are given below, procedure being grouped into logical steps.

Step 1: Illuminance Illumination values for various tasks are set out in the CIBSE Code for Interior Lighting.

Step 2: Luminaire type Consideration must be paid to all relevant factors, i.e. horizontal and vertical illumination requirements, glare prevention, efficiency, appearance, maintenance, economy, etc. For example, ordinary assembly work requires a certain amount of shadow to enable shape to be distinguished easily; therefore indirect lighting would be inappropriate, apart altogether from its relatively high running costs. The obvious choice for this class of work is the dispersive type of reflector.

Step 3: Mounting height The mounting height of fittings is usually dictated by the building construction, but, in general, it is advisable to make the height as great as possible compatible with good maintenance and installation facilities.

Step 4: Room reflectance It is now necessary to calculate, measure (or estimate) the effective reflectances of the three main room surfaces: (1) *the ceiling cavity* (area above the luminaires); (2) *the walls* (from the height of the working plane to the height of the luminaires); and (3) *the floor cavity* (area below the working plane).

Step 5: Room index From the room dimensions is calculated the 'room index' (RI), i.e. the ratio between twice the floor area and the area of the walls measured as the area between the working plane and the luminaires, i.e.

$$\mathrm{RI} = (L \times W)/[(L + W) \times H] \Leftarrow$$

where L and W are the room length and width, and H is the mounting height. Results may be rounded to the nearest value in the series 0.75, 1.25, 1.5, 2.0, 2.5, 3.0, 4.0 and 5.0. A room in which L and W are many times greater than H will have a high room index, while if L and W are less than H, it will have a low room index. It is important to realise that the room index is a measure of the relative, not the absolute, dimensions of the room.

Step 6: Utilisation factor and number of luminaires The room index affects the utilisation factor. The larger the room index, the higher the percentage of light reaching the working place. The utilisation factor will also be affected by the reflectances. The higher the room reflectances, the more light will be inter-reflected around the room. Utilisation factors are provided for different combinations of room reflectance and room index. In more recent publications utilisation factors (UF) may have the letters F, W or C in parentheses to indicate whether they refer to the floor, walls or ceiling, respectively, although normally only UF(F) values are published.

Table 21.4 gives an example of standard utilisation factors used in lighting calculations.

Allowance must be made for depreciation in light due to dust and dirt on fittings and surroundings: this is in the form of a maintenance factor. For normal interiors the factor is typically 0.8; for dirty locations it may be as low as 0.4.

For installations in high-bay foundries it may be necessary to make additional allowance for light absorption due to dirt and smoke. The average illuminance $E(\mathrm{F})$ over the working plane is given by

$$E(\mathrm{F}) = [\mathrm{UF(F)} \times n \times N \times \Phi \times \mathrm{MF} \times \mathrm{CF}]/A$$

where UF(F) is the utilisation factor for the reference surface S, N is the total number of luminaires in the installation, Φ is the light flux of each lamp (bare), MF is the maintenance factor of the installation, CF is the product of any additional correction factors necessary, n is the number of lamps per luminaire, and A is the area of the working plane.

Alternatively, the number of luminaires required is

$$N = [E(\mathrm{F}) \times A]/[\mathrm{UF(F)} \times n \times \Phi \times \mathrm{MF} \times \mathrm{CF}] \Leftarrow$$

Step 7: Spacing of luminaires and layout For a regular array of luminaires, $\mathrm{SHR}_{\mathrm{max}}$ is the maximum spacing/height ratio that will provide acceptable uniformity. When non-symmetric luminaires are used in long continuous runs, the maximum transverse spacing can be increased to $\mathrm{SHR}_{\mathrm{max,tr}}$. In addition to these limits the sum of the transverse

Table 21.4 Example of a standard utilisation factor (UF(F)) table for $\mathrm{SHR}_{\mathrm{nom}} = 1.5$

Room reflectance			*Room index*								
C	*W*	*F*	*0.75*	*1.00*	*1.25*	*1.50*	*2.00*	*2.50*	*3.00*	*4.00*	*5.00*
0.70	0.50	0.20	0.43	0.49	0.55	0.60	0.66	0.71	0.75	0.80	0.83
	0.30		0.35	0.41	0.47	0.52	0.59	0.65	0.69	0.75	0.78
	0.10		0.29	0.35	0.41	0.46	0.53	0.59	0.63	0.70	0.74
0.50	0.50	0.20	0.38	0.44	0.49	0.53	0.59	0.63	0.66	0.70	0.73
	0.30		0.31	0.37	0.42	0.46	0.53	0.58	0.61	0.66	0.70
	0.10		0.27	0.32	0.37	0.41	0.48	0.53	0.57	0.62	0.66
0.30	0.50	0.20	0.30	0.37	0.41	0.45	0.52	0.57	0.60	0.65	0.69
	0.30		0.28	0.33	0.38	0.41	0.47	0.51	0.54	0.59	0.62
	0.10		0.24	0.29	0.34	0.37	0.43	0.48	0.51	0.56	0.59
0.00	0.00	0.00	0.19	0.23	0.27	0.30	0.35	0.39	0.42	0.46	0.48

Rating: 65 W, 1500 mm.
Mounted: on ceiling.
Multiply UF values by service correction factors.

and axial SHR values should not exceed twice SHR_{max}. If SHR_{max} is not given, it can not be assumed to be greater than SHR_{nom} (the value for which the UF table is calculated).

SHR_{max} and $SHR_{max,tr}$ provide information only about the maximum spacing/height ratio that will result in acceptable uniformity on the horizontal working plane. In practical installations, obstructions or other factors frequently make closer spacing essential.

The spacing of units is limited by the mounting height, building structure and plant layout. In all cases it is recommended that spacing/height ratios (issued by the luminaire manufacturers) be not exceeded if even illumination is desired. Where freedom from shadow is required, closer spacings should be used. In calculating spacings it should be remembered that the mounting height should be taken from the working plane, be it the floor, a desk at 0.85 m or a work bench at 1.0 m.

Ceiling divisions, columns, shafting, ventilation trunking and other obstructions restrict possible layouts of outlets. It is thus desirable to draw a scale plan of the area showing all obstructions and plan the layout on this. In multistorey buildings the units should form, if possible, a symmetrical layout in the ceiling panels formed by the joints.

Especially where fluorescent lighting is concerned, the use of continuous trunking from which the luminaires can be suspended and which carries the wiring should usually be considered, as in many cases it can effect considerable economies in installation cost.

Where work benches or machines are located along the outer walls, the distance between the outside rows of luminaires and walls should not exceed one-third of the nominal spacing distance. If the wall space is a non-working area, this distance can be increased to one-half the nominal spacing. This should not be exceeded, as it is desirable to have a certain amount of light thrown upon the walls in order to maintain a reasonable brightness throughout the room.

Plant layout may determine the location of outlets, as in the case of the textile industry. Localised general systems are installed in such cases, the outlets being localised in relation to the plant.

An installation will not be satisfactory if the maximum spacing/height ratio is exceeded. Conformity, however, will still be unsatisfactory if there are obstructions (such as large machines); closer spacing is then essential.

21.7.1 Lighting systems

It is possible to categorise interior lighting into general lighting, localised lighting or local lighting:

General lighting A general lighting system is one that attempts to provide a constant illuminance across a working plane in an interior. It is however extremely unlikely that a uniform level of illuminance will be produced at all points across a horizontal working plane.

It has to be appreciated that general lighting does not take into account the visual tasks likely to be undertaken in an interior. This can have both advantages and disadvantages. One advantage is that it allows a degree of flexibility when locating individual areas within an interior where visual tasks can be carried out. A disadvantage of general lighting is that such systems can be very wasteful of energy since some areas within an interior are illuminated to a level greater than that required.

Localised lighting Localised lighting systems usually provide a required illuminance on the work areas in combination with a reduced level of illuminance in non-working areas, for example walkways. An often-used example of this is found in open plan offices where workstations are lit using uplighters. Simultaneously walkways and other non-work areas are lit by means of a number of ceiling-mounted downlighters. It will be evident that this system is correspondingly less wasteful of energy than the general lighting system.

Local lighting Local lighting systems provide illuminance over a relatively small area in which the visual task is located. It is often used in combination with general lighting so that together the local lighting and general lighting will produce the required illuminance on, and surrounding, the visual tasks.

Figure 21.36 shows the differences between the types of lighting system described.

21.7.2 Lighting surveys

Lighting surveys are often undertaken in response to adverse criticism from workers who complain that the lighting in their workplace is in some way unsuitable and/or insufficient. Whilst such an allegation may sometimes be justified, it is often the case that in an environment where there are reported visual problems the cause is non-lighting in origin.

Such non-lighting influences on the visual environment must therefore be investigated before any decision is taken to implement a full lighting survey. Typical non-lighting causes include:

(a) changes in the condition of the decor, e.g. room fabrics have become dirty and their corresponding reflectance values have decreased;
(b) changes in condition of fenestration e.g. windows have become dirty and therefore the transmission of natural daylight to the room interior will be impeded;
(c) changes in layout of the interior e.g. workstations have been repositioned;
(d) changes in working practices and activities.

21.7.2.1 Instrumentation

The major parameters of interest likely to be considered in a lighting survey include illuminance, luminance, reflectance and daylight factor. Equipment is available which is capable of displaying, either in analogue or digital form, more than one of the parameters listed.

21.7.2.2 Illuminance measuring equipment

Such instruments typically incorporate a selenium or silicon photovoltaic cell. The specification of the instruments typically includes details of spectral response, angular response, linearity of response and operational characteristics of instruments in adverse temperature conditions.

The spectral response of the cells typically used in the construction of illuminance measuring instruments will differ from the response of the human visual system. It is therefore necessary to correct for this differential by applying some form of compensation. This is often achieved by using filters and when such filters are incorporated into an instrument, the device is referred to as colour corrected.

The magnitude of illuminance recorded by, and displayed on, an instrument is influenced by the cosine of the angle the incident ray subtends with the normal to the plane of the instrument detector. It is necessary therefore to apply

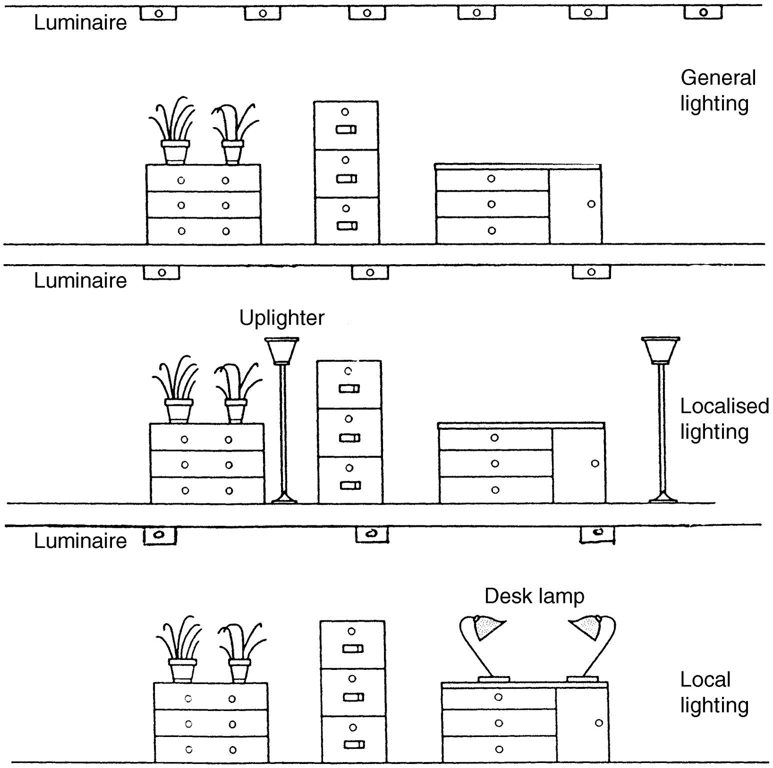


Figure 21.36 Types of lighting system

some form of compensation that takes into account the direction of the incident light falling upon the light-sensing detector. Instruments, which are capable of measuring illuminance values accurately from any incident direction, are referred to as cosine corrected.

The electronic components incorporated into the device will influence the linearity of response of the measuring equipment and the equipment may also be affected by adverse temperature conditions.

21.7.2.3 Luminance measuring equipment

Luminance measuring instruments incorporate photovoltaic cells, which also require to be colour corrected, and which are required to have a linear response.

21.8 Lighting applications

21.8.1 Office and interior lighting

It is possible to classify the main requirements for optimum lighting conditions within interiors as:

(a) health and safety;
(b) visual performance;
(c) aesthetics; and
(d) personal comfort.

When striving to obtain an office environment, which is both aesthetically pleasing and acceptable to the eye, it is usual to include:

(a) an analysis of those visual tasks that are likely to be encountered within an office;
(b) a balance between the quantity of daylight entering an office interior and the corresponding requirements from artificial lighting;
(c) a balance between direct and diffuse lighting contributions so that where practical no adverse three-dimensional and/or modelling effects will be produced;
(d) the provision of an environment that is free from both glare and any associated distractions.

When considering the health and safety of personnel, office lighting should ideally allow room occupants to carry out their normal duties in a manner that, under normal conditions, is considered to be safe. The lighting provided must not, under any circumstances, place any occupants of the room at risk. Furthermore it is necessary to provide for the safety of occupants in the event of an essential evacuation of the premises. In the United Kingdom emergency lighting is a requirement under The Building Regulations and is covered by BS 5266 Emergency Lighting.

It is important not to overlook both aesthetics and personal comfort when considering optimum office lighting conditions. There is a psychological influence on the occupants of an interior caused by the decor and room appearance and subsequently this is likely to have a 'knock on' effect upon worker behaviour that will ultimately affect output productivity.

21.8.1.1 Sick building syndrome and building related illness

The term sick building syndrome (SBS) is used to describe a host of symptoms which appear to have a high incidence in some buildings and which have a definite work relationship. It is essential to distinguish between sick building syndrome and other illnesses, which are connected with buildings, or building services. These building related illnesses are typically much less common; they almost always have a clearly identifiable microbiological cause and usually affect relatively few occupants. The term building related illness (BRI) is often preferred for such conditions as humidifier fever and Legionnaire's disease.

Sick building syndrome (SBS) can be caused by a host of factors, one of which is lighting. The symptoms of sick building syndrome (SBS) include headaches, nausea, dizziness, irritation of eyes, nose and throat and general lethargy.

21.8.2 Factory lighting

Typical outdoor lighting installations at factories include general external yard lighting, loading bays and storage areas.

The objective for factory roadway lighting is to provide suitable and sufficient illumination so as to allow the safe passage of personnel both on foot and in vehicles. To this end care must be taken when designing a lighting installation for such areas in order to provide the required illuminance but simultaneously to avoid the development of glare being experienced by works' personnel. It is normal to monitor and control values of average illuminance and point illuminance and in so doing maintain acceptable values of uniformity ratio so as to avoid producing 'patchy' lighting.

Loading bays and storage areas require special attention including due consideration given to the elimination of shadows.

21.8.3 Security lighting

The aims and objectives of security lighting are:

(a) to improve the likelihood of detection, identification and apprehension of intruders;
(b) to improve the efficiency of other security measures in use;
(c) to improve safety levels for authorised personnel.

Owners of shops will be concerned with the security of their premises especially where the sales rooms and associated areas contain stock, which could be of considerable financial value. It will be evident that in such situations the security lighting should pay particular attention to those areas where routine entry to the premises is gained, in addition to those areas of the interior that are visible externally.

Any lighting provided in the loading area will not only assist with the security of the premises but will also help with the normal operations of authorised personnel within the boundary of the premises.

One of the main considerations when contemplating the security lighting requirement for offices is the size of the premises. With a small suite of offices or a single lock-up office it is difficult to justify the use of dedicated security personnel. Conversely the situation applying with larger office blocks often demands the use of security guards whose sole function is to patrol and guard the offices.

Factories or larger scale premises will create different problems. In addition to the theft of finished goods raw materials are also targets for criminals and the security of such premises must therefore, of necessity, commence at the factory gatehouse or main entrance.

When designing a security lighting installation it is essential to avoid the production of shadows. In addition to posing a danger for authorised workers on a site, shadows are likely to provide areas in which criminals are likely to remain undetected and from which they can therefore subsequently make good an escape.

21.8.4 Floodlighting

Floodlighting can be applied to many installations e.g. buildings, industrial premises and for sport.

Floodlit buildings are commonplace in town and city centres. Thoughtful siting of luminaires and careful selection of the types of light sources used can produce a floodlit effect which is both striking and pleasing to the eye. It is however important to appreciate that floodlighting is not a procedure for illuminating a building during the hours of darkness to the same luminance level, and in the same manner, as that provided by natural daylight.

For optimum floodlighting of buildings there should be a flow of light across the front of a building. The direction of this flow should not be identical to that of the direction of normal viewing of the building front. Any contravention of this recommendation will lead to an absence of shadows producing a building appearance that is seen to lack character.

It will be evident that the colour output of the light source(s) used in any floodlighting scheme is critical, and the deliberate production of a colour difference can be used to advantage in the process of highlighting different areas of a building.

Industrial floodlighting is a necessary commodity for those locations where outside work continues during the hours of darkness. Adequate lighting of the correct type will ensure that maximum benefit is obtained from external work activities. Suitable and sufficient working illuminance combined with effective glare control should enable visual task details to be detected clearly, concisely and with speed which should subsequently allow work to proceed with safety and without creating visual problems for the workforce.

In many industrial activities it is essential to be able to discern the surface colour of engineering and production materials and it will be evident that floodlighting contributes markedly to the ease with which this identification process is carried out. In such situations those light sources with poor colour rendering properties are totally unsuitable and examples of such light sources include low pressure sodium lamps.

Sports floodlighting incorporates luminaires mounted on towers or grandstand fasciae. Light sources must have good colour rendering properties in particular when colour television transmission of events is likely.

21.8.5 Public lighting

Public lighting can be defined as any lighting provided for the public use, which is usually maintained at the public's expense. The functions of public lighting can be classified as:

(a) to ensure the continued safety of road users and pedestrians;
(b) to assist the police in the enforcement of the law;
(c) to improve the environment for the benefit of residents; and
(d) to highlight shopping areas and areas of civic importance.

When a driver is travelling along a road during the hours of darkness, objects are seen by one of two processes, either (a) direct vision (using surface detail) or (b) by silhouette vision. A driver will see as silhouettes either small objects at medium distances or large objects at greater distances. By directing the beams from the vehicle headlights on to the road so that they strike it at glancing angles (towards oncoming motorists) the whole of a road surface can be made to appear bright.

For optimum road lighting conditions the designer should strive for the following:

(a) a uniform road surface luminance;
(b) adequate and acceptable illuminance;
(c) suitable silhouette contrasts of the road ahead; and
(d) glare control from road lighting luminaires.

Low pressure sodium lamps have the highest luminous efficacy of artificial light sources and have been used extensively for road lighting. Their colour rendering properties are very poor and many public lighting engineers use such monochromatic sources only for minor road lighting installations. High pressure sodium lighting is now used on many trunk roads, other major roads and in town centres and areas of civic importance.

21.8.6 Light pollution

Extraneous light, in various forms, is a public enemy of increasing proportions. Astronomers are often particularly aggrieved in respect of this offensive light insomuch as they would prefer to view the night sky with as little stray light as possible. The situation occurs for a variety of reasons but in general terms it is due to either badly designed lighting installations or inaccurately aimed luminaires or a combination of both.

The likely consequences are sky glow or light pollution, where artificial lighting spills over, and therefore trespasses into, areas for which it is not intended. The situation becomes more pronounced when light is scattered through the night sky by particles of dust and droplets of water.

There appears little doubt that light pollution is a serious problem and unless the situation is addressed and subsequently reversed there is every possibility that the view of the night sky and all of the details contained therein, will be lost.

22

Environmental Control

R Tricker MSc, IEng, FITE(elec), FInstM, FIQA, MIRSE
Herne European Consultancy Ltd

Contents

22.1 Introduction 22/3

22.2 Environmental comfort 22/3
  22.2.1 Personal comfort 22/3
  22.2.2 Temperature and humidity 22/3
  22.2.3 Parametric limits 22/3
  22.2.4 Visual and acoustic parameters 22/5
  22.2.5 Widening the environmental specification 22/6
  22.2.6 Machines and processes 22/10
  22.2.7 Safety requirements 22/11

22.3 Energy requirements 22/11
  22.3.1 Steady-state loads 22/11
  22.3.2 Dynamic or cyclic loads 22/11
  22.3.3 Intermittent heating and cooling 22/12
  22.3.4 Plant capacity 22/13
  22.3.5 Computer-aided design 22/14
  22.3.6 Energy consumption 22/14

22.4 Heating and warm-air systems 22/14
  22.4.1 Radiators 22/14
  22.4.2 Convectors 22/15
  22.4.3 Warm-air systems 22/15
  22.4.4 Storage heating 22/15
  22.4.5 Air conditioning 22/15
  22.4.6 Cooling plant 22/16
  22.4.7 Cooling storage 22/18

22.5 Control 22/18
  22.5.1 Controllers 22/19
  22.5.2 Time controls 22/21
  22.5.3 Building management systems 22/21

22.6 Energy conservation 22/22
  22.6.1 Systems 22/23

22.7 Interfaces and associated data 22/24
  22.7.1 Electrical loads 22/26

22.1 Introduction

The environment in which we live and work has two basic elements: the external, over which we have as yet no control, although we are beginning to understand how our activities can moderate its characteristics; and the internal which can be maintained to specified conditions to meet our needs in terms of comfort and health.

Environmental control is still frequently considered only in terms of microprocessor- and software-based control systems which traditionally maintain specified conditions of temperature, humidity (in air conditioned spaces), illuminance and noise, to achieve comfort. This narrow concept of comfort has been affected by various aspects of the current debate on environmental and 'green' issues and a growing range of health-related criteria.

Until the energy crisis of 1973–1974, the specified criteria were achieved generally without any consideration of the efficient utilisation of energy used for the purpose. Since that time energy conservation and utilisation has become a major design parameter for all buildings and building-services systems. Because buildings in the UK use approximately one-half of the overall national energy consumption, there is considerable potential for saving by suitable integrated design and selection of equipment. Such savings affect both cost and environment.

The use of electrical energy is important in all forms of environmental control, whether it be for the supply of thermal power, circulation of air and water, or control. Apart from thermal power, where the choice of fuel is often governed either by its availability or the apparent economics during the design period, electricity will virtually always be involved with the other elements. The use of particular energy sources such as oil, gas, coal, electricity, etc., may be governed by the specific application. The comparison of fuels in terms of economics and costs to the client, as distinct from the primary energy consumption for each fuel is important and needs to be considered in selecting building-services systems, but as an indirect element of environmental control.

This review of environmental comfort is generally concerned with the internal or built environment, but the effect of design and operation of buildings on the external environment should not be forgotten. Any choice of materials or fuel used in buildings and the building's services affects the overall discharge of carbon dioxide (CO_2) into the atmosphere. The CO_2 is considered to be the classic greenhouse gas, contributing to global warming, and design and operating decisions which reduce CO_2 emissions will help to minimise long-term climatic effects. If current moderate scientific opinion is accepted, global warming could have a significant effect on the internal environment during the life-time of buildings which are being erected now. It is therefore important to identify the full range of parameters which may affect comfort and health within the built environment so that interaction between external and internal conditions can be identified. Many of the suggested comfort parameters are not energy related nor are they affected by the external environment.

22.2 Environmental comfort

The indoor environment should be safe, appropriate for its purpose and pleasant to inhabit. The parameters to be considered include the thermal, acoustic and visual conditions and are now expected to encompass health and psychological factors.

22.2.1 Personal comfort

In human terms an individual senses skin temperature, not room temperature, although the latter affects the former. The body loses heat by evaporation of moisture from the skin, convection to the surrounding air and radiation to, or conduction with, cold surfaces. These mechanisms, together with the degree of activity and the type of clothing worn, tend to maintain the skin temperature constant (except for exposed extremities) over a wide range of environmental conditions. However, real comfort occurs in a much narrower range of climate, and individual requirements differ considerably both intrinsically and, again, according to the activity and clothing. The narrow zone of real comfort conditions is often classified as neutral or comfortable. This, is shown in *Figure 22.3* in terms of temperature,[1] and illustrates the degree of satisfaction for a group of people in a particular space, about the optimum neutral temperature for the group. The specified space temperature is therefore always a compromise and is only one of the criteria affecting comfort.

Other parameters can have a marked effect on the space temperature necessary to provide optimum satisfaction to the occupants. One example of these effects (*Figure 22.4*)[1] illustrates the elevation in space temperature required to compensate for increasing air movement.

Beyond the parameters which are identified above, there are now other elements which have to be considered and they are identified later, after the following sections on more traditional criteria. For many of these additional elements more research is required before their full effects on comfort can be characterised.

22.2.2 Temperature and humidity

The term 'space temperature' has been used so far to avoid confusion. Most people assume that space temperature specified in terms of the dry bulb air temperature defines levels of warmth. The previous comments indicate that this may not be valid, although temperature detectors in common use are mainly calibrated for, and measure, dry bulb air temperature. Alternative temperature indices may provide better definitions of comfort conditions or are used for design calculations: these include equivalent, effective, globe, dry resultant and environmental temperatures. Environmental temperature is used for calculation, and resultant temperature θ_{res} is considered to be a measure of comfort dependent on internal dry bulb temperature θ_{ai}, mean radiant temperature θ_{r} and speed of air movement u as defined in

$$\theta_{\mathrm{res}} = [\theta_{\mathrm{r}} + \theta_{\mathrm{ai}}\sqrt{(10u)}]/[1 + \sqrt{(10u)}]\Leftarrow$$

Given that the recommended values of θ_{res} are those listed in *Table 22.1*, it is possible to adjust room temperature detectors or thermostats to a level suitable for comfort, for any mean radiant temperature and air velocity.

22.2.3 Parametric limits

Limits need to be applied to any specified comfort conditions, particularly in the case of temperature and humidity. Generally, in terms of human comfort, limits of ±2 °C about a specified temperature and a relative humidity (r.h.) of ±10%, about a mean of 50%, will be acceptable. Limits more critical than is necessary will create additional and unnecessary costs.

There may also be statutory limits which have to be applied in terms of energy conservation. In the UK since

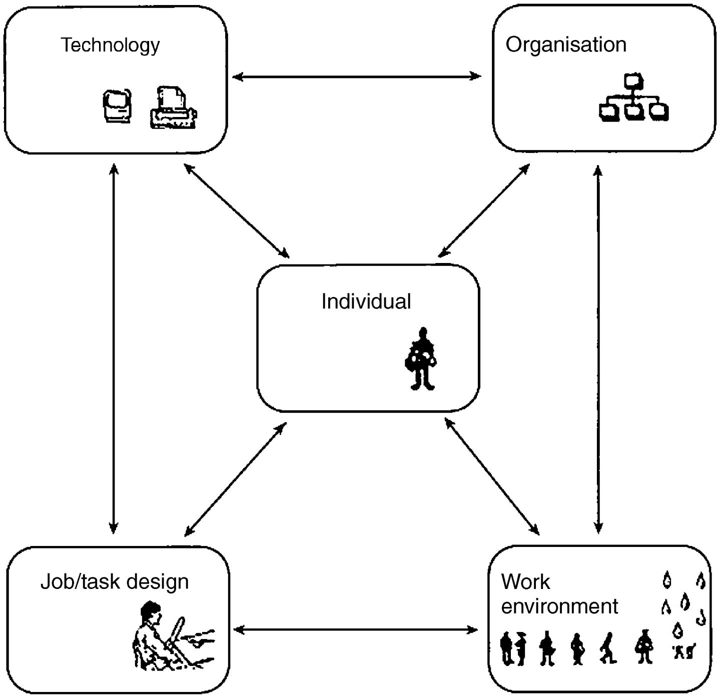

Technology
Organisation
Individual
Job/task design
Work environment

Figure 22.1 Model of the work system

1974, outside the domestic sector, space temperatures[2] now have an upper limit of 19°C. This is only a partial limit, because the reference covers only the heating. To complete the limit the regulation would have to specify an upper limit of 20°C for heating cycles and a lower limit of, say, 25°C for cooling cycles. Between these two limits there would be neither heating nor cooling input.

The level of humidity in a space can have a considerable impact on comfort, but in a heated building there is a limited range of control over its value. Artificially increasing

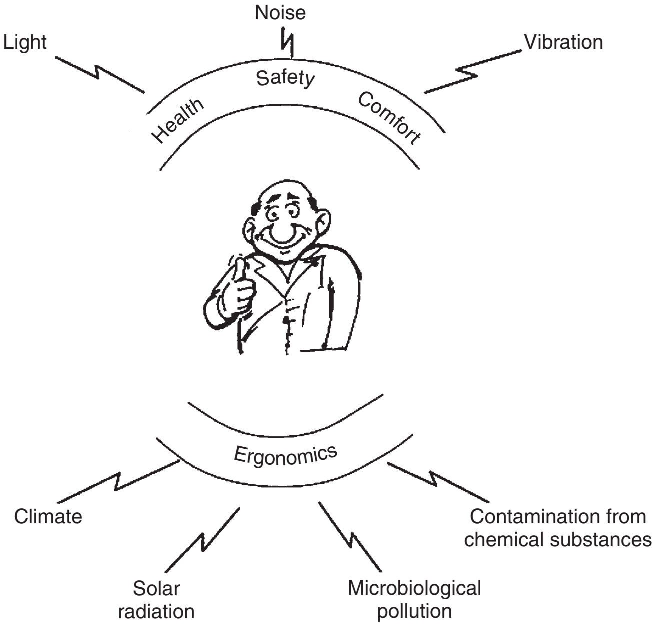

Noise
Light
Vibration
Safety
Health
Comfort
Ergonomics
Climate
Contamination from chemical substances
Solar radiation
Microbiological pollution

Figure 22.2 The operator/engineer's environment

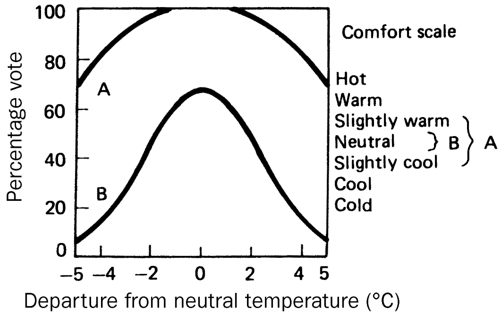


Figure 22.3 Comfort vote for personnel at around the neutral temperature for varying criteria. Curve A is for people giving any of the three central descriptions. Curve B is for the central description alone. (Courtesy of CIBSE)

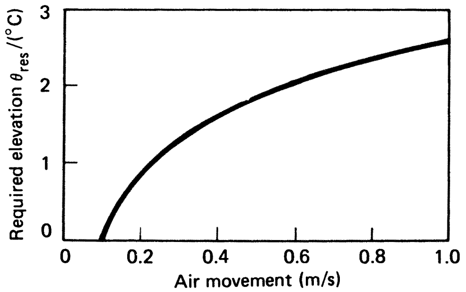


Figure 22.4 Corrections to dry resultant temperature for air movement. (Courtesy of CIBSE)

Table 22.1 Recommended design values for dry resultant temperature

Type of building	θ_{res} (°C)
Assembly halls, lecture halls	18
Canteens and dining rooms	20
Churches and chapels:	
$\leq 7000\,\text{m}^3$	18
$> 7000\,\text{m}^3$	19
Dining and banqueting halls	21
Factories	
Sedentary work	19
Light work	16
Heavy work	13
Flats, residences and hostels	
Living rooms	21
Bedrooms	18
Bathrooms	22
Hospitals	
Corridors	16
Offices	20
Operating theatre suite	18–21
Wards and patient areas	18
Hotels	
Bedrooms (standard)	22
Bedrooms (luxury)	24
Laboratories	20
Offices	20
Restaurants and tea-shops	18
Schools and colleges	18
Shops and showrooms	18
Swimming baths	
Changing rooms	22
Bath hall	26
Warehouses	
Working and packing spaces	16
Storage space	13

*Extracted from CIBSE Guide A1. Part of Table A1.3. Courtesy of CIBSE.

the air change rate by opening doors and windows is unlikely to be acceptable in winter and is a palliative in summer. Fortunately the band of comfort conditions in humidity terms is fairly broad for most people, and a range of 40–60% r.h. is usually acceptable and, in the UK, often occurs in practice internally. Below approximately 35% r.h., static electricity effects occur and noses and throats may be affected and above, say, 65% the effect of stickiness may be felt. Air-conditioned systems are normally designed to avoid these extremes.

22.2.4 Visual and acoustic parameters

The acoustic and visual impact on comfort conditions is extremely important. The correct level of illuminance for a particular task is important in its own right, but the overall aesthetics are a combination of the lighting level, the lighting source, the architectural finishes and their reflecting properties, and furniture and equipment. These aesthetics contribute to the comfort of the occupant.

Sound, in terms of personal comfort, is also related to the particular task. People's reaction to sound varies according to age and situation. Acoustics is both complex and subjective, and only a broad outline is included for the purpose of defining general criteria for comfort. Because the response of the ear is non-linear and less sensitive at low frequencies, it perceives equal loudness for various combinations of frequency and sound pressure levels, units of loudness being defined in phons. Sound pressure is a fluctuating air pressure sensed by the ear. The fluctuations are minute in relation to atmospheric pressure: sound pressure levels are specified in decibels. The levels are created by the sound power (the power transmitted by the sound waves) which is normally considered only in reference to a sound source. Sound power levels (L) are also referred to in decibels (dB).

$$L = 10\,\lg(W/W_0)\ \text{dB}$$

where W is the source power (in watts) and W_0 is the reference level (normally 1 pW).

Sound pressure is proportional to the square root of sound power.

A series of equal loudness curves (*Figure 22.5*)[1] is split into three sectors defined by A, B and C, which correspond to the sensitivity of the ear under varying conditions and can be measured by instruments with weighting networks corresponding to these bands. The subjective reactions for comfort in buildings are normally related to the A scale and the noise levels are quoted in dB-A. It is common to specify acceptable background noise levels for annoyance and speech intelligibility by means of NR (noise rating) or NC (noise criteria) curves, the former being most commonly used in Europe. Both sets of curves attempt to express equal human tolerances to noise across the audible frequency spectrum and are based on subjective experimental data. Normally the curves specify noise levels between 4 and 8 units below the measured dB-A values, although the relationship is not constant. *Figure 22.6* shows the NR curves and *Table 22.2* lists the recommended noise ratings for various situations.

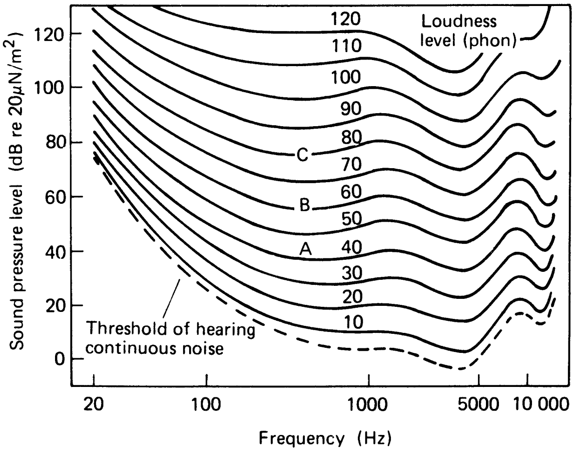


Figure 22.5 Equal-loudness-level contours. (Courtesy of CIBSE)

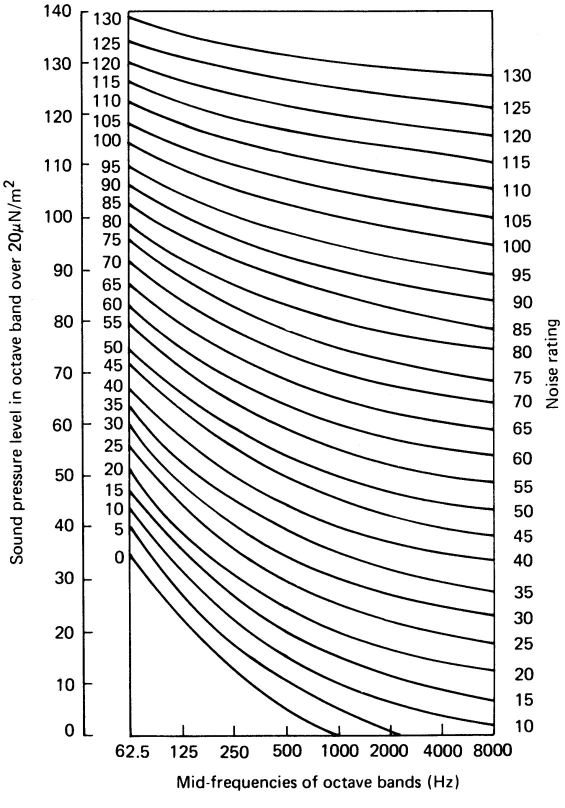


Figure 22.6 NR curves. Each curve is classified by a number corresponding to the speech interference level which was originally defined as the average of the sound pressure levels measured in the octave bands 600–1200, 1200–2400, and 2400–4800 Hz. The maximum permissible loudness level is taken to be 22 units more. Thus NR 30 has a speech interference level of 30 dB and a loudness level of 52 phons; this means that, for effective speech communication, the loudness level in a space designed to have a background level complying with NR 30 must not exceed 52 phons

Table 22.2 Recommended noise ratings*

Situation	*NR value*
Concert halls, opera halls, studios for sound reproduction, live theatres (> 500 seats)	20
Bedrooms in private homes, live theatres (< 500 seats), cathedrals and large churches, television studios, large conference and lecture rooms (> 50 people)	25
Living rooms in private homes, board rooms, top management offices, conference and lecture rooms (20–50 people), multipurpose halls, churches (medium and small), libraries, bedrooms in hotels, etc., banqueting rooms, operating theatres, cinemas, hospital private rooms, large courtrooms	30
Public rooms in hotels, etc., ballrooms, hospital open wards, middle management and small offices, small conference and lecture rooms (< 20 people), school classrooms, small courtrooms, museums, libraries, banking halls, small restaurants, cockail bars, quality shops	35
Toilets and washrooms, drawing offices, reception areas (offices), halls, corridors, lobbies in hotels, etc., laboratories, recreation rooms, post offices, large restaurants, bars and night clubs, department stores, shops, gymnasia	40
Kitchens in hotels, hospitals, etc., laundry rooms, computer rooms, accounting machine rooms, cafeteria, canteens, supermarkets, swimming pools, covered garages in hotels, offices, etc., bowling alleys, landscaped offices	45
NR50 and above NR50 will generally be regarded as very noisy by sedentary workers but most of the classifications listed under NR45 could just accept NR50. Higher noise levels than NR50 will be justified in certain manufacturing areas; such cases must be judged on their own merits	

Notes

1. The ratings listed above will give general guidance for total services noise but limited adjustment of certain of these criteria may be appropriate in some applications.
2. The intrusion of high external noise levels may, if continuous during occupation, permit relaxation of the standards but services noise should be not less than 5 dB below the minimum intruding noise in any octave band to avoid adding a significant new noise source to the area.
3. Where more than one noise source is present, it is the aggregate noise which should meet the criterion.
4. NR ≈ dB-A value − 6.

*Courtesy of CIBSE.

In conditions of adverse background noise the acceptability of differing noise sources may not depend on their absolute level and frequency but on their relationship with one another.

Outside the areas of normal sedentary or light industrial environments noise can rise to levels which may be injurious to health. At 90 dB-A or higher, exposure to the noise in confined spaces can be tolerated only for specific periods, e.g. 8 h at 90 dB-A, 2 h at 96 dB-A and 0.8 h at 100 dB-A.

22.2.5 Widening the environmental specification

The traditional criteria for comfort have already been identified as temperature, humidity, illuminance and noise.

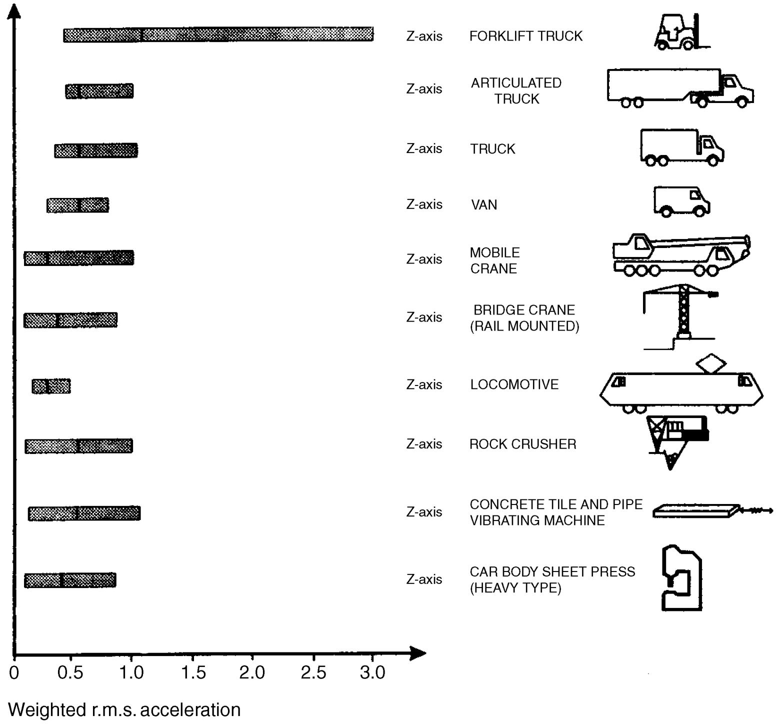


Figure 22.7 Examples of extra noise systems

There is much debate on the need to extend the criteria for defining a modern working environment which is comfortable and healthy. The range of additional possibilities is vast and it will take time before any of them become fully recognised and accepted nationally and internationally.

22.2.5.1 Comfort criteria

There are several comfort criteria which are commonly accepted as being important but do not form part of the current environmental specification.

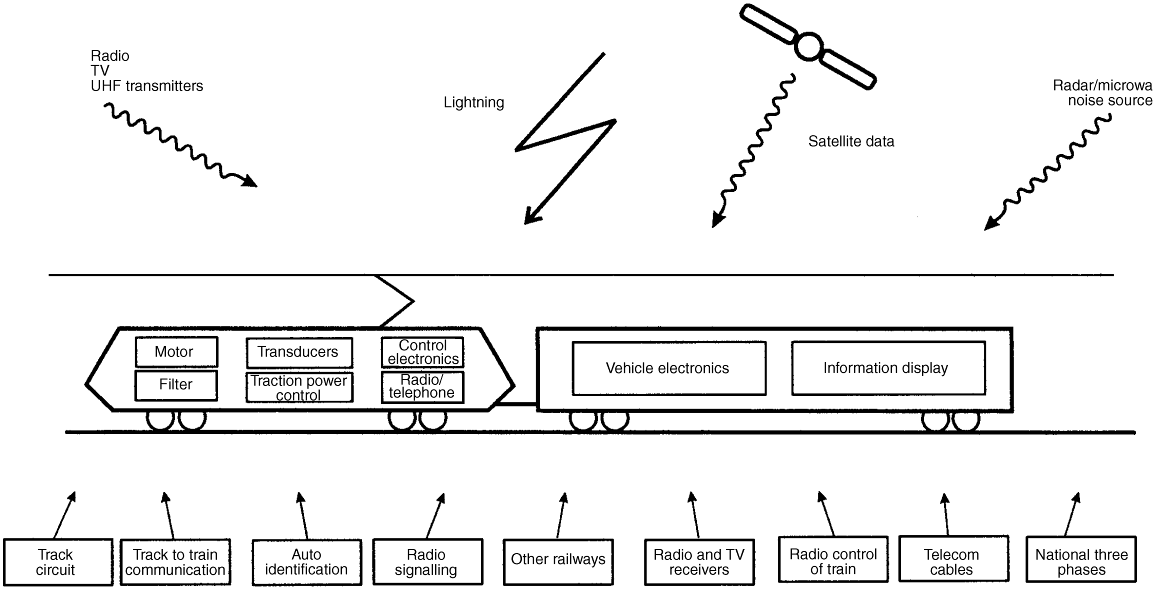

Radio
TV
UHF transmitters
Lightning
Satellite data
Radar/microwave noise source
Motor
Filter
Transducers
Traction power control
Control electronics
Radio/ telephone
Vehicle electronics
Information display
Track circuit
Track to train communication
Auto identification
Radio signalling
Other railways
Radio and TV receivers
Radio control of train
Telecom cables
National three phases

Figure 22.8 Examples of noise sources and disturbances in the railway environment

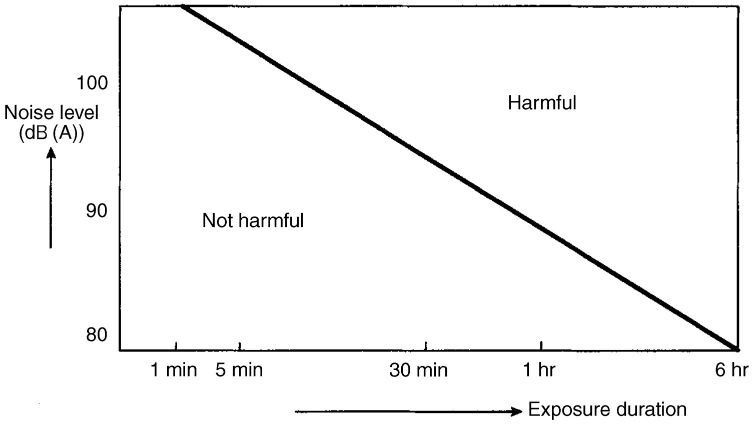


Figure 22.9 Harmful and non-harmful noise level

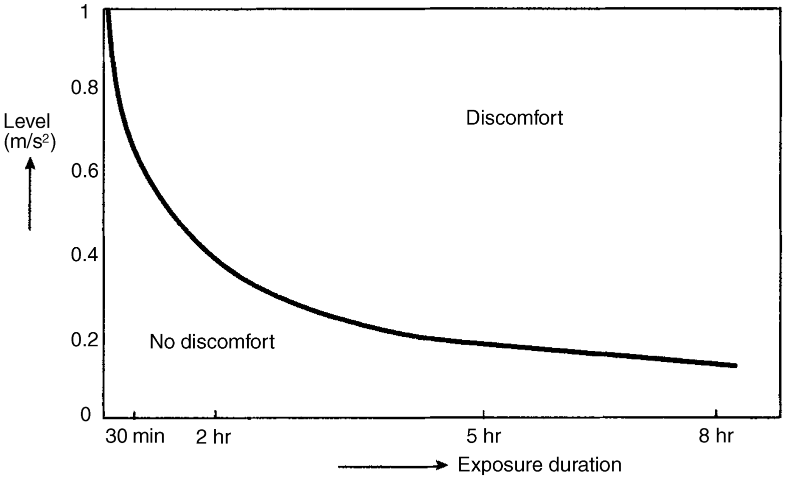


Figure 22.10 Body vibration levels (comfort and discomfort)

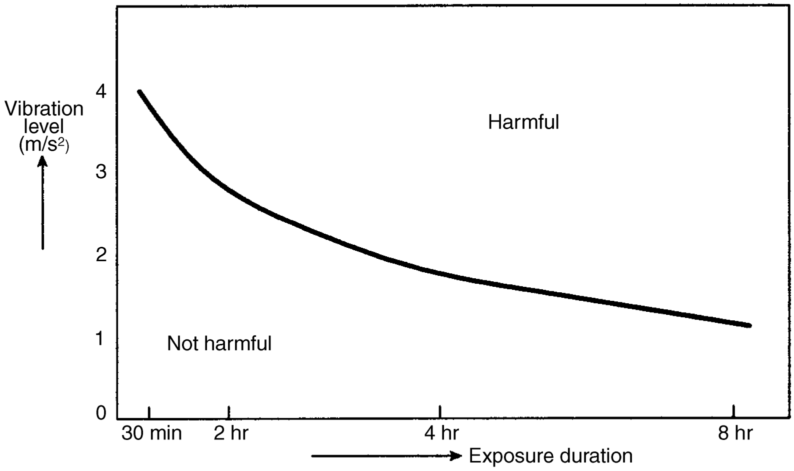


Figure 22.11 White finger vibration

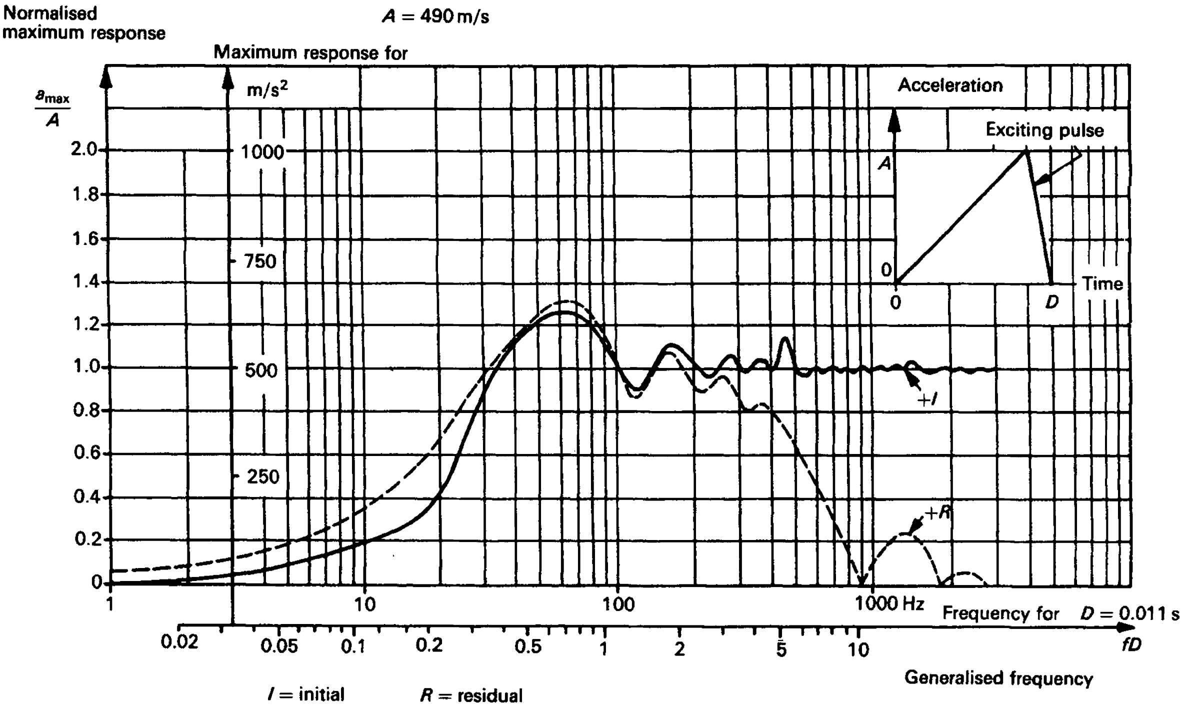


Figure 22.12 Maximum (annoyance) noise levels

Fresh air change and air movement These criteria are less likely to be part of the project brief and are commonly selected by the designer. Design guidance is available and quantitatively defined in professional handbooks and selection is a function of user needs and design experience. Their more explicit specification should be addressed.

Glare, veiling reflections, daylighting, luminance, etc. This group of criteria, like fresh air and air movement, are rarely specified in a project brief or specification. They too are part of the designer's evaluation and selection process and suitable guidance is available. There are current research and study projects examining a range of postulated potential problem areas relating to light sources, systems and the health of the working population. A familiar example is the unsuitable lighting arrangements for occupants using display terminals. There is a publication from the Chartered Institution of Building Services Engineers (CIBSE) entitled *Lighting for Visual Display Terminals*, which addresses this problem.

22.2.5.2 Colour finishes

There is another group of parameters which needs more consideration, some within the control of the architect, others the responsibility of the building-services engineer, and all of them require a co-ordinated approach by both disciplines. Colour finishes in a building and their reflectances, i.e. the amount of light which strikes the surface and is then reflected, are extremely important in terms of occupant satisfaction, and this applies as much to the furniture and equipment as it does to the walls, ceilings and floors. The same point could be made about the colour tint chosen for many window solutions, with the added point that it can detrimentally affect the occupants' perception of the outside environment. This group of criteria may sound somewhat esoteric in terms of comfort and health, but they do require objective review and further research.

The recognisably engineering factors which may enter more directly into future environmental standards are listed below.

(1) *Mean radiant temperature*—the effect of large hot or cold surfaces.
(2) *Infra-sound*—the possible effect of low-frequency vibration
(3) *Mains flicker and ultraviolet radiation from luminaires*—current studies and improvements in technology suggest that these are not, and need not be, problems.
(4) *Ionisation*—the effect of negative or positive ions in the air; an area of continuing debate and argument.
(5) *Information technology (IT) acoustics*—the effects of the various levels and frequencies of noise emanating from IT equipment may have a marked effect on occupancy comfort and fall outside the traditional specified noise criteria.

The parameters listed above are probably the most likely candidates to have a direct bearing on conventional environmental comfort, but there are other factors which may affect our comfort and health at work. They can be split into three further groupings.

22.2.5.3 Pollutants

Modern society recognises the multiplicity of pollutants in our environment generally, but the following identified groups specifically affect the built environment and comfort:

(1) tobacco smoke;
(2) fibres and dusts (mineral, paper, ink from printers, etc.);

(3) volatiles and organic vapours (from adhesives, etc.);
(4) micro-organisms (bacteria, viruses, spores, dust mites, etc.); and
(5) carbon dioxide.

All these pollutants, which may in some cases appear as odours, are created within the working environment, and very little is known about their concentrations, and the short- or long-term exposure effects at differing concentrations. Even less is known about mixtures of these contaminants and the possible 'cocktail' effect. It may be that a very small percentage of the working population is allergic to one or more of these pollutants. As the identified pollutants can exceed three figures, any summation of such allergies could conceivably affect a relatively large percentage of the occupants. On a longer term view, if any pollutants are identified as causing such problems,[3] their source materials could be banned, or fresh air change rates could be increased to lower the concentrations to safe levels—but only to a limited extent.

A recent study and tests by the Building Research Establishment, Garston on a UK building with a history of sick building syndrome (SBS) has indicated that a reduction in the dustmite population has also reduced the level of SBS complaints in the treated area.

22.2.5.4 Psychological factors

The debate on comfort and health covers both the obvious and the less frequently addressed criteria which affect our physical and physiological reactions to our environment. It also extends to consideration of the psychological factors which are outside the professional competence of the architect and engineer. Designers may in future be guided by medically orientated research into design solutions which could alleviate the effect of factors, such as:

(1) lack of job fulfilment;
(2) repetitive/boring work;
(3) lack of privacy;
(4) lack of individual identity;
(5) perceived lack of control over 'personal' environment; and
(6) poor management.

22.2.5.5 Sick building syndrome

Until recent times the design of building services was a function of an environmental specification which had to be met to achieve comfort. Nowadays the designer has to deal with several medically classified criteria apart from the, as yet, unresolved problems associated with SBS. Engineering designers are not professionally qualified to assess medical risks, but they can design to alleviate or eliminate such risks once they have been identified. Thus, design of buildings and building services can certainly contain the risks associated with Legionnaires' disease (and its associated family of illnesses), and radon, and they must always be considered in the design process. Less clearly defined, but possibly hazardous, are the effects from electromagnetic radiation generally, and high-voltage overhead cables in particular. Design teams may need to use consultant medical input as part of the design processes of the future.

The subject of SBS and its relationship with environmental comfort deserves some specific attention. As it is widely mentioned but not clearly understood. The World Health Organisation (WHO) identifies a range of symptoms for SBS which cause genuine distress to some building occupants, but cannot be clinically diagnosed and therefore cannot be medically treated. They typically include:

(1) stuffy nose;
(2) dry throat;
(3) chest tightness;
(4) lethargy;
(5) loss of concentration;
(6) blocked, runny or itchy nose;
(7) dry skin;
(8) watering or itchy eyes; and
(9) headache.

Affected individuals may suffer from one or more of these symptoms and the syndrome is characterised by the additional feature that the symptoms are said to disappear soon after the affected people leave the building.

Is it important? Sick building syndrome was reported as long as 30 years ago, but its significance, or apparent tendency to increase, has only been apparent over the last few years. Recent research[4,5] has indicated that it occurs more often in air-conditioned buildings than others, but as it does occur in heated and naturally ventilated buildings, the cause cannot simply be ascribed to air conditioning. Overall comfort and health in buildings cannot be satisfied by the heating, ventilating and air conditioning (HVAC) industry in isolation.

While SBS may be classified as an illness, there is no absolute proof that it is caused by any one factor or combination of factors. A large number of studies are being carried out,[6] often initiated because of the high level of occupant dissatisfaction or SBS symptoms in the working environment.

There is a body of evidence which suggests that levels of comfort, health or SBS may be caused by a combination of factors, but the mix and the weighting against each is indeterminate at present. Any of the non-medical parameters previously identified may individually, or in combination, contribute to SBS. To the list must be added design which does not conform to the current best practice and the quality of maintenance and hygiene in buildings.

22.2.6 Machines and processes

Apart from comfort criteria for personnel, there are two other areas where environmental conditions may be important: machine rooms and process plants. The former covers such spaces as computer rooms and medical machine areas; and the latter, areas such as electronic manufacturing or food processing factories. In many ways the criteria are similar to those for personal comfort, but there may be requirements for closer limits and, in particular, air filtration becomes a significant factor, the number and size of particles being very closely defined according to the process. In certain critical processes the air movement patterns are also specified, laminar flow probably being the most difficult to achieve. While temperatures are often specified with limits of not greater than ±0.25 or ±0.5°C, with humidities to ±2% r.h. or better, two points should be made. Limits specified may be unnecessarily stringent and need to be questioned. As an example, computer rooms in the past needed close limits if the machines were to operate correctly; but nowadays this is not normally necessary. The second point concerns the achievement of conditions throughout the treated space. Strictly, the specified conditions can normally be achieved only at the point of detection and control, the variation throughout the remainder of the space being largely dependent on good plant design and distribution of the heating and cooling media.

22.2.7 Safety requirements

In environmental control the specified conditions and their limits have to be achieved, but generally only during normal periods of occupation. It is therefore necessary to consider the environmental conditions necessary under abnormal circumstances. Examples include: (1) emergency or maintained lighting levels when the normal system breaks down or during unoccupied hours, (2) the low limit temperatures to be maintained to prevent freezing or damage to equipment and furniture, and (3) the maintenance of humidity below a specified dewpoint condition to prevent condensation on cold glazing.

22.3 Energy requirements

To achieve comfort conditions energy is required. The selection of electricity, gas, oil, solid fuel or alternative energy sources is a function of the required conditions, the plant selected, the economics, availability of supply, client preference and, to some extent, crystal-ball gazing.

The calculation of the energy necessary to achieve a particular set of environmental conditions is based on a number of concepts ranging from the simple to the complex and covering both the actual loads for heating and cooling the spaces and the plant sizing to deliver these loads. More sophisticated techniques have been developed to improve the accuracy of calculations and predictions in terms of heating or cooling loads and energy consumption. It is necessary to remember that the more precise figures derived from these techniques are valid only so long as the buildings to which they are applied are built with the same precision and with materials having the same indices used in the calculations. Care must therefore be exercised to ensure that the calculation procedures do not aim for an order of certainty that cannot be achieved in normal construction.

22.3.1 Steady-state loads

In its simplest form the steady-state heat load for a space within a building may be defined as

$$Q = [\Sigma UA + \tfrac{1}{3}NV]\Delta\theta$$

where Q is the energy required (watts), U is the thermal transmittance of any element surrounding the space (W/m^2 K), A is the area of the element (m^2), $\Delta\theta$ is the temperature differential across the element (°C), V is the volume of the space (m^3), and N is the number of external air changes per hour (h^{-1}).

The thermal transmittance of a wall, roof, floor, etc., is based on an electric circuit analogue. Normally each wall, etc., is a laminar structure of parallel layers of different materials and air spaces, each having a thermal resistance depending on its composition, thickness and surface properties. Given the thermal resistances (in m^2 K/W) as R_{si} and R_{so} for the inner and outer surfaces, R_1, $R_2, \ldots,$ for the component layers, and R_{a} for the air spaces, the thermal transmittance (in W/m^2 K) is

$$U = 1/[R_{\mathrm{si}} + R_1 + R_2 \cdots + R_{\mathrm{a}} + R_{\mathrm{so}}]\Leftarrow$$

Tabulated U values for a wide range of construction elements[7] are available for different exposures: some are given in *Table 22.3*. There are certain factors for which allowances may have to be made to the tabulated figures. The values quoted are for homogenous areas of construction, whereas in practice edge details and corners affect the figures. In addition, cold bridging can affect the values, i.e. the effect of wall ties between the inner and outer skin of the building or the framing round windows. Some examples of the ventilation rates used in normal calculations are detailed in *Table 22.4*.

The use of this method enables the heating load (or sensible cooling load) to be calculated for steady state conditions based on a minimum (or maximum) outside design condition and the additional assumption that the heating (or cooling) system will operate continuously. In practice systems generally operate intermittently and the building behaves dynamically. It is to cater for these factors that the more sophisticated calculations are introduced.

22.3.2 Dynamic or cyclic loads

To examine the dynamic performance of buildings, i.e. the energy requirements or loads under cyclic conditions, a procedure is used where the factors of admittance (Y), surface factor (F) and decrement factor (f) are introduced, the admittance having the greatest effect. The factors are functions of the thickness, thermal conductivity, density, specific heat capacity, position and frequency of energy inputs of each of the materials used in the construction. These have analogues with reactive loads in electric circuits. Consequently, there are phase changes (ϕ_y, ϕ_{F}, and ϕ_f) associated with them which, because the fundamental frequency is one cycle per day, are expressed as time lags/leads to the nearest hour.

The use of these factors leads to some complex equations which define the cyclic heat requirements for the building. The admittance can be thought of as the thermal elasticity of the structure, i.e. its ability to absorb heat; the decrement factor is a measure of how a cyclic heat input is attenuated as it passes through the structure; and the surface factor is a measure of how much of the cyclic input at a surface is readmitted to the space.

On thin structures the admittance equals the static U value; on multilayer constructions the admittance is largely determined by the internal layer. Thus, insulation on the inside of a concrete slab gives an admittance close to that of the insulation alone, whereas if the insulation is within or on the outside of the slab, the admittance value is virtually that of the slab alone. Decrement factors range from unity for thin structures of low thermal capacity, decreasing with increasing thickness or thermal capacity. Surface factors decrease with increasing thermal capacity and are virtually constant with thickness. Sample values of these three factors are shown in *Table 22.3*.[7]

Other factors affecting the load requirements include environmental temperature, solar gains, internal gains and the latent load for air conditioning plants, i.e. the amount of moisture that has to be removed (or added) to the treated air.

Environmental temperature (θ_{ei}) has already been mentioned and it is a concept used in carrying out load calculations, as it defines the heat exchange between a surface and an enclosed space. Its precise value depends on room configuration and the convective and radiant heat transfer coefficients of the surfaces. For the UK and hot climates it may be shown that

$$\theta_{\mathrm{ei}} = \tfrac{1}{3}\theta_{\mathrm{ai}} + \tfrac{2}{3}\theta_{\mathrm{m}}$$

Table 22.3 Thermal transmittance, admittance, decrement and surface factor for various constructions*

Construction *(outside to inside)*	*U* (W/m^2 K)	*Admittance*		*Decrement*		*Surface factor*	
		Y (W/m^2 K)	$\omega\psi$ (h)	*f*	$\phi\psi$ (h)	*F*	(h)
Brickwork							
220 mm brickwork, unplastered	2.3	4.6	1	0.54	6	0.52	2
220 mm brickwork, 13 mm dense plaster	2.1	4.4	1	0.49	7	0.53	1
105 mm brickwork, 25 mm air gap, 105 mm brickwork, 13 mm dense plaster	1.5	4.4	2	0.44	8	0.58	2
105 mm brickwork, 50 mm urea-formaldehyde foam, 105 mm brickwork, 13 mm lightweight plaster	0.55	3.6	2	0.28	9	0.61	1
Concrete blockwork							
200 mm heavyweight concrete block, 25 mm air gap, 10 mm plasterboard (on dabs)	1.8	2.5	1	0.35	7	0.64	0
200 mm lightweight concrete block, 25 mm air gap, 10 mm plasterboard (on dabs)	0.68	1.8	2	0.47	7	0.82	1
Roofs—pitched							
5 mm asbestos cement sheet	6.5	6.5	0	1.0	0	0.35	0
5 mm asbestos cement sheet, loft space, 10 mm plasterboard	2.6	2.6	0	1.0	0	0.74	0
10 mm tile, loft space, 25 mm glass-fibre quilt, 10 mm plasterboard ceiling	0.99	1.1	2	1.0	1	0.90	0
Roofs—flat							
19 mm asphalt, 75 mm screed, 150 mm cast concrete (dense), 13 mm dense plaster	1.9	5.7	1	0.34	8	0.50	2
19 mm asphalt, 13 mm fibreboard, 25 mm air gap, 25 mm glass-fibre quilt, 10 mm plasterboard	1.0	0.97	2	0.99	1	0.92	0

* Extracted from CIBSE Guide Section A3. Sample values from schedules. Courtesy of CIBSE.

where θ_{ai} is internal air temperature and θ_m is mean surface temperature.

Solar gains affect load calculations in two ways. First, there is the effect of solar radiation on the heat transfer characteristics of the building fabric, which is covered by the use of a parameter known as sol-air temperature (θ_{eo}). The definition of θ_{eo} is that temperature which, in the absence of solar radiation, would give the same rate of heat transfer through the wall or roof as exists with the actual outdoor temperature and the incident solar radiation, i.e. it is an artificial outside temperature to take into account the effects of solar radiation. The second factor covers the direct solar radiation gains through windows, some of which have an immediate effect and some of which is absorbed into the internal structure and readmitted subsequently. Both types of gain affect energy consumption, but in terms of load they are ignored for heating calculations and included for air conditioning load purposes.

Internal gains from lights, machines and occupants may be substantial. Again, in heated buildings the gains affect the energy consumed by the environmental plant but are not normally taken into account in calculating the design load (the energy input required for the coldest day). For air conditioning loads the inclusion of these gains is most important.

Latent gains are ignored in heated buildings but for air conditioning a considerable proportion of the maximum cooling load may consist of latent cooling, i.e. the removal of excess moisture from the air because of the use of fresh air for ventilation and internal gains from occupants and, possibly, processes.

22.3.3 Intermittent heating and cooling

The calculation of loads for steady-state and cyclic situations leads naturally to consideration of the effects of running the plant intermittently to satisfy only the specified environmental conditions during periods of occupation. Inherent in this are the following.

(1) The preheat period necessary to bring the building up to temperature under varying climatic conditions—in particular, on the coldest day for which the load is calculated.
(2) The thermal response of the building during preheat, which will depend on its construction, insulation and ventilation.
(3) The thermal response of the plant when first switched on.
(4) The ratios between preheat, normal heating and plant-off periods.
(5) Relative running and capital costs.

Table 22.4 Air infiltration rates for heated buildings*

Building	*Air infiltration rate* (h^{-1})
Assembly hall, lecture halls	1/2
Canteens and dining rooms	1
Churches and chapels	1/2
Dining and banqueting halls	1/2
Flats, residences, and hostels	
Living rooms	1
Bedrooms	1/2
Bathrooms	2
Hospitals	
Corridors	1
Offices	1
Operating theatre suite	1/2
Wards and patient areas	2
Hotels	
Bedrooms	1
Laboratories	1
Offices	1
Restaurants and tea-shops	1
Schools and colleges	
Classrooms	2
Lecture rooms	1
Shops and showrooms	
Small	1
Large	1/2
Department store	1/4
Fitting rooms	$1\frac{1}{2}$
Swimming baths	
Changing rooms	1/2
Bath hall	1/2
Warehouses	
Working and packing spaces	1/2
Storage space	1/4

* Extracted from CIBSE Guide A4. Extracts from Table A4.12. Courtesy of CIBSE.

While the actual calculations can be complex and related to the dynamic states already mentioned, there are some basic points which illustrate the situation. In a steady state or dynamic analysis with continuous plant operation, loads for particular spaces may be calculated and the output equipment sized on this basis. When intermittent plant operation is introduced, the situation illustrated in *Figure 22.13* is typical.

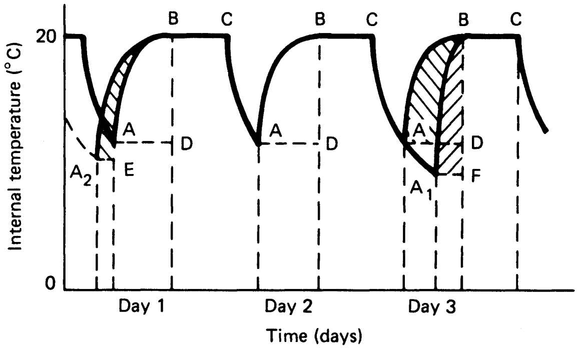


Figure 22.13 Intermittent-heating temperature/time curves

Here BC represents the period of occupation for which an internal dry bulb temperature of 20°C is required. CA represents the normal temperature decay in the space for a particular set of external weather conditions, and the area ABDA is the energy required to restore the temperature to 20°C at the start of the occupation period.

Because the normal load calculations are based on a constant temperature internally, it is obvious that if the temperature falls as shown, the energy input has to be increased above the steady state requirement, in order to raise the temperature to the required level. Calculating the amount of additional energy becomes part of the load calculations and may be based on variations of the cyclic state techniques already mentioned. The thermal response of the building and the time at which the plant is to be switched on are major factors in the calculations: *Figure 22.13* illustrates some of the variants which have to be taken into account. On the coldest day for which the system is designed the energy input is calculated in terms of the power to satisfy a switch-on time designated by point A. The actual position of A is temperature and time dependent, as indicated by points A_1 and A_2.

Point A_1 represents a situation where the temperature decay has been allowed for longer than that required for A. It is clear that the power input capacity to achieve point B from A_1 must be greater than that for point A, because the differential temperature is greater and the time allowed is reduced. But the actual energy for the purpose is the comparison of areas ABDA and A_1BFA_1. Therefore, apart from the calculations for the load requirements the economics of plant costs against cost in use (energy costs in this case) have to be evaluated.

The second point A_2 represents the situation after a non-standard shutdown (e.g. weekends) when the normal load is based on heating from A. Again it is necessary to calculate the load for condition A_2 or to arrange for alternative operation of the plant so that, for example, the temperature is not permitted to drop below A.

22.3.4 Plant capacity

While it is possible to calculate all the heating and cooling loads for achieving environmental comfort conditions in various spaces, including the sizing of individual output terminals, the selection of main plant to supply the necessary energy is another matter. The actual terminal size for a space has to take into account all the elements already considered, which clearly illustrates that capacity is rarely based on the steady-state load for continuous plant operation. Economic factors can sometimes inhibit optimum selection. The choice of suitable terminal sizes is reflected in the main plant selection, which has to cover the following points:

(1) heating plant capacity for intermittently heated buildings is normally based on simultaneous peak loads for all the spaces in the building, with such exceptions as the domestic sector, whereas cooling plant assumes diversity between the peak loads in various conditioned spaces (some diversity is permitted for continuously heated buildings).
(2) heating or cooling plant capacity should be sufficient to cater for process requirements in addition to the environmental load; e.g. the domestic hot water load may be purely for washing but may also include kitchen or restaurant requirements.
(3) sizing of source units should be such as to permit efficient operation under part-load conditions.

In respect of (3), particularly for multiple boiler or chiller installations, the environmental load in most cases only reaches the design peak for a small percentage of the total operating hours per annum, and the efficiency of boilers and chillers normally falls as their output decreases from the specified design level. It is therefore important to choose the units so that at, say, 25% of design load the operating source units are matched to the requirement to maintain a high efficiency.

The choice of refrigerant in chillers is environmentally important. Chlorofluorocarbons (CFCs) are currently the most common refrigerants in chillers used for commercial purposes and they deplete the atmospheric ozone layer and act as greenhouse gases. Under the internationally accepted Montreal Protocol, CFCs are scheduled to be phased out within the next 5–10 years and alternatives have to be provided which are environmentally safer. The current alternatives may be less efficient refrigerants than the CFCs they replace and more energy will then be required to produce the same cooling output, with a consequent increase in CO_2 released to the atmosphere.

22.3.5 Computer-aided design

The use of the computer for environmental comfort design is generally restricted to calculation of loads and annual energy consumptions, and to check whether the summertime temperature in the building rises to a point where air conditioning is essential.

Computer programs enable far more sophisticated techniques to be employed without laborious arithmetic. However, they should be used only as a design tool for the project, which still requires practical knowledge.

The programs for UK use are generally based on detailed extensions[8] of the techniques already outlined. Among the criteria which can be incorporated are the effects of shading provided by building configurations and overhangs, which affect both the load and energy consumption.

Apart from providing the calculated loads, energy-consumption programmes can provide the annual figures based on hourly weather data over a full year for any location and type of plant, so that a comparison between plants is rapidly available.

22.3.6 Energy consumption

There are three specific points to be identified in considering energy consumption: (1) degree-day figures, (2) energy budgets, and (3) energy targets.

The degree-day is a concept that permits energy consumption in a building to be monitored from year to year against a monthly datum. It is normally used only for the comparison of consumption in heated buildings, although a modified form is being considered for use with air conditioning. Degree-days measure the interval for which the outside temperature drops below a specific value (normally 15.5°C) and the amount by which it does so. The monthly figures are published for various areas. The base of 15.5°C is used in the UK, but other figures are used elsewhere to broadly represent that outside condition for which no system heating is required to maintain a suitable internal temperature: internal gains, etc., are always assumed to provide a rise of several degrees.

The energy budget for a building is the estimated annual energy consumption of a building. It is necessary for the cost-in-use evaluation.

Table 22.5 Energy targets for heated buildings

Building	*Consumption per annum*	
	(GJ/m^2)	($kW\text{-}h/m^2$)
Offices[1]	0.47–1.19	130–330
Factories[2]	0.68–1.08	190–300
Warehouses[2]	0.54–0.97	150–270
Schools[3]	0.36–1.26	100–350
Shops[4]	0.61–1.44	170–400
Hotels[4]	0.86–1.55	240–420

Notes:
(i) The figures are for heated and naturally ventilated buildings.
(ii) The range of consumption is a function of the thermal insulation, air sealing and efficiency of the heating system.
(iii) Figures are for conventional hours of occupancy and lighting levels. The factory figures include a 20% allowance for process gain.
(iv) The figures are for the UK.

References:
1 BRESCU, *Energy Consumption Guide No. 19, Energy Efficiency in Offices*, October, 1991
2 EEO, *Energy Efficiency in Buildings, Factories and Warehouses*, 1988
3 BRRCSU, *Energy Consumption Guide No. 15, Energy Efficiency in Schools*, September, 1991
4 CIBSR, *Applications Manual AMS: Energy Audits and Surveys*, 1991

Energy targets define the design aim for energy use in buildings, related to type, usage and location. The figures may be quoted in terms of power or energy per unit area (e.g. W/m^2 or kWh/m^2 p.a.). The official energy unit in SI terminology is GJ/m^2 p.a., where $1\,GJ \approx 278\,kW\text{-}h$. Buildings are now being designed to meet such energy targets and there is a trend towards making such figures mandatory in certain parts of the world. *Table 22.5* lists the targets for a variety of buildings in various locations.

22.4 Heating and warm-air systems

The majority of heating systems in the UK use water as the means of distributing thermal power (with steam as one variant of water). Electricity as a direct source of thermal power is an alternative which has advantages in certain situations but is frequently dismissed on the grounds of comparative energy cost. Water may be utilised for pure heating systems via emitters which produce radiative and convective heating, or for air heating systems where fan assisted devices produce warm air via a heat exchanger. The characteristics of boilers, by both type and use, are generally known in the engineering professions and are not covered here.[9]

22.4.1 Radiators

Radiators are the most common form of heating system. There is a roughly equal split between the radiant and convective heat output. On water systems cast iron for radiators has been largely replaced by mild steel. The radiant emission Q is a function of the difference $\Delta\theta$ between the temperature of the ambient air and the mean of the internal liquid, according to the expression

$$Q = k(\Delta\theta)^n$$

where k is a constant depending on dimensions and $n = 1.3$ for radiators. Electric radiators and tubular heaters are rarely used outside domestic or small commercial premises and even then only because no other form of heating is available.

22.4.2 Convectors

Convectors may be natural or fan convectors: both are common. The natural versions may be in upright cases with top and bottom grilles for air circulation, or used as skirting heating. Their emission may be designated in a form similar to that for radiators, but $n = 1.35$ for upright types and 1.27 for skirting versions. Water flow rate can have a considerable effect on emission. Fan convectors provide a form of air heating. They are normally controlled by switching on and off by means of a thermostat. In the emission formula, n is unity.

Convectors using electrical power for the thermal output are unusual compared with the water power versions.

22.4.3 Warm-air systems

Warm-air systems vary from domestic units to industrial systems. Some sophisticated versions which duct warm air through a large building complex may be treated as air conditioning systems without cooling and humidification elements, but such systems are rare.

A substantial proportion of heaters, with nil or minimal ductwork, are freestanding and distribute air locally on a recirculation basis without the introduction of fresh air. If the air volume is freely distributed, the complete unit is usually started and stopped by means of a space thermostat, or one mounted in the recirculation air inlet to the unit. Where the air flow is restricted, by manual or thermostatically controlled dampers to control the temperature by varying the air flow, it is necessary to ensure that the unit and fan are switched off when the air volume is reduced to a predetermined level.

The primary source of thermal energy for these units may be oil, gas or electricity. The latter is used in domestic units and in some commercial applications.

22.4.4 Storage heating

Underfloor heating can be operated with hot water or electrical energy as the thermal medium. The floor construction, covering, thermal time constant, temperature control and the idiosyncrasies of the user make system calculation difficult, and underfloor heating is consequently uncommon. Its future may depend on using low-grade heat spread over large emitting surfaces at temperatures of 23–30°C, possibly with solar panels as the heat source.

Storage heaters, electrically fed, are employed in domestic and commercial premises, which avoids the need for central heating plant. For adequate thermal capacity the units are unavoidably heavy (60–300 kg), with ratings of 0.5–6 kW. The refractory heat-storage blocks have heating element temperatures up to 900°C. The heaters are normally run on off-peak supply, emission from the units being regulated to match the periods during which the heat output is required. 'Natural' storage units radiate during the off-peak charging period. They are generally less satisfactory than 'fan assisted' units, which are insulated to restrict output during charging, the output being controlled by timers and thermostats to start and stop the fan, which controls most of the heat output.

The charge during the permitted period should be regulated in accordance with internal and external temperatures. Energy regulators are available for this purpose (see Section 22.5.2).

Other systems are based on large and well-insulated water storage tanks, electrically heated off-peak by boilers or immersion heaters. In some cases the storage temperature may exceed 100°C, which necessitates a pressurised vessel. Water is circulated (or, for the high-temperature case, injected) into conventional heating or air conditioning systems during the periods required.

22.4.5 Air conditioning

Air conditioning is the filtering, washing, heating and cooling of air to achieve specified temperature and humidity levels. In temperate climates building design and services can usually achieve reasonable comfort conditions without air conditioning, but a system may be found necessary (a) if the extreme conditions are not tolerable, (b) if the building design requires it, (c) if urban noise and dirt have to be reduced, and (d) if internal heat gains (e.g. computer rooms) have to be accommodated.

In an air conditioning system air is moved by fan power through the relevant space, which results in a physical sensation of air movement quite unlike that in normal heated spaces and with simple air heating. Clients and occupants have to be forewarned of this, otherwise there is a likelihood of complaints about the environment which are unwarranted.

22.4.5.1 Systems

Most systems have a section of plant which adjusts the humidity, to ensure that air passed into the conditioned spaces is suitable for the specified humidity (a 'dewpoint' condition of about 10 °C). The dewpoint plant is described in Section 22.5.1.3. After dewpoint treatment, air is ducted at high or low velocity, for which duct size, fan and acoustic treatment differ. Common systems available are given in the following paragraphs.

Constant volume Normally this uses branched ducting, each branch with a reheater controlled by a space temperature detector.

Dual duct Air from the main plant is split into two duct systems: one carries cold and the other carries preheated air. Both ducts traverse the building. Each space has a mixing unit, connected to both ducts and adjusting the hot/cold ratio in accordance with a room temperature detector.

Variable volume Cold air is distributed to individual spaces through terminals, which throttle the rate of air supply and are controlled by room temperature detectors. Throttling raises the static pressure in the ductwork, the effect being used through a control system to vary the air flow through the supply and extraction fans. Methods of control may be mechanical, or by the several speed control methods applicable to electric motors. As some air is always necessary in a space for ventilation, the terminals do not close in normal operation, and continue to feed in some cooling air. The system is therefore best applied to buildings that require cooling throughout the year, or to systems that incorporate small reheaters (although in the latter case the variable-volume system may be inappropriate). In general, the control of air volume and consequential reduction of air treatment provides a low energy system.

When the air supply from any terminal is throttled back to the minimum design volume, the air distribution pattern will obviously differ from the maximum volume situation. In extreme cases this can create environmentally unacceptable conditions. Many systems are now designed so that the terminal supplies a constant air volume consisting of a variable cooling component from the main air plant with the remainder provided by recirculated air from the controlled space. This has the effect of maintaining both the correct air distribution pattern and the requisite amount of cooling.

Induction Air is supplied at high velocity to terminal induction units, mainly perimeter mounted, which are fitted with heating and cooling coils. Air is forced out of nozzles in the unit at a velocity high enough to induce entrainment of recirculation air from the space in the discharge jets. Thus, the space air is circulated through the unit, which provides the necessary quantity of conditioned fresh air from the main air plant. The heating and cooling coils are fitted with control valves and sequenced according to the requirements of the space temperature detector, to maintain the correct conditions. An extract system removes the equivalent amount of air to that supplied by the main plant. This type of system is rarely used nowadays.

Fan coil This has some similarity to the induction system but the air is supplied at relatively low velocity to the units, each of which has its own fan. The coil configuration and control is similar to that for the induction unit, but electrically there is a dual requirement for a distribution system to serve the fractional kilowatt one-phase fans and to switch them off during plant-off periods.

Reversible heat pump cycle This system is often used when the main plant air is distributed through a ductwork independent of the units, and sometimes with a non-air-conditioned fresh air supply. Each unit contains a reversible cycle compressor so that it can produce either cooling, or heating in a heat pump mode. The energy transfer medium to and from the units is by a circulating water system with the water temperature controlled at approximately 24°C. Each unit then extracts heat from, or supplies heat to, the circulating water, depending on whether the unit is on the heat pump or chilling cycle mode. The circulating temperature is maintained by sequencing a cooling tower heat exchanger to lower the temperature, or a non-storage calorifier to raise it. The latter may frequently be electrically fed.

Units are generally ‘packaged’ to include controls. The most common have a 1 kW compressor with a one-phase motor and a fan of less than 100 W; but 3 kW (one- or three-phase) units with fans of power more than 100 W are available.

22.4.6 Cooling plant

In essence, cooling plant for comfort conditioning is indicated schematically in *Figure 22.14*, which shows a single machine and tower, but multiple systems are more common. The temperature of the chilled water is controlled by T_{P1} operating the evaporator system and the heat extracted from the primary water appears in the condenser. The cooling tower dissipates this heat and returns the water to the condenser at a fixed temperature dictated by T_C, which by varying the position of valve V_1 controls the amount of tower cooling.

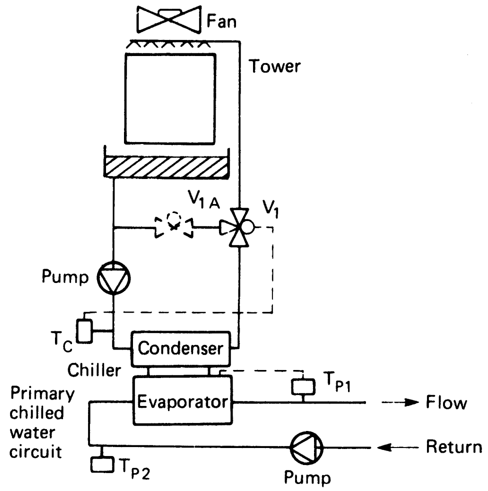


Figure 22.14 Schematic diagram of a single chiller and cooling tower. V_{1A} is an alternative valve position for V_1, and T_P is an alternative to T_{P1} for specific cases only

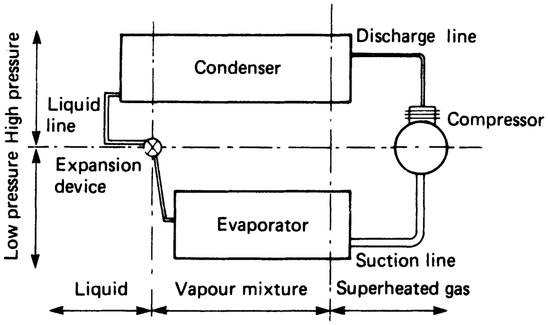


Figure 22.15 Scheme of a vapour-compression refrigeration cycle. (Courtesy of ASC Ltd)

Cooling towers are basically forced or induced draught types, the terms describing the method by which air is drawn past the sprayed water in the tower for cooling. Air-cooled condensers are also used where, in simple terms, the action of the tower is replaced by air blast cooling. Chillers can be of reciprocating, centrifugal, absorption and screw forms, all of which operate on a refrigeration cycle (*Figure 22.15*). The system circulates a refrigerant which has liquid and gas phases and a boiling point at atmospheric pressure well below that of water. Heat is absorbed by the gaseous and liquid mixture in the evaporator which then becomes a superheated gas: the absorbed heat is extracted from the primary chilled water circuit (*Figure 22.14*), and this is the prime function of the whole system. The gas is then passed through the compressor, which raises the gas pressure and also its boiling (or condensing) point. The condenser then extracts heat from the gas at this higher temperature and condenses it to a liquid again at a relatively high temperature and pressure: this process extracts heat from the refrigerant using the higher temperature water available in the tower circuit (*Figure 22.14*). The liquid then passes through the expansion valve which reduces both the temperature and pressure of the liquid as it expands to the gaseous and liquid mixture state for the cycle to repeat. The use of the expansion device does not alter the total heat content of the fluid.

Chillers[10] can be made by manufacturers operating internationally, motors may be rated (as in the USA) on a basis of 60 Hz, and may not be suitable for working on a frequency of 50 Hz. Again, 60 Hz control circuitry for one voltage may require modification for 50 Hz and a different voltage. Both input and output are expressed in kilowatts. This may cause confusion, because the output is 3–4 times the input, indicating the ‘coefficient of performance’ of the system.

Cooling towers have caused some outbreaks of Legionnaires' disease and there are certain situations where their use may be forbidden or restricted, particularly for hospitals and medical facilities. The alternatives, which are evaporative condensers or air blast coolers, use more energy per unit of heat rejection. Where cooling towers are properly maintained, with suitable water treatment, the risk is minimal.

22.4.6.1 Chillers

Reciprocating machines or compressors These operate on the compression of the refrigerant in a system analogous to that of an internal combustion engine, except that the 'fuel' is contained in a closed loop and an auxiliary motor drives the pistons. Compressors have various numbers of cylinders (up to 16) and methods of control. Machines are available for inputs up to 150–220 kW, corresponding to outputs up to 500–700 kW.

Centrifugal chillers Centrifugal machines use a rotating impeller which performs the compression operation on the gas by centrifugal force. *Figure 22.16* shows how the impeller is included in the system. One- or two-stage compression is normal. The control of output is normally by means of inlet guide vanes (not shown in the figure), which are actuated according to the load demand and alter the angle of entry of the gas into, and the performance of, the impeller.

Standard machines are available for outputs of 500–700 kW or greater, but not below 500 kW. Electric motor drives are most common and they may be hermetically sealed into the machine. The electrical rating is from 1/3 to 1/4 of the output rating and for machines above 500 kW input motors, operated economically at 3.3 kV or higher.

Absorption machines Absorption machines also rely on a refrigeration cycle. Analogy with other systems is best considered with the low-pressure section created by a permanently evacuated or high-vacuum system, and compression achieved by heating. The refrigerant fluid is a mixture of water and a fluid (normally lithium bromide) with an ability to absorb water. However, it is the water that acts as the refrigerant, as it will boil at low temperatures in a partial vacuum.[10]

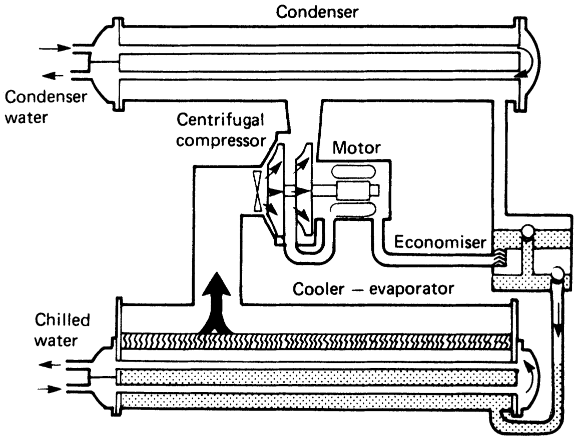

Condenser
Condenser water
Centrifugal compressor
Motor
Economiser
Cooler – evaporator
Chilled water

Figure 22.16 Scheme of a centrifugal chiller

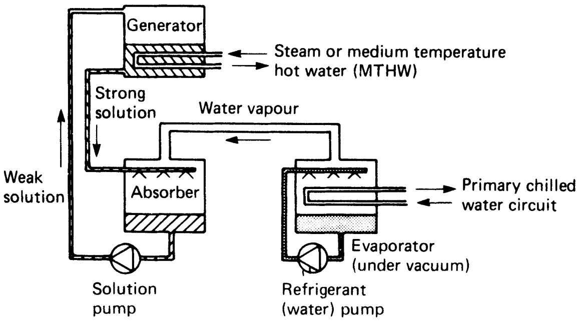

Generator
Steam or medium temperature hot water (MTHW)
Strong solution
Water vapour
Weak solution
Absorber
Primary chilled water circuit
Evaporator (under vacuum)
Solution pump
Refrigerant (water) pump

Figure 22.17 Scheme of a basic absorption cycle

Water evaporates more quickly if the surface of a given volume is extended. Rather than using a vessel with a large surface area, this is best achieved by spraying. *Figure 22.17* shows how the heat from the cooling load can be picked up by the action of water boiled in vacuum, the chilled water coils being immersed in the sprayed water. Because of its affinity to lithium bromide the water vapour is carried away from the evaporator to the absorber, where it is mixed to provide a lithium bromide solution. If these were the only cycle components, once the lithium bromide had become diluted its capacity for absorbing water vapour would be reduced and an equilibrium position reached whereby no further evaporation of the water and no further useful water cooling could take place.

The weak solution from the absorber can be pumped to a generator where, with the addition of heat, the water vapour can be boiled out of the lithium bromide to produce a strong solution which is returned to the absorber. Here it is sprayed to increase the surface area and, as in the evaporator, increase the capacity to absorb the water vapour from the evaporator. This secondary cycle maintains the absorbent at an operating level, but a water supply is required to replace the water vapour evaporated from the evaporator. If the rejected water vapour from the generator were passed to a fourth vessel, a condenser as shown in *Figure 22.18*, the vapour at a high temperature could be condensed and returned to the evaporator to complete the cycle. In addition to eliminating the need for make-up water, the fourth vessel provides a vacuum-tight system.

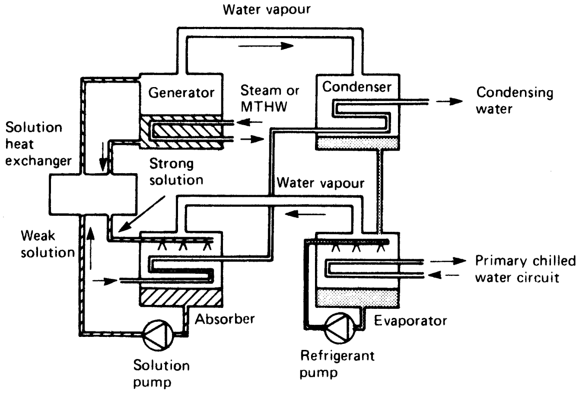

Water vapour
Generator
Steam or MTHW
Condenser
Condensing water
Solution heat exchanger
Strong solution
Water vapour
Weak solution
Primary chilled water circuit
Absorber
Evaporator
Solution pump
Refrigerant pump

Figure 22.18 Scheme of a full absorption cycle

When the lithium bromide solution absorbs water, heat is generated: it consists of the heat of condensation of the absorbed water plus the reaction heat between the lithium bromide and water vapour. To increase the capacity of the lithium bromide to accept the water vapour, it is kept cool by passing the condenser water first through the absorber and then onto the condenser.

Because the generator is hot and the absorber cool, the cycle efficiency can be increased by a heat exchanger which heats the weak solution pumped from the absorber to the generator and cools the strong solution returning.

In the diagrams used to describe the absorption cycle the flow of water vapour is restricted between evaporator and absorber and between generator and condenser, by the size of the connecting pipes. In practice this is overcome by housing the evaporator and absorber in one shell and generator and condenser in a second. Alternatively, all can be housed in one common shell with a division plate.

The machines are made in two basic size ranges. At the lower end the most common unit is the gas fired domestic refrigerator and direct gas fired units are also made for commercial use in outputs from 10 to 100 kW. The upper range is for outputs of 350 kW and higher, although 1500 kW is generally the minimum.

Energy for heating cycles associated with absorption machines is normally back-pressure steam from a primary process at 200 kPa (2 bar) or from medium-temperature water at 120°C. The steam system makes efficient use of energy which might otherwise be wasted.

Screw machines These belong to a range of positive displacement compressors, claimed to have advantages over conventional compressors in terms of reduced operating noise, lower operating speed and increased thermal efficiency. The machine essentially consists of two mating helically grooved rotors, a male (lobes) and a female (gullies), in a stationary housing with suitable inlet and outlet gas ports (*Figure 22.19*). The flow of gas in the rotors is both radial and axial. Compression is obtained by direct volume reduction with pure rotary motion. For clarity, the description of the four basic compression phases is here limited to one male rotor lobe and one female rotor interlobe space.

(1) *Suction*: as a male lobe begins to unmesh from a female interlobe space, a void is created and gas is drawn in through the inlet port. As the rotors continue to turn, the interlobe space increases and gas flows continuously into the compressor. Just prior to the point at which the interlobe space leaves the inlet port, the entire interlobe space is filled with gas.
(2) *Transfer*: as rotation continues, the trapped gas pocket is moved circumferentially around the compressor housing at constant suction pressure.
(3) *Compression*: further rotation starts meshing of another interlobe space at the suction end and progressively compresses the gas in the direction of the discharge port. Thus, the volume of the trapped gas within the interlobe space is decreased and the gas pressure consequently increased.
(4) *Discharge*: at a point determined by the built-in volume ratio, the discharge port is uncovered and the compressed gas discharged by further meshing of the lobe and interlobe space. During the remeshing period of compression and discharge, a fresh charge is drawn through the inlet on the opposite side of the meshing point.

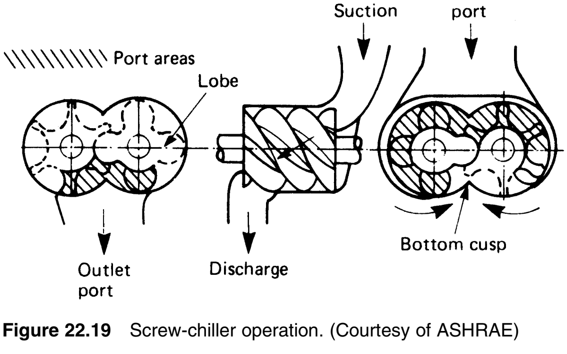


Figure 22.19 Screw-chiller operation. (Courtesy of ASHRAE)

Machines are available for outputs ranging from 250 to 2000 kW with both open and hermetic motor drives. The electrical input is from 1/3 to 1/4 of the output.

22.4.7 Cooling storage

Thermal storage systems have traditionally been used for the purpose of heating hot water overnight, see Section 22.4.4. The major advantage is reduced heating costs, not a reduction in overall energy consumed.

The principle has now extended to thermal storage for cooling systems where ice or low temperature phase change materials are used for the purpose, also in well insulated tanks. The charging of the stores is normally carried out at night by electrically powered chillers, although absorption chillers using gas firing or steam/high pressure hot water could be used. The system has several advantages in that smaller chillers can be selected than would be the case if they had to meet the peak daytime cooling load, the power supply system and maximum demand is therefore smaller and, again, cheap tariff electricity can be used for most of the operating load. The use of the systems also reduces the amount of chlorofluorocarbons (CFCs) or hydrochlorofluorocarbons (HCFCs) used in the air conditioning systems and can claim to be environmentally attractive.

All thermal storage systems need to be evaluated in terms of their pros and cons. The advantages mentioned above have to be carefully weighed against the costs of additional storage space, which can be large.

22.5 Control

The importance of controls to achieve comfort conditions has always varied according to the sophistication of the plant to which they are applied, the conditions specified and the economics of providing them. With cheap energy, simple control systems were the norm and comfort conditions were often of low priority and specified only as minimal levels. Energy is now expensive and comfort conditions are considered a more critical factor of human tolerance. Control systems are therefore commonly applied to all types of environmental system, performing the dual role of maintaining comfort conditions and conserving energy. In this combined role it is significant that for heated buildings in the UK a change of 1°C in normal space temperature will affect the energy consumption by as much as 10%.

The most common parameters considered are temperature, humidity and time. Generically, controls are either electric/electronic or pneumatic. The latter system uses clean dry compressed air as the motive power and is rarely independent of electric/electronic elements, whereas the former system uses electric motive power alone. Pneumatic systems tend to be used where there are many terminal controls, normally on air conditioned systems. Electric/electronic systems are now competitive with pneumatic terminal controls and it is rare to find new commercial buildings with pneumatic control systems. The systems described below are based on non-pneumatic systems. Where run and standby air compressors are supplied for pneumatic control systems, the electrical supply must be capable of starting and running both compressors together.

22.5.1 Controllers

Intrinsically, controllers operate in a two-position or a modulating mode. The former is recognised by on/off thermostats operating a device such as a fan or two-position control valve; the latter is a combination of detector and controller which can vary the position of the control valve over its full range of travel. The type of controller is selected according to the application.

22.5.1.1 Boilers and chillers

The controls for heating and cooling sources are normally supplied as part of the equipment. When multiple sources are used and sequential control is required to achieve a constant heating or cooling flow temperature, the electrical interlocking requirements are significant. *Figure 22.20* shows an arrangement for boilers in parallel. The interlocking necessary to achieve a constant flow temperature, for sequential operation which matches the load requirements, permits any source to lead the sequence, and various standard safety features need to be included. The control interlocks are extremely complicated when using relays and timers, etc., but modern usage normally adopts a software-driven solution which drastically reduces the hardware content of the interlocking package.

22.5.1.2 Heating systems

The control of heating covers the majority of systems in the UK. There are two elements to be considered: central plant control and terminal control. Apart from controlling the flow temperature from the boiler(s), central plant control for radiator and convector systems (the greatest number of heating systems) is carried out mainly by weather compensators. *Figure 22.21* illustrates such a system where the temperature to the load is controlled in accordance with external temperature. In cold weather water is supplied to the load at the boiler flow temperature, i.e. with valve V closed to the bypass and the setting of T_i corresponding to the boiler control temperature. As the outside temperature rises, a signal from T_e, passed via the controller, resets T_i downwards. T_i controls the position of the motorised valve V to mix water from boiler with water from the bypass (which is the return from the load and at a lower temperature than the boiler flow), to correct the temperature. As the temperature T_i decreases, the output from the terminals is reduced and may be matched to the load by selection of the temperature characteristic. *Figure 22.22*, curve 1, illustrates a typical characteristic.

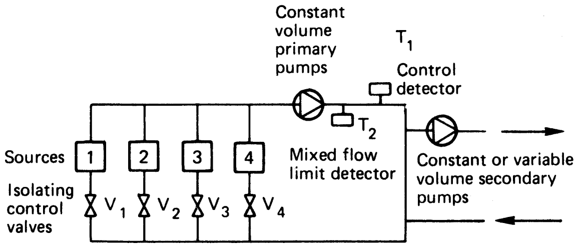

Constant volume primary pumps
T_1
Control detector
T_2
Sources
1
2
3
4
Isolating control valves
V_1
V_2
V_3
V_4
Mixed flow limit detector
Constant or variable volume secondary pumps

Figure 22.20 Flow-controlled multiple modulating boilers in parallel. Sequential control is from the detector in mixed flow-constant volume through boilers. The system controls any three of four sources.

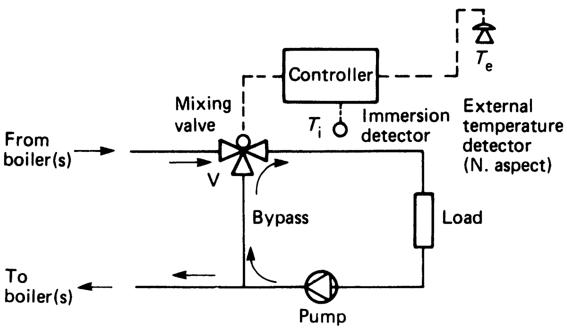


Figure 22.21 Basic scheme of a weather compensated system

The actual characteristic varies according to the system design parameters and to heat losses from the building, and may be varied in a number of ways. The characteristic may be generated and adjusted from point A (curve 2) or point B (curve 3) or points A and B (curve 4). A–C represents the limit condition which is governed by the boiler temperature, and may also be used for warm-up situations during which the compensator system is overridden.

A single compensator is incapable of dealing with varying internal loads or solar gain, and additional controls are employed for this purpose. These may be zonal controls, where an additional thermostat and control valve acts as a local trimming device to detect local gains, or terminal controls on each emitter. The latter permits individual temperature control of each space and might appear to make the compensator redundant. However, the compensator

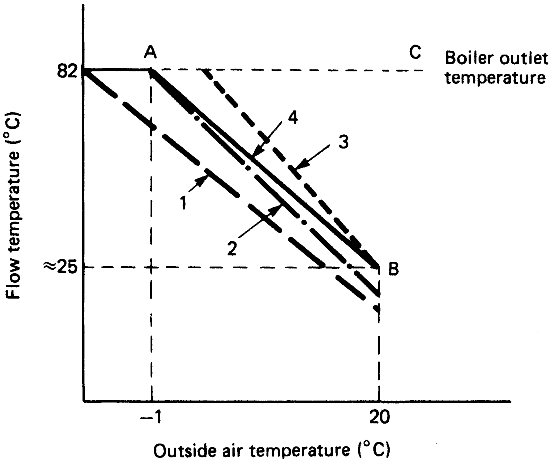


Figure 22.22 Typical compensator characteristics

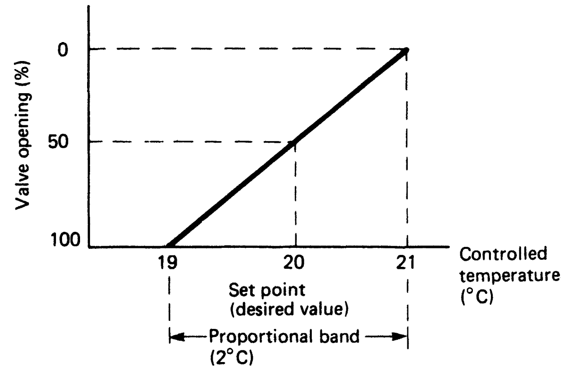


Figure 22.23 Proportional band effect on control setting. The temperature is shown for a set point of 20°C with a proportional band of 2°C

still provides two advantages: heat loss from the distribution mains is reduced as the outside temperature rises, and the reduced circulating temperature prevents individuals from calling for excessive temperatures in mild weather.

The most common terminal control for these systems are thermostatic radiator valves (TRVs), which are self-acting devices requiring no external supply. TRVs have one characteristic which is common to many control systems: they are proportional controllers. This means that they provide their set temperature only at one position of the valve, normally the mid-position as indicated in *Figure 22.23*. A specific change in temperature from this setting is required for the valve to take up any new position, which means that the valve moves from fully open to fully closed over a range of space temperature and that a particular valve position is related to a specific temperature. The sustained deviation from the set point which this describes is known as 'offset'. The range of temperature over which the valve performs its full travel is the 'proportional band' and TRVs have a fixed band which varies according to source and type. Typically the band is 2–3°C: thus, if the setting is 20°C, the space temperature rises to 21°C with the valve fully open and drops to 19°C with it fully closed (a 2°C band). This characteristic is useful in energy conservation terms, as it decreases the internal temperature at times of greatest load, thus reducing energy consumption, always assuming that the 19°C used in the example is acceptable to the occupants.

Controls for domestic heating systems[11] would not normally include a compensator, and TRVs or room thermostats controlling small motorised valves are more common. The use of a single room thermostat starting and stopping the complete system is common, but should be upgraded for both energy conservation and comfort. Warm air systems may be controlled by thermostatically controlled motorised dampers interlocked with the plant heat exchanger and fan.

22.5.1.3 Air-conditioning systems

Air-conditioning system controls are similar to those for heating in respect of the final emitters, i.e. a temperature detector-cum-controller, often proportional in character, which modulates either a control valve or a damper (or both in sequence) to maintain the required temperature. If the terminals depend on the sequential operation of heating and cooling, it is again possible to use the width of the proportional band to provide both comfort and energy conservation. The control may be adjusted to provide heating from 19 to 21°C and cooling from 22 to 24°C, with a dead zone between 21 and 22°C with neither heating nor cooling action.

The control of conditions in the main air-handling plant is a vital element for the overall space environment and is probably the most complex requirement in the heating, ventilating and air conditioning (HVAC) field. The aim is to achieve a stable dewpoint (fully saturated air) condition, which defines the amount of moisture in the conditioned air. The relation between dry bulb temperature, wet bulb temperature, humidity and total heat (enthalpy) is fully defined in psychrometric charts. Attainment of a particular dewpoint condition is sufficient to provide a specified set of relative humidity conditions in a space, taking into account the moisture pick-up (latent gain) from the occupants. *Figure 22.24* illustrates a typical dewpoint plant which contains basic controls and the necessary override and safety features. The details are related to the systems described in Section 22.4.5.

The dewpoint is controlled by T_1, which sequentially modulates a preheater battery control valve V_1, dampers D_{1a}, D_{1b} and D_{1c}, which operate in parallel, and a cooler

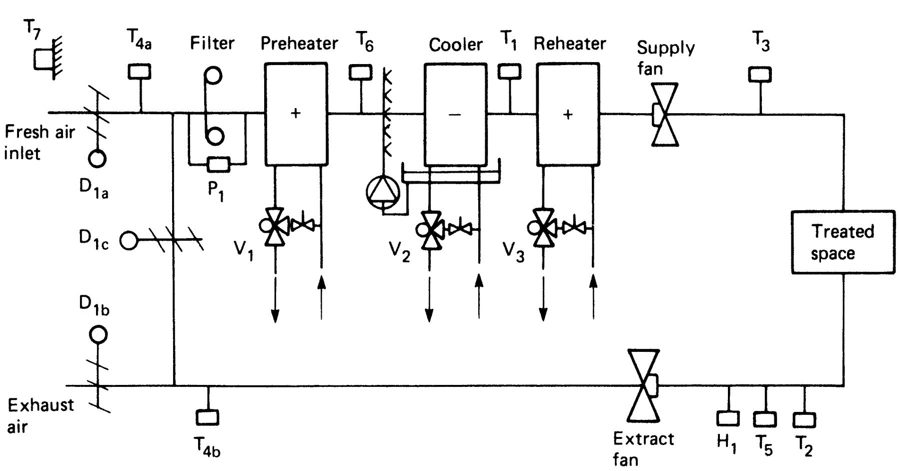


Figure 22.24 Scheme of the dewpoint. T_1, dewpoint detector; T_2, return-air-temperature detector; T_3, low-limit detector; T_4, enthalpy detectors/comparators; T_5, boost-limit thermostat; T_6, frost-protection thermostat; T_7 free cooling enthalpy detector; H_1, return-air-humidity detector; P_1, differential pressure switch; V_1, preheater valve; V_2, cooler valve; V_3, reheater valve

battery valve V_2 to maintain a constant saturated temperature condition. The reheater, which is part of many systems, is controlled by the extract temperature detector T_2, modulating the control valve V_3 to maintain a constant space temperature. The low-limit detector T_3 is sometimes employed in the discharge duct to override T_2 and maintain the discharge temperature above a predetermined limit.

The modulation of the dampers is the 'free cooling mode', using air for cooling prior to the use of mechanical cooling. Conditions can occur where the dampers need to be overridden. When the external temperature rises above the return air temperature (or, more precisely, when the enthalpy, or total heat, of the outside air exceeds that of the return air), it is more economic to cool return air than outside air. A detection device is therefore required to measure the total heat. The enthalpy of the outside air can be measured directly (T_{4a}) as being above the design room value, or by dual detectors (T_{4a}/T_{4b}) which compare the room and outside air total heat conditions. In either case, when the room total heat is exceeded by the outside air, a signal from the device drives the dampers to the minimum fresh air position, determined by the amount of air necessary to satisfy the fresh air ventilation requirements, which may be a statutory design parameter.

At night or during other shut-down periods the dampers are normally driven to the zero fresh air position. They remain in this position after plant start-up until the space temperature reaches a predetermined level as detected by T_5. During this boost period all the air is recirculated, valves V_1 and V_3 are fully open and the spray coil is de-energised. Thermostat T_6 protects the plant in cold weather conditions. If for any reason the temperature of the preheater drops to approximately 2°C, T_6 operates to shut down the plant and give an alarm.

It is also possible to satisfy the dewpoint condition without mechanical cooling, whenever the enthalpy of the external air is below that approximating to the dewpoint condition. This is accomplished by detector T_7, which holds off the chiller and cooling-tower plants whenever the enthalpy is below the predetermined level.

In rare cases a humidity detector (H_1) mounted in the extract duct from the conditioned space is provided to monitor excessive or reduced latent gains in the space. It then resets the dewpoint detector (T_1) to compensate.

A differential pressure detector (P_1) fitted across the plant filter is a standard control item to provide a warning of high pressure across the filter for both maintenance and energy conservation purposes.

Some interlocking features are desirable on all dewpoint plants. One is associated with fire defence. It is becoming standard practice, sometimes mandatory, to ensure in the event of smoke or fire that the plant shuts down and that firemen can start the extract fan independently of the supply fan, with the dampers run out of sequence. A frequent associated requirement is for recirculation dampers to be fully closed to facilitate smoke exhaust. The basic dewpoint control, and the various additional functions described, illustrate the possible complexity of the interlocking diagrams, which parallel those of the multiple boiler systems described previously. The sustained offset with proportional controllers makes them unsuitable for this application because the dewpoint temperature control is often critical for the comfort conditions required. Thus, floating or two-term controllers are used which do not suffer from offset.

In energy conservation terms, systems which treat all the air and bring it to a dewpoint condition are inefficient. Alternatives are frequently used, where the cooling and heating coils may be controlled in sequence from the detector in the return air duct, and humidity control is only employed at the upper and lower limits of the specified conditions, by using the latent cooling capacity of the cooling coil, or injecting moisture, respectively.

22.5.2 Time controls

The use of time switches is generally accepted. On large plants optimum start controls are becoming commonplace. The principle of operation of optimisers is to compare the external conditions with a representative internal condition so that the plant switch-on time is varied to achieve the required internal comfort conditions only at the time of occupation. In contrast, time switches are set to ensure that the conditions can be achieved on the coldest day, which means that in milder weather the building reaches comfort conditions earlier than necessary. Optimisers are therefore used for energy conservation, and can save 7–10% of the energy used with time switch control. The operation is shown graphically in *Figure 22.25*. As in *Figure 22.13*, BC represents the period of occupation and CA the temperature drop for the design coldest day. As the external conditions improve, the decay curve moves to CA_1, CA_2, etc., and the switch-on time is calculated by the optimiser and is delayed, moving from S to S_2, etc. A conventional time switch operating at S when the decay curve is CA_2 would waste energy equivalent to area XYA_2BX. Optimised stop facilities are also available.

22.5.3 Building management systems

Control systems for building services have moved steadily from stand-alone control loops towards arrangements where a high degree of central supervision is provided—largely automatic, with manual override applied for specific circumstances. There has been a progression from electromechanical supervisory data centres, through electronic building automation systems to the current energy management (EMS) and building management (BMS) systems which depend on microprocessor/software technology. Energy management is just one, albeit important, of the facilities offered by a BMS, which acts as the controller,

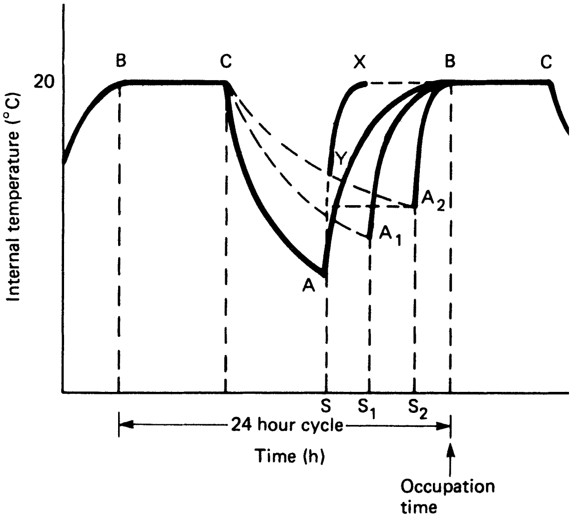


Figure 22.25 Optimised start of heating plant

monitor and fault locator for all aspects of the built environment and can provide a wide range of other management functions.

The latest generation of microprocessor-based systems are now used to replace the traditional stand-alone controllers used in control loops and the means of control has passed from analogue to direct digital control (DDC). On any but the smallest BMS (which can be a single programmable controller with a number of analogue and digital inputs and outputs), the equipment configuration will be similar to that shown in the block diagram of *Figure 22.26*. The data-gathering panels (DGP) in modern systems will be microprocessor based and incorporate levels of software which make them largely independent of the central processor. The main criteria and possible uses are listed briefly below.

Data transmission This may be multiplexed multicore distribution, single or two wire trunks for pulse coded messages, or a fibre optic system. Where remote buildings are coupled to one system British Telecom or Mercury lines may be used for transmission.

Scanning Typically scan times are between 2 and 30 s, although the point-to-point scan may be much faster. Multiple scan times can be employed, one for analogue signals and others for high- and low-priority alarms.

Hardware and peripherals These include the following:

(1) data inputs from two-position and analogue devices, and data outputs for control switching and set-point adjustment;
(2) outstations, which may be relatively simple data processors or intelligent systems, with stand-alone capability;
(3) intercom, which may be a feature additional to the transmission system;
(4) central processor, which contains the memory for automation and alarm functions, often with a back-up power supply;
(5) operator's keyboard and display unit;
(6) printer(s) for common or separate logs and alarms;
(7) visual display units in monochrome or colour;
(8) permanent displays such as annunciator panels or mimic diagrams.

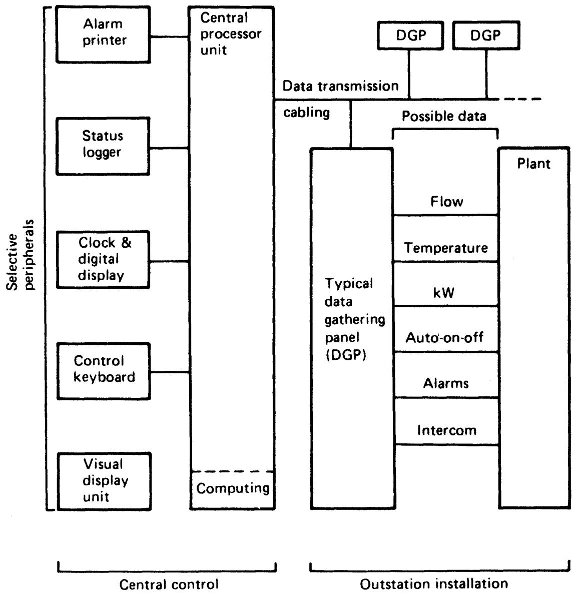


Figure 22.26 Block diagram of a typical building automation system. DGPs will normally include their own software capability for control functions. The central processor unit may range from a main frame machine to a microprocessor unit

Software This covers the following:

(1) alarm priorities;
(2) alarm inhibiting;
(3) analogue alarms;
(4) integration, e.g. energy consumption;
(5) totalisation, e.g. summation of motor run times, etc;
(6) time switch, including optimised start;
(7) event initiated sequences, e.g. an alarm which initiates a specific sequence of operations;
(8) load shedding and lighting control;
(9) restart after power failure: prevents electrical overload on restart;
(10) process control, e.g. the use of the centralised system as the controller for individual loops (DDC);
(11) optimum damper control (see Section 22.5.1.3);
(12) security, e.g. patrol tours and card entry;
(13) interlocking, e.g. the use of software in place of conventional relays, timers, etc;
(14) fire, i.e. alarms and specific event initiated sequences;
(15) programmed maintenance, i.e. the use of stored data to produce a work schedule for maintenance and service;
(16) facilities management.

An example in block form of a building automation system is shown in *Figure 22.26*.

Economics The cost evaluation for a proposed scheme should include consideration of the following:

Savings	*Expenditure*
Reduction of energy	Capital cost of system
Reduction in maintenance staff	Interest on capital cost Additional specialist staff
Increased plant life with programmed maintenance	Maintenance of the automation system
Use of software for interlocking in place of relays, timers, etc.	Preparation of detailed programmed maintenance format
Avoidance of duplicate systems (e.g. fire and security central consoles)	Collection of data and producing particular software

22.6 Energy conservation

The subject of energy conservation and the efficient use of energy is of interest in all branches of engineering design. The provision of satisfactory environmental conditions consumes such a large fraction of the national energy budget that energy conservation cannot be divorced from comfort. The means of reducing energy consumption range from using less energy by more effective building design and more efficient HVAC systems, to lowering comfort standards and reclaiming energy that would otherwise go to waste. Lowering comfort standards is a difficult objective to promote. Heat reclaim methods are numerous and in many cases are applicable also to industrial processes, and they often improve the efficiency of existing systems.

There is also the problem of effective maintenance and servicing of existing installations, which can contribute to both energy conservation and satisfactory environmental comfort standards.

22.6.1 Systems

The importance of heat reclaim systems has always been recognised but in practice their application has been very restricted until recently. Because a heat recovery system is frequently part of a sophisticated installation, it should be integrated into the more comprehensive requirements of the design philosophy. Some of the more common systems are described below.

Thermal wheel These are rotating air-to-air heat transfer devices between two separate air streams in parallel and adjoining ductwork. The speed of rotation will not normally exceed 20 rev/min and the heat recovered decreases with speed. The control of energy transfer is effected by varying the speed or the exhaust air quantity passing through the wheel. Normally, only sensible heat is recovered, but versions exist which reclaim latent heat as well. A standard arrangement is shown in *Figure 22.27*. The temperature of the air supplied to the space (or other elements of the plant) is controlled by varying the speed of the drive motor. Control can also be achieved by bypass dampers to reduce the air passing through the exhaust air section. It is necessary to check that the energy saved by heat transfer is not exceeded by the additional fan power required.

Liquid coupled indirect heat exchanger (run-round coil) This system is simple. The general arrangements are shown in *Figure 22.28*. The pump may be controlled by an externally mounted thermostat T_1, which runs the pump whenever the external temperature is below the design exhaust air temperature. Alternatively, a more sophisticated arrangement would be to use a differential thermostat T_{2E}, T_{2S} which runs the pump only when the temperature of the exhaust air is higher than that of the supply air. The capital cost of the plant and the additional pump and fan horsepower must be equated to the energy saved before such a system is adopted.

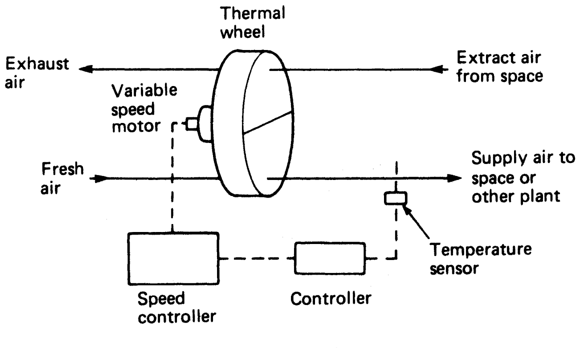

Thermal wheel
Exhaust air
Extract air from space
Variable speed motor
Fresh air
Supply air to space or other plant
Temperature sensor
Speed controller
Controller

Figure 22.27 Variable speed control of thermal wheel

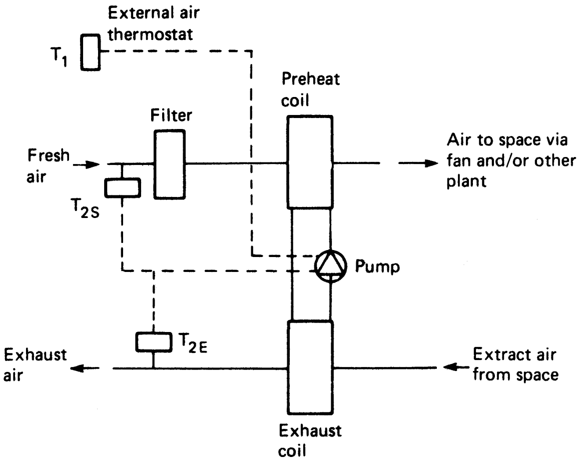

External air thermostat
T_1
Preheat coil
Filter
Fresh air
Air to space via fan and/or other plant
T_{2S}
Pump
Exhaust air
T_{2E}
Extract air from space
Exhaust coil

Figure 22.28 Scheme of a heat-reclaim run-round coil. T_1, T_{2s} and T_{2e} are detectors for different methods of control

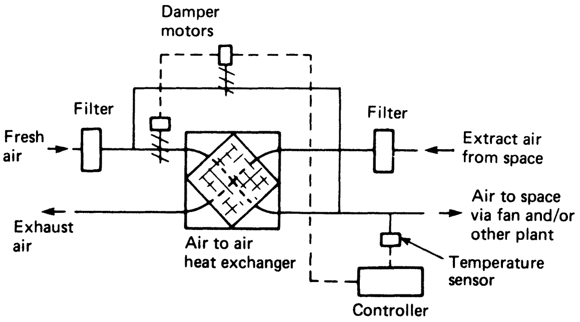

Damper motors
Filter
Filter
Fresh air
Extract air from space
Air to space via fan and/or other plant
Exhaust air
Air to air heat exchanger
Temperature sensor
Controller

Figure 22.29 Scheme of heat reclaim by an air-to-air heat exchanger

Cross-flow stationary recuperator (air-to-air heat exchanger) This device is an alternative to the thermal wheel but will provide only sensible heat transfer. The general arrangement is shown in *Figure 22.29*. The control system is similar to the bypass damper arrangement described for the thermal wheel.

Heat pump The use of heat pumps in HVAC systems is increasing. The most common form incorporates a refrigerant compressor unit with evaporator and condenser where the functions of the latter two elements may be reversed to give heating or cooling cycles as required. This arrangement is shown in *Figure 22.30*, where the heating cycle mode illustrates the generic definition of a heat pump system. This type of unit is described in Section 22.4.5.

Heat pumps are frequently used purely for heating, but whether in this mode or as reversible units, they are available with air-to-air, water-to-water and air-to-water heat transfer. The selection of refrigerant is important to ensure that maximum efficiency is achieved for the specified range of inlet and outlet temperatures.

In some heating applications, particularly where heat is being extracted from outside air, the external coil (acting as the evaporator) will tend to ice up and a defrost control system must be used. This requires an arrangement for

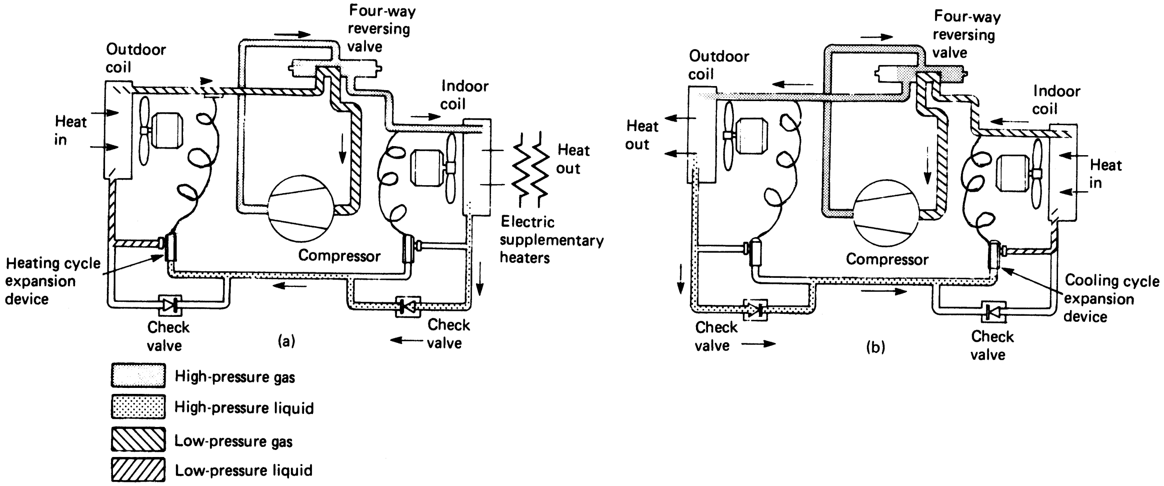

Four-way reversing valve
Outdoor coil
Indoor coil
Heat in
Heat out
Electric supplementary heaters
Heating cycle expansion device
Compressor
Check valve
Check valve
(a)
High-pressure gas
High-pressure liquid
Low-pressure gas
Low-pressure liquid
Four-way reversing valve
Outdoor coil
Indoor coil
Heat out
Heat in
Compressor
Cooling cycle expansion device
Check valve
Check valve
(b)

Figure 22.30 Reversible cycle units. (a) Heating cycle (heat pump); (b) Cooling cycle (chiller)

allowing hot gas to be passed through the coil for a short period, or the use of separate electric heaters.

Heat pipe This simple device is becoming an accepted heat reclaim component; it contains no moving parts and can be fitted in a manner analogous to a thermal wheel. Essentially, it consists of a sealed and evacuated tube containing a refrigerant, e.g. Freon, and lined with a wick. The action is shown in *Figure 22.31*: (*a*) shows that the application of heat vaporises the liquid refrigerant, which is then cooled at the top of the tube (giving up its latent heat), absorbed in the wick and returned to the bottom of the tube. This system is utilised in heat reclaim as in (*b*), where banks of tubes replace the thermal wheel in *Figure 22.27*.

Double-bundle condenser One common method of heat reclaim normally considered in all air conditioning systems is the double-bundle condenser. *Figure 22.14* shows the conventional condenser, but it is obviously advantageous to use the rejected heat rather than discard it in the cooling tower. The arrangement is shown schematically in *Figure 22.32*, where one tube bundle is used for the tower heat rejection circuit and the other for heat reclaim to the heating systems in the building. The system is controlled in sequence so that the tower circuit is brought into use only when there is no requirement for reclaimed heat. Temperatures of 45 °C can be attained from the reclaim circuit without difficulty.

22.7 Interfaces and associated data

Modern buildings continue to become more complex and sophisticated, a fact which is generally accepted. Whether all the complexities are necessary is discussable, but not relevant here. What does need clarification is the treatment of the various interfaces which occur between all the parties involved in the building process starting with the developers/designers and ending with the tenants. If these interfaces are not clearly identified and treated, the environment and its control in the final building will be unsatisfactory and inefficient.

The first stage in this process is recognition that all the design disciplines should be involved with the client from

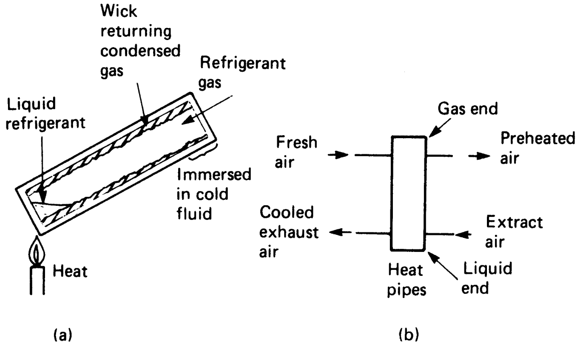

Wick returning condensed gas
Refrigerant gas
Liquid refrigerant
Immersed in cold fluid
Heat
(a)
Gas end
Fresh air
Preheated air
Cooled exhaust air
Extract air
Heat pipes
Liquid end
(b)

Figure 22.31 Heat pipes for heat reclaim

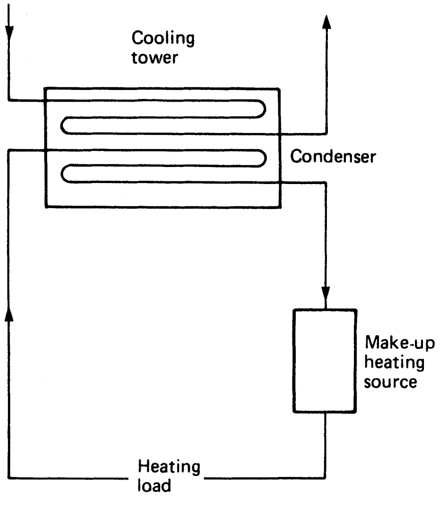
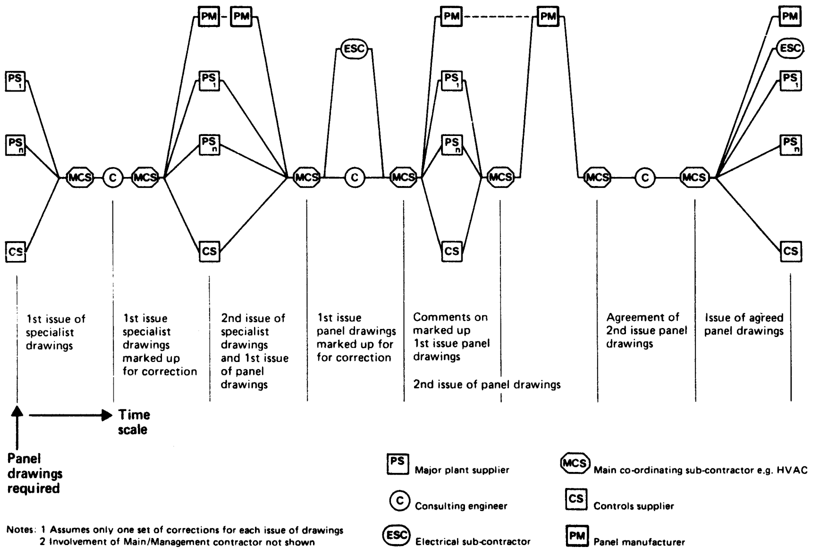

Cooling tower
Condenser
Make-up heating source
Heating load

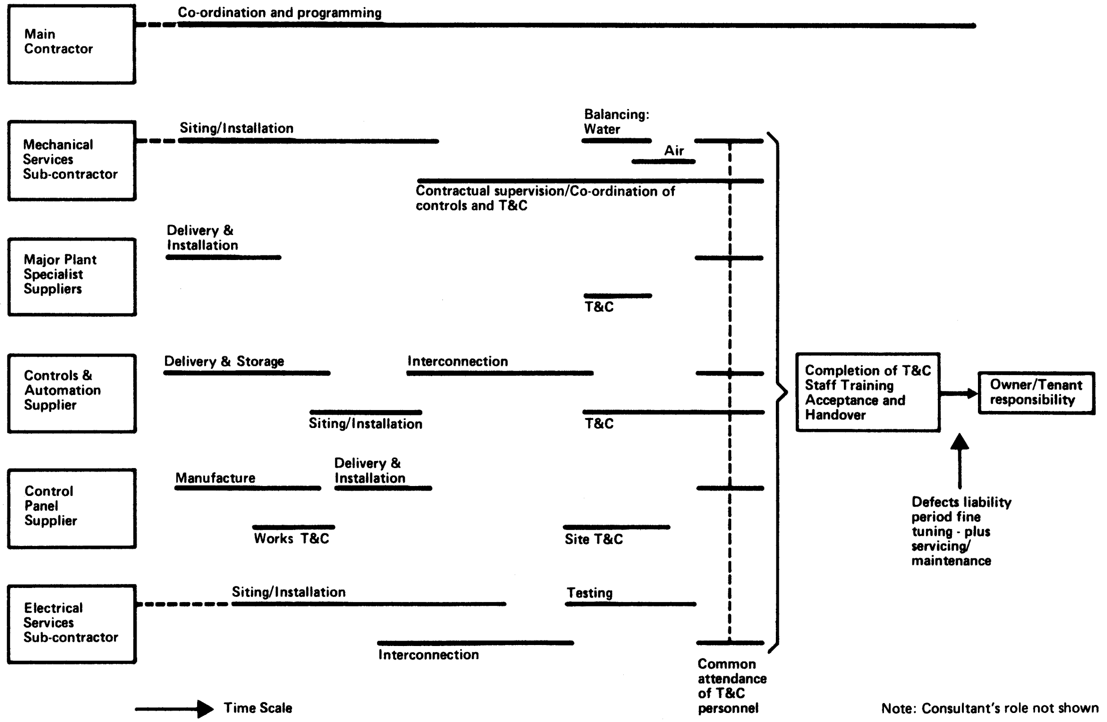
Figure 22.32 Basic scheme of a double-bundle condenser for heat reclaim

BASIC CONTROL PANEL WIRING DIAGRAM PROGRAMME


PS 1
PS n
CS
MCS
C
MCS
PM
PM
PS 1
PS n
CS
MCS
ESC
C
MCS
PM
PM
PS 1
PS n
CS
MCS
MCS
C
MCS
PM
ESC
PS 1
PS n
CS
1st issue of specialist drawings
1st issue specialist drawings marked up for correction
2nd issue of specialist drawings and 1st issue of panel drawings
1st issue panel drawings marked up for for correction
Comments on marked up 1st issue panel drawings
2nd issue of panel drawings
Agreement of 2nd issue panel drawings
Issue of agreed panel drawings
Time scale
Panel drawings required
PS Major plant supplier
MCS Main co-ordinating sub-contractor e.g. HVAC
C Consulting engineer
CS Controls supplier
ESC Electrical sub-contractor
PM Panel manufacturer

Notes: 1 Assumes only one set of corrections for each issue of drawings
2 Involvement of Main/Management contractor not shown

BASIC SITE CO-ORDINATION FOR CONTROLS AND T&C


Main Contractor
Co-ordination and programming
Mechanical Services Sub-contractor
Siting/Installation
Balancing: Water
Air
Contractual supervision/Co-ordination of controls and T&C
Major Plant Specialist Suppliers
Delivery & Installation
T&C
Controls & Automation Supplier
Delivery & Storage
Interconnection
Siting/Installation
T&C
Control Panel Supplier
Manufacture
Delivery & Installation
Works T&C
Site T&C
Electrical Services Sub-contractor
Siting/Installation
Testing
Interconnection
Common attendance of T&C personnel
Completion of T&C Staff Training Acceptance and Handover
Owner/Tenant responsibility
Defects liability period fine tuning - plus servicing/ maintenance
Time Scale

Note: Consultant's role not shown

Figure 22.33 Examples of interfaces and co-ordination

the concept stage onwards. Very frequently architects are the only discipline initially involved, followed by the structural engineers and then the mechanical and electrical (building services) engineers. The ultimate tenants and their requirements are all too rarely involved until after the design is completed. With the complexity of modern buildings, the increasing pressures of environmental issues and the multiplicity of interfaces, it is essential that design teams are multidisciplinary in nature (with builder and tenant involvement if possible) from the very inception of the project. Just two examples of how these interfaces affect the way environmental control is achieved are shown in *Figure 22.33*. The first shows the quickest route to obtain approved control panel diagrams and at best it takes many weeks. The second shows how the setting up and commissioning of the environmental controls and automation systems has to be planned.

The interface requirements of the systems shown in *Figure 22.33* refer to well-defined elements in the environmental-control process. When the considerations have to include the more subjective environmental control issues of the greenhouse effect, ozone depletion by chlorofluorocarbons (CFCs), internally/externally created pollutants and the physiological and psychological reactions of occupants to their working environment, the interface problems can be overwhelming. Only a co-ordinated approach by all concerned can achieve the necessary environmental control.

22.7.1 Electrical loads

Environmental control and issues are often inextricably related to energy and its effective use. Mechanical and electrical engineers working on the building services of buildings, often worth half the total cost of the building, have an extremely important role in minimising energy consumption. At one level they need to work with the architect to ensure an efficient fabric design and at the other they need to define electrical loads at the concept stage so that transformers and switchgear can be sized and suitable space provisions made.

Electrical-load figures are supplied in many forms, and at the concept stage in the design process considerable expertise is required to correctly assess suitable figures which will generally withstand the various changes during the design process. This has to be borne in mind when considering *Figures 22.34* and *22.35*. The curves in *Figure 22.34* represent typical maximum demands for fans and pumps for different types of air-conditioning system, while those in *Figure 22.35* indicate maximum demands for different types of chiller. Lighting load data are voluminous and are best obtained from CIBSE Lighting Division publications, or manufacturers.

References

1 *CIBSE Guide Book A*, Section A1, CIBSE, London (1986)

2 *Control of Fuel and Electricity, The Fuel and Electricity (Heating Control) Order—Statutory Instrument*, 1974 *No. 2160* (1974)

3 *Health and Safety. The Control of Substances Hazardous to Health Regulations (COSHH)—Statutory Instrument 1988*, *No. 1657* (1988)

4 WILSON, S. and HEDGE A., *A Study of Building Sickness,* Building Use Studies Ltd, London (1987)

5 WILSON, S., O'SULLIVAN, P., JONES, P. and HEDGE, A., *Sick Building Syndrome and Environmental Conditions*, Building Use Studies Ltd, London (1987)

6 RAW, G., *Sick Building Syndrome*. The review of the evidence on causes and solutions. HSE Contract Research Report No. 42/1992

7 *CIBSE Guide Book A*, Section A3, CIBSE, London (1986)

8 *CIBSE Guide Book A*, Sections A5 and A9, CIBSE, London (1986)

9 *Fuel Economy Handbook*, National Industrial Fuel Efficiency Service, Graham and Trotman, London (1980)

10 *Applied Air Conditioning and Refrigeration*, 2nd edition, G. T. Gosling, Applied Science

11 *Controls for Domestic Central Heating Systems*, British Gas (October 1978)

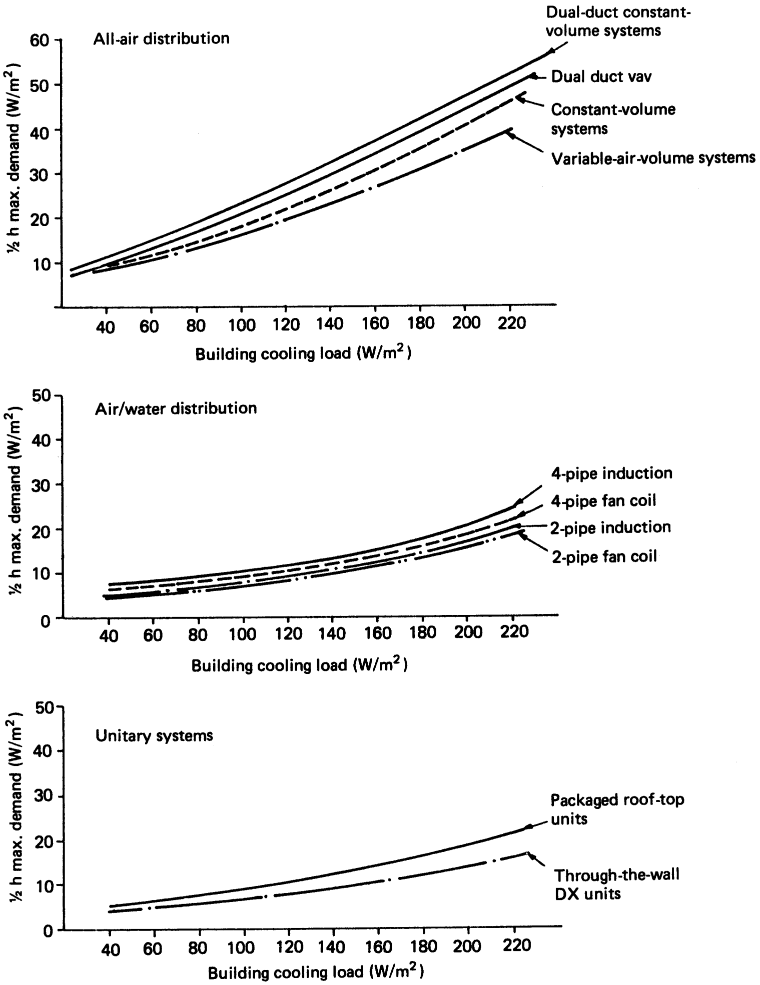


Figure 22.34 Maximum electrical demand for fans and pumps. VAV=variable air volume; DX=direct expansion. (Courtesy of Ove Arup Partnership)

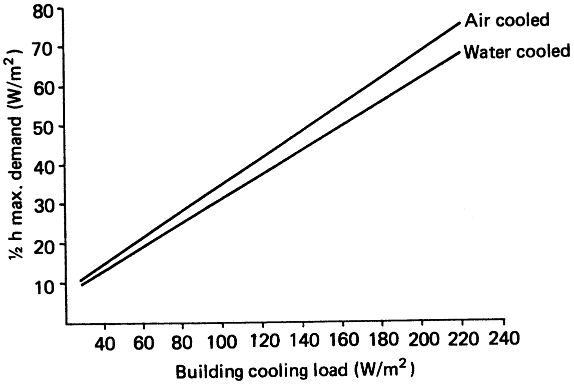


Figure 22.35 Full-load maximum demand characteristics for water- and air-cooled chillers. (Courtesy of Ove Arup Partnership)

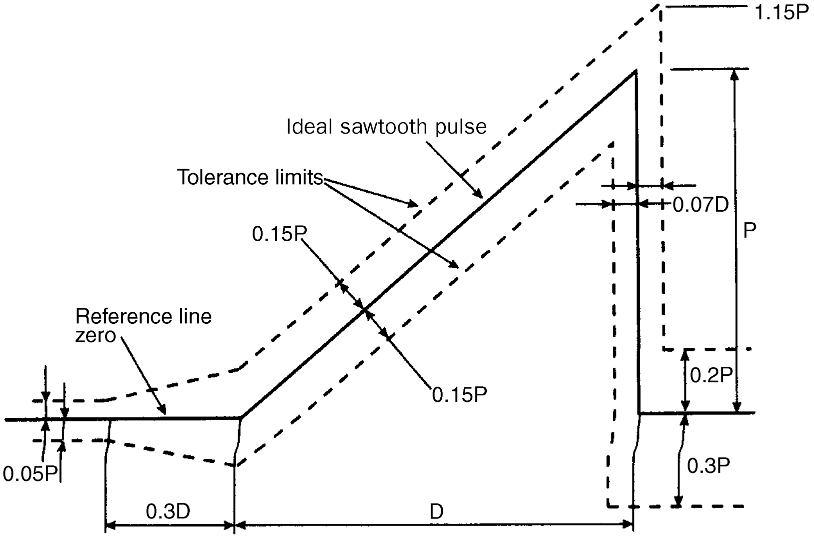

1.15P
Ideal sawtooth pulse
Tolerance limits
0.15P
0.07D
P
Reference line zero
0.15P
0.2P
0.3P
0.05P
0.3D
D

Figure 22.36 Shock pulse configuration and its tolerance limits (courtesy of Ove Arup Partnership)

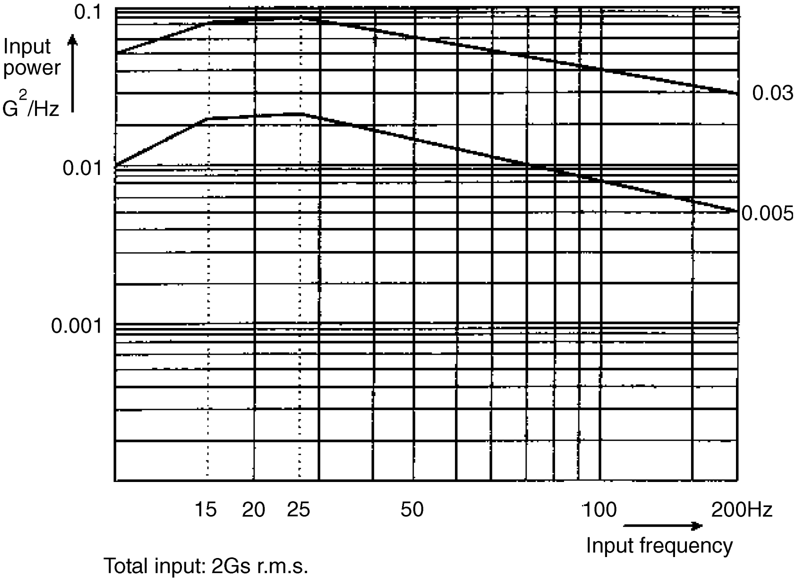


Figure 22.37 Power spectral density graph for random vibration testing

23

Electromagnetic Compatibility

A Maddocks

Contents

23.1 Introduction 23/3

23.2 Common terms 23/3

23.3 The EMC model 23/3

23.4 EMC requirements 23/5

23.5 Product design 23/6

23.6 Device selection 23/6

23.7 Printed circuit boards 23/6

23.8 Interfaces 23/7

23.9 Power supplies and power-line filters 23/7

23.10 Signal line filters 23/8

23.11 Enclosure design 23/8

23.12 Interface cable connections 23/9

23.13 Golden rules for effective design for EMC 23/11

23.14 System design 23/11

23.15 Buildings 23/13

23.16 Conformity assessment 23/13

23.17 EMC testing and measurements 23/14

23.18 Management plans 23/15

23.1 Introduction

Electromagnetic Compatibility (EMC) has now become a major consideration on any project involving the design, construction, manufacture and installation of electrical and electronic equipment and systems. Electrical equipment must be designed not only to meet a functional technical performance specification but due consideration must also be given to the interaction the equipment has with the electromagnetic environment in its intended operating location. If the equipment is expected to operate reliably in a steel works, for example, it is imperative that the designers, purchasers, installers and operators are aware of the nature of the electromagnetic environment and the potential for unwanted coupling to the equipment which could cause equipment mis-operation or malfunction. For example, for safety-related equipment, any interference to the operation of the system could have serious consequences. Equally, electromagnetic disturbance generated by the equipment itself could cause interference to radio reception. In the case of domestic radio and TV reception this may adversely affect the quality of reception, but can also block emergency channels and in some cases, e.g. a radio controlled crane, could cause malfunction with potential reliability and safety implications. Generally, both aspects of controlling emissions from the equipment and providing adequate immunity to the expected electromagnetic disturbances in the intended operating environment are key features of any equipment or system design.

The need for emission control has been recognised for many years and most countries have introduced legal regulations to support the efficient utilisation of the electro magnetic spectrum. This is manifest in the form of requirements to comply with EMC emission standards in product certification. In some territories such as Europe and Australia, the legal requirements also address immunity aspects. For most manufacturers of good quality equipment this is not an added burden because they are keen to demonstrate to their customers that the equipment will prove to be reliable in the field, and moreover, recognise that any actual incidents of malfunction could be damaging and costly to the business.

Fortunately there are advantages for both equipment suppliers, purchasers and operators in the availability of nationally or internationally approved high quality EMC standards for both emissions and immunity which can be referenced in contractual agreements as well as providing an excellent foundation for the designers of the equipment. Manufacturers and users alike can apply these standards to their mutual advantage, particularly in the planning of major projects.

EMC considerations need to be addressed at the outset of the development. It is well known that the costs of achieving EMC conformity rise almost exponentially with the delay in first considering the requirements. The key issue in addressing EMC matters is to adopt a strategic approach where the EMC requirements are recognised and clearly understood at the design concept stage, through the product development to in service use, and throughout the lifetime of the equipment.

23.2 Common terms

There are a number of common terms used in the science of EMC. EMC itself can be defined as the ability of equipment to operate satisfactorily within its intended electromagnetic environment without contributing to the disturbance level in that environment such that radio communication is not adversely affected. Other related terms are RFI and EMI: RFI is radio frequency interference, which is usually defined as interference to radio services in the radio bands, 9 kHz to 300 GHz; EMI, electromagnetic interference, is generally accepted as interference both in the radio frequency bands and in the low frequency region d.c. to 9 kHz.

Because of the large ranges of values that are dealt with in EMC it is customary to express emission limits and system performance in logarithmic ratios, i.e. in dBs. For a voltage or current value, the value in dBs is given by $20 . \log_{10}$ (ratio) and for power ratios, $10 . \log_{10}$ (ratio). It is important to recognise that 1 dB in voltage is equivalent to 1 dB in power, although the linear ratios are different, 1.12 and 1.26 respectively.

Electric and plane wave fields are expressed in Volts/metre or dB(μV/m) and magnetic fields are expressed in Amps/metre or dB(μA/m). (1 Amp/metre is equivalent to 1.25 micro Tesla of magnetic induction in free space.)

23.3 The EMC model

An example of electromagnetic coupling between two items of electrical equipment is illustrated in *Figure 23.1*. Electrical disturbances generated by equipment #1 may be coupled to equipment #2 by a variety of means: conduction via a common connection to the mains supply, inductive and capacitive coupling between interface and power conductors, and by direct radiation. Conducted coupling tends to be the dominant coupling mechanism at lower frequencies e.g. below 1 MHz, where conductor impedances are low; capacitive and inductive coupling is more important at higher frequencies where the capacitive impedance between long parallel runs of conductors is relatively low. Radiation coupling dominates at frequencies where the length of the radiating conductor is comparable with a wavelength. For a small computer system, radiation from cables will be prominent in the range 30–300 MHz but at higher frequencies, direct radiation from circuit board tracks dominates. Due consideration must be given to these effects in designing and installing equipment and systems.

There are many types of coupling that may occur in any particular application and the key factor for the designer is to recognise and understand the various mechanisms. The simple but effective model source-path-receptor model shown can be applied effectively in dealing with overall requirements (*Figure 23.2*).

For the equipment designer, his product must be considered both as a source of disturbance and as a potential receptor to disturbances in the intended environment. Emission control is achieved by recognising the potential sources of disturbance within the equipment and the paths by which they may couple to the outside world and cause interference in radio communications. Examples of significant sources of internally generated electromagnetic disturbance are given in *Table 23.1*.

Emission control is achieved by effective design, filtering and suppression measures. The level of control that is normally required is that sufficient for interference free reception of radio communication and radio and TV broadcast services.

But the designer of electronic equipment must also consider his product as a receptor of electromagnetic disturbance in the intended operating environment. For the equipment to work reliably in the field the designer

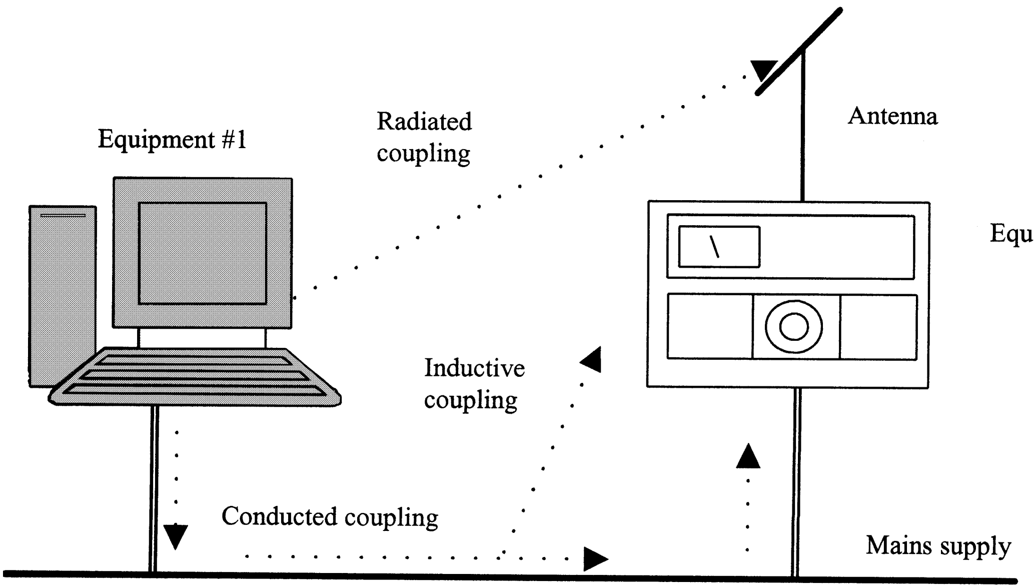


Figure 23.1 Example of electromagnetic coupling between two items of equipment

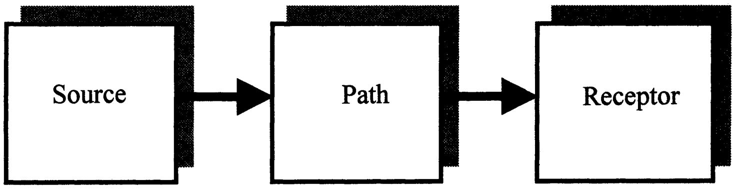


Figure 23.2 Source–Path–Receptor model

must consider the types and nature of electromagnetic disturbances which are likely to be present, their magnitude, and even their probability of occurrence if it is a safety related system. Generally there are two categories, man made and naturally occurring, as shown in *Tables 23.2* and *23.3*.

For radio transmitters and overhead power lines, the level experienced by the receptor circuit will depend on the separation distance from the source. Some VHF and UHF rf transmitters are designed so that the field strength at ground level does not exceed 1 Volt/metre but mobile and portable transmitters such as mobile phones can generate field levels of up to 100 Volt/metre close to the antenna. In cases where the equipment is intended to be operated close to one of these sources, compliance with commercial EMC standards may be insufficient and additional measures may be required, either in the equipment design or the installation, to ensure that the equipment will operate reliably in practice.

The electrostatic discharge is a high voltage disturbance, typically 8 kV, with a very short rise time, about 1 nanosecond, and generates high level spectral components extending to high frequencies and careful equipment design is often required to provide adequate protection. The residual effects of a lightning event are high voltage surges of about 2 kV but with longer durations and much greater energy.

The degree to which electronic equipment may be affected by the electromagnetic disturbances in the environment depends on many factors but is primarily dependent on: a) the coupling between the equipment, b) the magnitude of the external source, and c) the sensitivity of the internal electronics. In broad terms, any semi-conductor device in the equipment must be considered as a potential receptor.

Most digital electronic devices generally require a signal of at least 0.5 Volt to change state and for protection against ESD with a peak amplitude of 5 kV, the designer must provide a circuit isolation in excess of 10 000 to one or 80 dB. The first 60 dB may be relatively easy to achieve through natural cicuit losses but the next 20 dB will inevitably require careful design. Analogue semiconductor devices have higher sensitivities with devices in the control application requiring only millivolts of disturbance to cause interference. Here the main concern is the continuously applied or long duration disturbance such as from radio transmitters. An applied field of 3 Volt/metre could induce a voltage of 3 Volts into a conductor connected to a remote analogue sensor causing an rf current of up to 10 mA to flow in the wiring. To reduce this level of disturbance to acceptable proportions high degrees of isolation, possibly in the order of 60–70 dB are required. Most of the coupling at high frequencies is in common mode and there will be an inefficient coupling to a differential voltage that could be confused with the wanted signal, but inevitably some form of circuit protection, often in the form of filtering will be required.

Table 23.1 Man made sources in equipment

Source or device	*Electromagnetic disturbance*
Digital electronic circuits	Harmonics of clock oscillators
Commutator motors	High frequency switching transients
Contact devices	Showering arc discharges
RF oscillators	RF fields and voltages
Luminaires and Lighting equipment	Arc discharges
Switch mode power supplies	Harmonics of the switching frequency

Table 23.2 Man made sources in the environment

Source or device	*Electromagnetic disturbance*
Radio transmitters	Rf fields
Power distribution	Surges, fast transients, dips and interruptions
Overhead power lines and railway traction	Magnetic and fields, corona discharges

Table 23.3 Naturally occurring sources in the environment

Source or device	*Electromagnetic disturbance*
Human body electrostatic charging	Electro-static discharges
Lightning	Fields and power surges

23.4 EMC requirements

The overriding requirement in the supply of equipment to the customer, or in placing the equipment on the market, is to meet the accepted or agreed conformity assessment requirements. For supply of defence equipment for example, the manufacturer will be required to demonstrate that the equipment complies with the EMC specification for the project, such as DEF STAN 59-41 in the UK. The manufacturer ensures that the equipment is tested to the standard and submits the test report to the project office for approval. In the case of equipment for residential, commercial or industrial application, emission control regulations apply in most territories, and compliance with a relevant EMC standard must be demonstrated. In Europe, Member States have transposed the provisions of the EMC Directive 89/336/EEC into their own legislation. For the vast majority of equipment types, the manufacturer can apply a relevant standard, which is one harmonised by the Member States and published in the *Official Journal* of the EC, and make a self declaration of conformity also applying the CE Mark. It should be noted that the CE Mark denotes compliance with all applicable directives which are likely to include at least the Low Voltage Directive in addition to the EMC Directive. Market entry and free circulation of the equipment within the EC is then permitted.

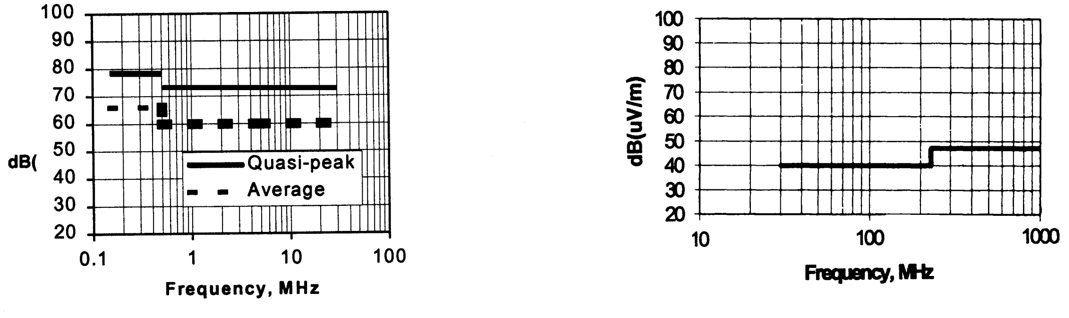
The relevant EMC standards are of three types, high frequency emissions standards for the protection of the radio spectrum, low frequency emissions standards for the protection of the power distribution network, and immunity standards for demonstrating the equipment's robustness in the presence of electromagnetic disturbances. Typical emission limits, in this case for Information Technology Equipment are presented in *Figure 23.3*(*a*) and (*b*).

Limits for equipment intended for residential application are generally more severe by 10 dB because of the shorter separation distances and tighter coupling that can occur in domestic premises.

Standards are available for the protection of the power distribution network. Mains harmonics such as generated by high power semiconductor switching devices can cause problems in the supply of electricity for which limits apply for the frequency range 0–2 kHz. Voltage fluctuations due to switching of heavy loads, which can cause lighting flicker is another low frequency phenomenon for which relevant standards exist. For example, equipment containing switch mode power supplies or switching heating loads as in air handling units will be subject to the standards listed in *Table 23.4*.

The European immunity requirements are primarily based on international IEC standards. Both product specific, and generic standards (to be applied where no product standard exists), are based on a raft of basic standards addressing fundamental EMC phenomena. These are listed in *Table 23.5* together with typical levels for a commercial or light industrial application.

The descriptions for these two environments may be insufficient in some applications. For example if the equipment is to be installed in an airfield environment, immunity may be required against radar transmitter fields of up to 100 V/m and these requirements would need to be incorporated in the technical specification for the project.



a) Conducted emissions limits

b) Radiated emissions limits at 10 m

Figure 23.3 Harmonised European conducted and radiated emissions limits for ITE (EN 55022) Class A

Table 23.4 Low frequency emissions

EMC phenomenon	*European Basic Standard*
Mains harmonics	EN 61000-3-2
Voltage fluctuations (flicker)	EN 61000-3-3

23.5 Product design

In designing equipment for EMC conformity it is important to identify the primary paths whereby electromagnetic energy may be transferred. These are conduction and induction via the power lead and interface cables, radiation from the PCB circuit tracks and radiation leakage through apertures and discontinuities in any shielded enclosure. The aim of designing for EMC conformity is to provide sufficient attenuation in each of the relevant paths such that emission control and provision of adequate immunity is achieved at an economic cost to the product. For many protection measures e.g. shielding and filtering there is reciprocity between measures for emission control and for immunity, although the required degree of attenuation in a particular path may be different. Experience with previous generations of similar products and some exploratory work in determining the immunity margins by testing to the level of equipment misoperation, can provide valuable information as what is the more critical, aspect emissions or immunity. Further development work can then concentrate on testing to only the critical EMC phenomenon.

For equipment contained in metal enclosures which can provide a reasonable degree of shielding, there are a number of important considerations. As shown in *Figure 23.4*, the barrier for the power cable is a filter mounted at the point of entry for the cable.

Similarly, unshielded signals lines to inputs and outputs entering the enclosure should also be filtered at the point of entry. Where screened cables are employed, usually for the protection of sensitive circuits, or for emission control for high frequency signals, the cable screen should be regarded as an extension of the shield of the enclosure.

Where metal enclosures are not employed, the transfer of high frequency electromagnetic energy must be controlled by good design at the circuit board level. The use of multiplayer PCBs offers such a facility but sometimes there appears to be a cost penalty associated with the more complex structure and fabrication. These costs should be compared with the costs of retrospective remedial measures should the equipment fail to comply with radiated emission and radiated immunity requirements. Generally, most designers accept that the multi-layer board provides a good margin of compliance if properly configured and is the preferred option.

23.6 Device selection

The selection of the type of semiconductor device or integrated circuit technology can have an effect on overall EMC performance. Although the trend is for faster devices for higher processing speeds, emission control is easier with slower devices. *Figure 23.5* shows the waveform of a digital signal and its associated spectrum. The spectrum starts to decay at 20 dB/decade at just below $1/(\pi\varsigma T)$ where $1/T$ is the repetition rate, but decays even faster at frequencies above $1/(\pi\varsigma t_r)$ where t_r is the rise-time. Lower clock rates and longer rise times thus reduce high frequency emissions.

For improved immunity, circuit bandwidth is important, the narrower the bandwidth, the lower the energy from impulsive disturbance and there is a greater probability of discriminating against high frequency single frequency disturbances. Although device selection may not be a prime consideration in designing for EMC conformity because of performance constraints, knowledge on the basic principles can be extremely valuable.

23.7 Printed circuit boards

Tracks on circuit boards can act as antennas and can radiate signals efficiently at high frequencies. There are essentially two basic mechanisms, monopole and loop type.

Where for example the signal from a high frequency clock oscillator is taken to another device on the circuit board and the return conductor is not immediately adjacent to the signal line, a loop is formed which will radiate the signal and generate a field at a remote point (*Figure 23.6*(a)). Taking the example of TTL logic for a 1 mA signal at 200 MHz in a 3 cm^2 loop the field at 10 m would equal the limits of EN 55022 Class B. The areas of all such loops are to be minimised for lowest radiated field. Similarly, immunity to rf radiation and possibly other types of disturbance such as indirect electrostatic discharges (ESD) would also be improved by this measure.

However a more important mechanism is where a voltage is generated across the length of the ground return on the

Table 23.5 Immunity levels

EMC phenomenon	*European basic standard*	*Typical disturbance level a.c. power port*	
		Residential, commercial and light industrial environment	*Industrial environment*
Electrostatic discharge	EN 61000-4-2	8 kV (air discharge)	8 kV (air discharge)
Radio frequency field	EN 61000-4-3	3 V/m	10 V/m
Fast transients	EN 61000-4-4	1 kV	2 kV
Surge	EN 61000-4-5	1 kV	2 kV
Common mode rf voltage	EN 61000-4-6	3 V	10 V
Power frequency magnetic field	EN 61000-4-8	3 A/m	30 A/m
Dips and interruptions	EN 61000-4-11	30% for 10 ms 60% for 100 ms	30% for 10 ms 60% for 100, 1000 ms

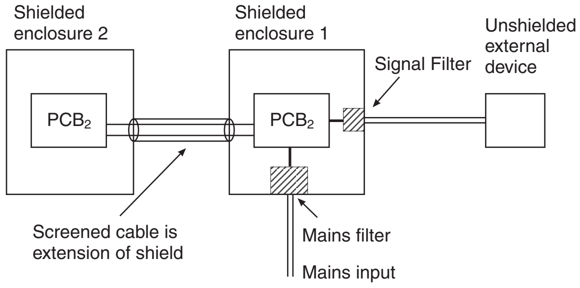


Figure 23.4 Design concepts

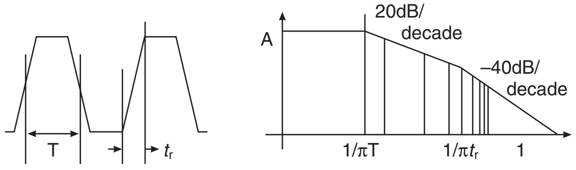


Figure 23.5 Waveform and spectrum of a digital signal

board where the voltage drives into an attached wire or cable causing it to radiate as a monopole (*Figure 23.6*(b)). This is a far more efficient radiation mechanism and can give rise to high emission levels. The key factor is to ensure that the cable shield termination is connected to the metalwork of the enclosure and not to the zero volt conductor/plane of the printed circuit board.

Multiplayer PCBs are effective at achieving the design aim of reduced loop area and close tracking. The ground plane acts as the high frequency earth return for all signals and all the high frequency energy is effectively contained within the layer of insulations. Compared with a single sided board, EMC performance is improved by up to 30 dB by this measure.

23.8 Interfaces

The multi-layer PCB circuitry communicates with the outside world through attached shielded interface cables and it is important to minimise the amount of high frequency energy such as high order clock harmonics from exiting the equipment via the interface. One method of control (Reference 1) is to isolate the interface connectors on a piece of 'clean' earth ground plane having only a single point of connection to the main part of the board.

The clean interface ground is bonded to the equipment chassis at the cable port; the decoupling capacitors are thus far more effective than they would be if terminated to the noisy ground of the main part of the pcb.

Digital signal interfaces may require additional protection against incoming transient disturbances particularly where the interface cable is unshielded or the pins are exposed to ESD. Here transient suppression in the form of gas tubes and Transorbs may applied to good effect, provided a good low impedance earth is available.

23.9 Power supplies and power-line filters

Many power supplies are of the switch mode types. These are well-known sources of high levels of high frequency disturbance, comprising harmonics of the switching frequency and cover the spectrum up to 30 MHz. These power supplies usually contain well designed high performance filters which attenuate the internally generated disturbance to levels below the most stringent emission limits. Any disturbance generated by the electronic circuit powered by the SMPSU is also attenuated strongly and there is usually no requirement for additional power line filtering for either emissions control or immunity.

The power supplies are usually supplied compliant to national and international EMC standards and the selection of the power supply can then be based solely on functional requirements.

Where a separate power line filter is installed in equipment a single stage filter can often be selected by choosing the largest line inductance at the rated current. However, overall EMC performance is generally determined by EMC tests on the complete system because the practical terminating impedances are usually unknown for all the modes of propagation, and calculations or estimates based on manufacturers' data can often be accompanied by high uncertainties.

The performance of the filter will be critically dependent on the filter installation method. *Figure 23.9* shows an example of bad installation practice. The performance is severely degraded by cross coupling via the bundled input and output wires and the earth is provided by a wire. The impedance of the wire is in series with the suppression capacitors between the power conductors and the equipment chassis and the high frequency performance of the filter will be degraded. The metal body of the filter should be bonded directly to the chassis via a shake-proof washer and not by a wire and the input and output wires should

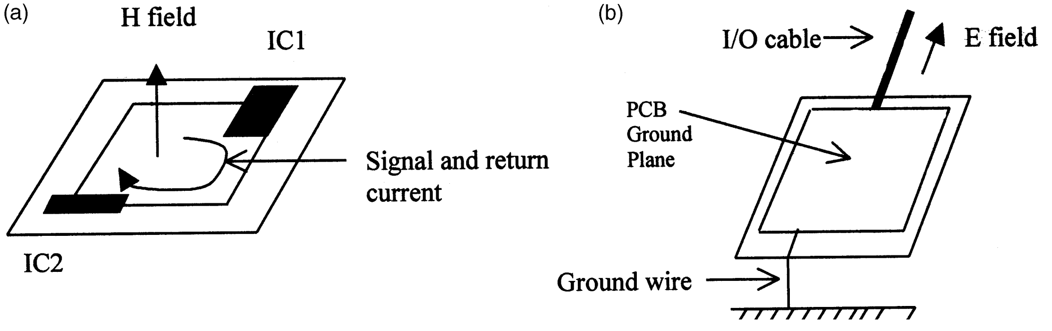


Figure 23.6 Radiation from printed circuit board tracks: (a) differential mode radiation (loop type); (b) common mode radiation (monopole type)

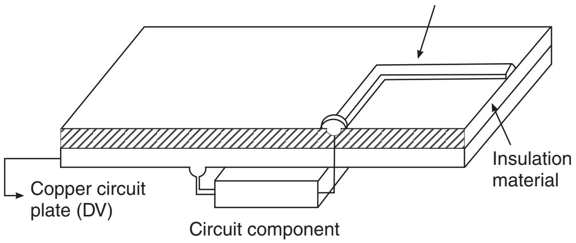


Figure 23.7 Ground plane board

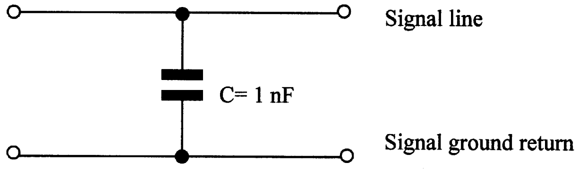
be separated and aligned 180 degrees in opposition to one another.

23.10 Signal line filters

Where there is no shield for the signal cable a signal line filter may need to be employed and should be installed at the point of entry to the equipment enclosure. It is important to ensure that the introduction of the filter does not cause too much attenuation of the wanted signal. Generally low pass filters are employed and for digital data circuits the half power point of the filter response should be no lower than the 9th harmonic of the fundamental data rate in order to preserve the quality of the waveform. Installation is also important for signal line filters. The decoupling capacitors should be connected to a good chassis ground.

The simplest form of signal line filter is the capacitive filter comprising a series of 1 nF capacitors from each line to ground at the point of entry of say an RS232 cable to the equipment enclosure. These may be obtained incorporated within the 'D' connector.

In most cases an 'L' filter can be employed using series inductance (denoted 'L' in the figures) but where extra stages are required e.g. for additional protection, a 'T' filter is usually preferred to a 'Pi' filter because it is less dependent on a good earth.



Capacitive filter

For some high integrity systems, the signal line filters are used in conjunction with screened cables, the filter components often being incorporated within the plug or socket at the end of the cable. However, the small size available will limit the value of series inductance that can be employed and the series elements often comprise a ferrite sleeve or series of beads that give good performance, but only at the higher frequencies.

Additional protection for analogue circuits such as operational amplifiers can be achieved by the use of a balanced (equal value) pair of series resistors at the input pins and by 1 nF decoupling capacitors at the input. It is important to de-couple the supply rails using capacitors having a good rf performance to minimise unwanted potential differences and to reduce the size of loops to avoid coupling with applied electromagnetic fields.

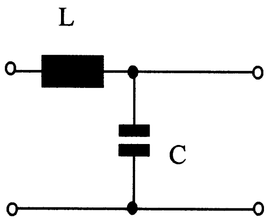
23.11 Enclosure design

Metal equipment enclosures can be configured to provide an effective barrier to electromagnetic waves. Basic shielding theory states that an incident wave is partly reflected at the surface and is then attenuated in its passage though the medium. The total shielding loss is the sum of the reflection and absorption losses. There is another factor known as the secondary reflection loss but this is only taken into account for very thin shields where the absorption is less than 10 dB.

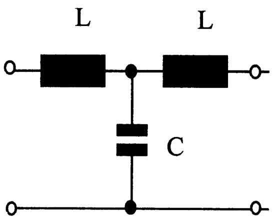
For most practical low frequency applications, a reliable minimum measure of shield performance can be obtained by calculating the absorption loss factor.

$$A = 131 \cdot t\,(f\mu_r\sigma_r)^{1/2}\ \mathrm{dB}$$

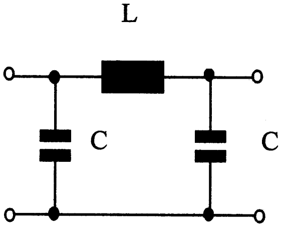
where t is the thickness of the material in millimetres, *f* is the frequency in MHz and μ_r and σ_r are the permeability and conductivity relative to copper respectively. For a 0.5 mm copper sheet at 1 MHz, the absorption loss is 65 dB, more than adequate for most commercial/industrial requirements of 20–40 dB. However the same copper sheet provides only 0.5 dB at 50 Hz. If effective protection is required against power frequency magnetic fields, higher permeability materials such as steel ($\mu_r = 300$–1000) will be required. For a 20 dB attenuation at 50 Hz the thickness of the steel will need to be about 5 mm. The high frequency performance of practical enclosures is not so much dependent on the inherent properties of the material but far more on the apertures and discontinuities in the surface.



'L' Filter



'T' Filter



'Pi' Filter

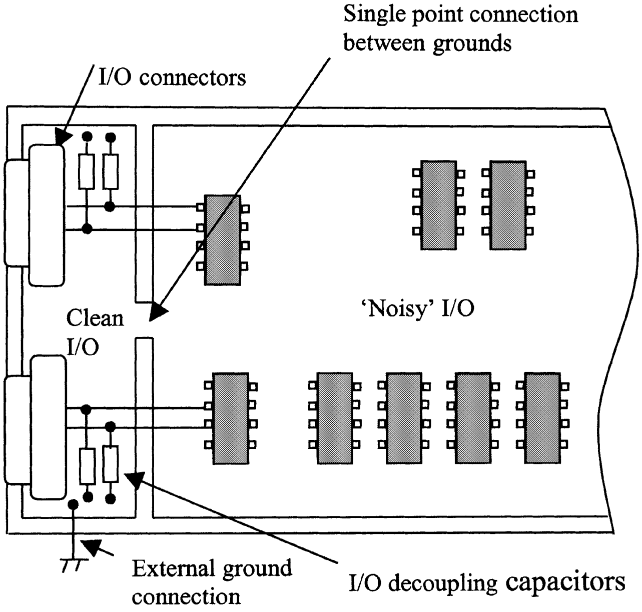

Single point connection between grounds
I/O connectors
Clean I/O
'Noisy' I/O
External ground connection
I/O decoupling capacitors

Figure 23.8 PCB interface design

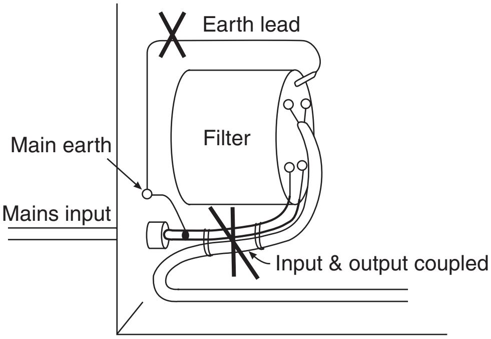


Figure 23.9 Example of very poor filter installation

Electromagnetic energy will leak through apertures to a degree dependent on the ratio of the aperture diameter to wavelength. Where the aperture diameter is one half wavelength it is virtually transparent. The degree of attenuation at lower frequencies can be easily derived using the expression $SE = 20\log_{10}(2D/\lambda)$, where λ is the wavelength at the frequency of interest and D is the largest dimension of the aperture. For high attenuation D must be as small as possible. For large apertures for ventilation etc., a mesh may be used to cover the area thus providing an effective shield. In general terms, a large number of small apertures is preferred to a small number of large apertures.

Discontinuities in the surface can also degrade shielding performance, especially where the length of the slot is one half wavelength. The length of the slot should be reduced by providing more points of contact between the sections. Ideally there should be a good surface to surface bond at the connections between modules in a rack enclosure, and the additional expense of modules with rf spring fingers along the edge will pay dividends both for reducing the transfer of electromagnetic wave energy but also in protecting against ESD. Gaskets can also be used to achieve good high frequency electrical bonding between metal panels and enclosure sections. However, the enclosure must be designed from the outset to accept the gasket.

Many enclosures are not metal, but plastic or other non-conducting material. These would provide no shielding to electromagnetic waves but it is possible to introduce a shielding effect by various means. These include painted, sprayed, sputtered or plated metallic coatings to the inside surface of the plastic enclosure or by using metallic loaded plastic material. Generally these measures produce only low levels of shielding, e.g. less than 30 dB, but they can be extremely effective in improving the situation for marginally non-compliant equipment.

23.12 Interface cable connections

Cable characteristics can have an impact on the EMC performance of a system. Unshielded twisted pair cable is often used for low and medium speed data links. In the presence of an incident electromagnetic wave, voltages will be induced in the cable both in common mode and differential mode. Although the differential voltage is minimised at low frequencies by the twists in the cable, the common mode voltage remains and will be applied to the circuit to which the cable is connected. The balance of the circuit is therefore crucial and if the impedances to ground at the circuit are equal, the conversion to unwanted differential mode noise is minimised. The difficulty is that parasitic components such as stray capacitance tend to unbalance the circuit at high frequencies and the ability to reject common mode disturbance reduces rapidly with increasing frequency, thus causing interference.

Additional protection can be provided by the addition of an outer screen usually of copper braid or aluminium foil. At high frequencies, the electromagnetic energy is confined to the inner and outer surface layers of the screen by

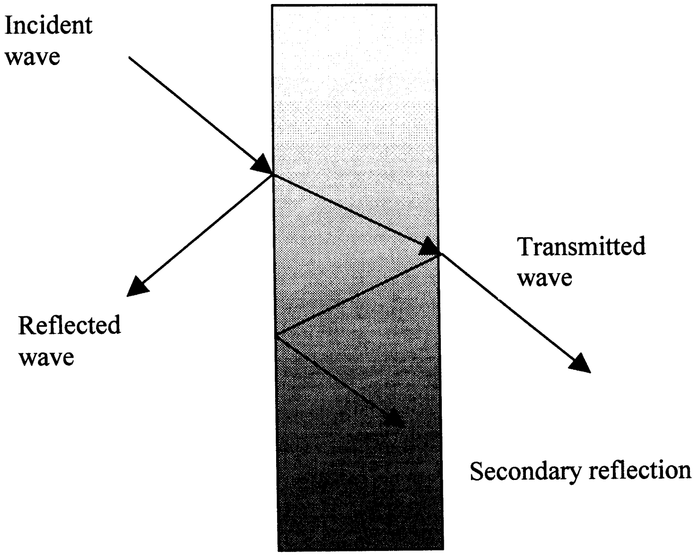


Figure 23.10 Shielding theory

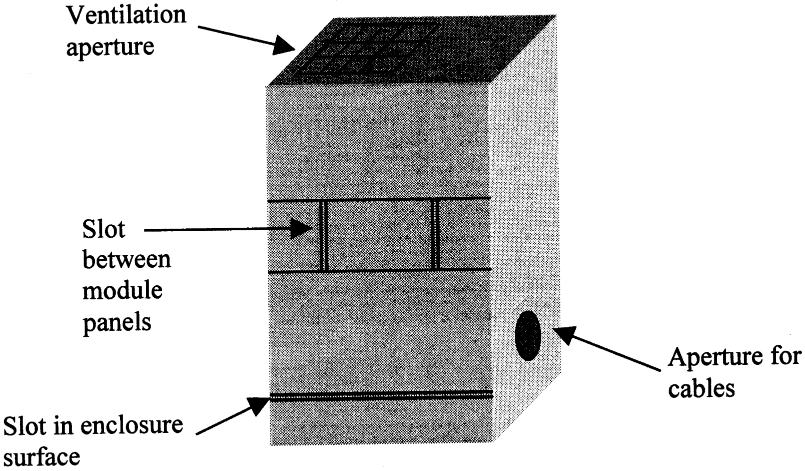


Figure 23.11 Apertures and discontinuities in Enclosures

skin effect and isolation between the inner conductors carrying the wanted signal and the external environment can be readily achieved through good cable screen design. Screened cable performance can be characterised by its surface transfer impedance (Z_t) which is the ratio of the voltage induced on the inner conductor to a current in the outer surface of the outer conductor. The shielding effectiveness of coaxial cables is given approximately by $SE = 20 \cdot \log_{10}(50/Z_t)$. The transfer impedances of various types of screened cable are shown in *Figure 23.12*.

At low frequencies performance is determined by the self impedance of the screen conductor. For the single copper braid for example, the weave allows energy to pass through the shield and high frequency performance is reduced, although the level of performance is usually adequate for most purposes. For special applications such as for cables carrying very low level signals in an intense electromagnetic environment such as a nuclear power station the outer screen is composed of many layers including inductive materials as in super screened cables, and excellent performance is achieved.

Having the correct termination for the cable shield is an important consideration in achieving the maximum performance of the cable. Ideally the cable screen should be terminated in a 360° peripheral glanded connection as shown in *Figure 23.13*.

If the cable screen is made up into a ‘pigtail’ connection, performance at high frequencies will be severely degraded and pigtail screen connections should be avoided. At high frequencies the impedance of the pigtail connection becomes significant and is a point of common coupling between the inner conductors and the external environment, increasing emissions and reducing immunity to external disturbances.

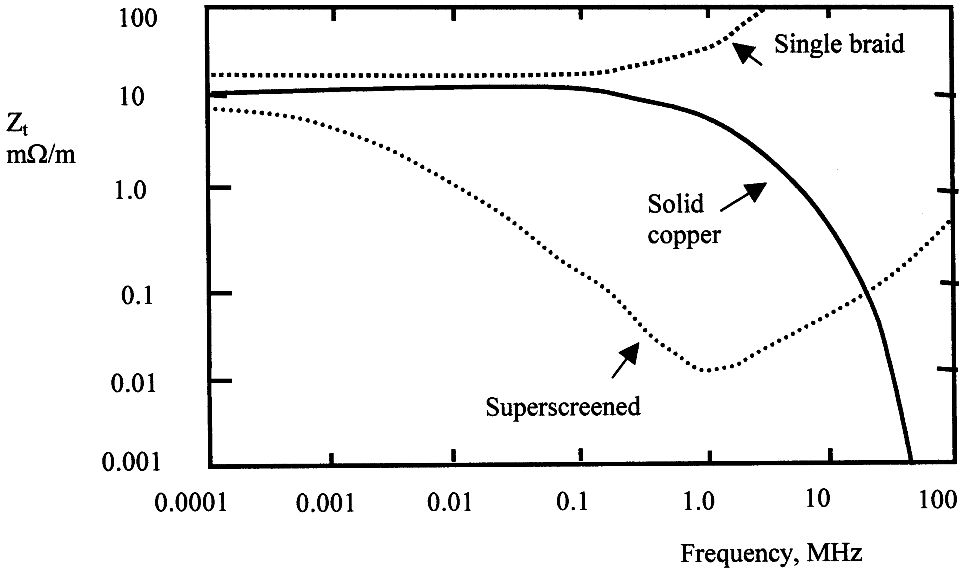


Figure 23.12 Transfer impedance of various types of screened cable

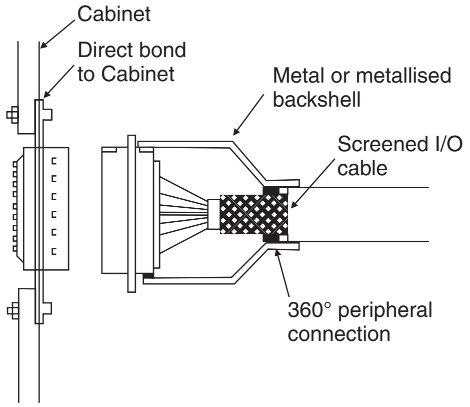


Figure 23.13 Correct 360°peripheral termination of screened cable

23.13 Golden rules for effective design for EMC

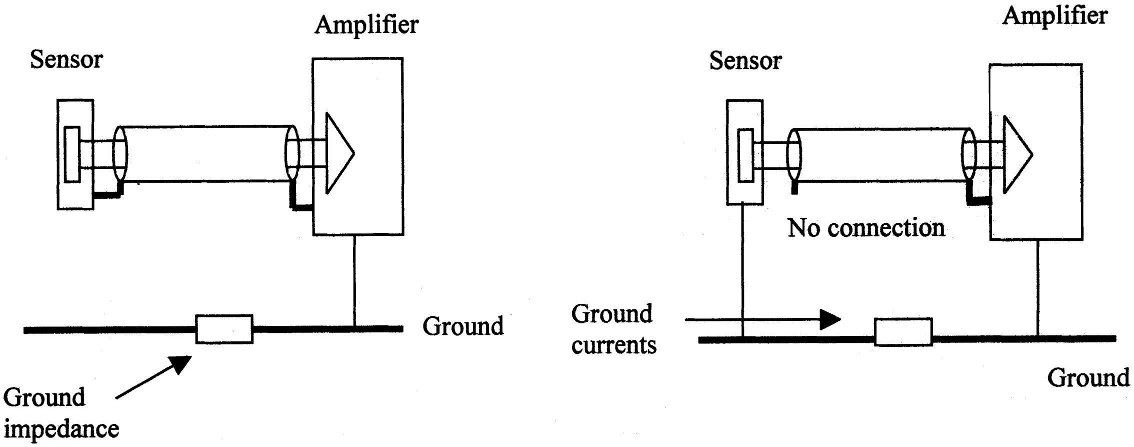
(1) Consider the location and design of EMC 'barriers' at the concept stage.
(2) Consider the use of multi-layer printed circuit boards with good layout and interface design.
(3) Select filters for unscreened power and/or signal cables entering the enclosure.
(4) Ensure correct filter installation at the point of entry to the enclosure.
(5) Connect the screens of interface cables at the point of entry to the enclosure by a glanded 360°peripheral connection.
(6) Bond the enclosure sections and minimise the size of apertures for best enclosure shielding.

23.14 System design

For distributed systems comprising several or many electronic equipment interconnected by cables, good cable installation practice combined with other earthing and grounding measures can be applied to good effect to improve system reliability in the presence of electromagnetic disturbances. These measures for improving immunity are equally effective in reducing emission levels from an installation and will ease the task of operators and system installers in ensuring compliance with the essential requirements of the EMC Directive. The main methods of control are: cable segregation, where cables of various types are routed in groups so that cross talk or disturbance pick up is minimised, cable separation where cables with high levels of disturbance are spatially separated from the cables carrying sensitive signals and isolation by the use of earthed conductors which reduce coupling between cable types.

In a first step towards good EMC engineering practice, the following general guidelines should be observed:

- *Cable segregation*: Sensitive cables such as signal cables may be grouped together; mains cables, including power feeds and lighting circuits, carrying up to 250 V may also be grouped together but the cable types should not be mixed.
- *Cable separation*: In addition to the groupings, data, telecoms or sensitive cabling should be separated from three-phase cables used for heavy electrical inductive load switching e.g. air conditioning, welding equipment and motors by the largest practicable distance.
- *Isolation*: Metal cable trays, if not already in use, should be implemented. Having an adequate low impedance for the frequencies in use, and with good earthing, the tray will effectively become a partial screen for the cables.
- *Shield termination*: Cable shield termination is also a key factor in controlling EMC but the best practice is often dependent on the particular circumstances. For low frequency applications the shield may be terminated at only one end in order to mitigate against ground



a) Sensor not grounded

b) Sensor grounded

Figure 23.14 Screened cable termination methods

Table 23.6 Cable type classification

Cable type	*Characteristics*
Very sensitive	Low-voltage or low-currents as from sensors
Sensitive	Signalling cables. 24 V flat cables for parallel data transfer
Indifferent	AC power between 100 V and 250 V, depending on the EMC properties of the apparatus connected
Noisy	AC and DC relay feed line without protection measures such as filters or diodes
Very noisy	Leads to DC motors with brushes, switched power lines, cables and earth wires in high voltage switchyards, etc.

noise currents but this will reduce its effectiveness, particularly against magnetic fields.

For cables carrying low frequency signals, cable terminations have to be designed carefully to avoid coupling with noise currents flowing in the ground. If the sensor is not directly connected to ground as in *Figure 23.14*(*a*) above it may be possible to terminate the screen at both ends, thus providing maximum protection against inductively coupled disturbances. If the sensor is grounded as in *Figure 23.14*(*b*), the voltage drop across the ground impedance to noise currents in the ground will give rise to high currents in the shield and noise voltages may be present at the input to the amplifier. This can be overcome as shown by terminating the cable at one end only, thus avoiding the ground loop, but the performance of the shield may be reduced. If high performance is required under all conditions e.g. with the sensor grounded it may be necessary to introduce transformer coupling or opt-isolation in order to minimise unwanted coupling.

Cables transporting similar signals can often be bundled together. With cables transporting different signals it is possible to differentiate between cables as shown in *Table 23.6*.

The five types of cables listed above should be separated in sequence from each other by at least 150 mm on cable trays or racks. That means that a very sensitive cable should he separated from a sensitive cable by at least 150 mm, and from a very noisy cable by at least 600 mm. In this latter case, a minimum separation distance of 1 m is recommended if cable racks are used.

The use of a parallel earthing conductors (PEC) can reduce common mode currents in signal leads by reducing the system common mode impedance and the loop area. Routing cables through trays, conduit or ducting has the effect of introducing a parallel path for disturbance current, which not only is capable of diverting and withstanding high currents but also of providing the necessary low impedance thus protecting the signal cable. Additional connections to earth should be made at regular intervals for very long cable runs.

Using trays or racks of sufficient wall thickness when used to separate cables, can provide both PEC and differential mode reduction in cross-talk. They can often be laid next to each other. Another solution is to keep some

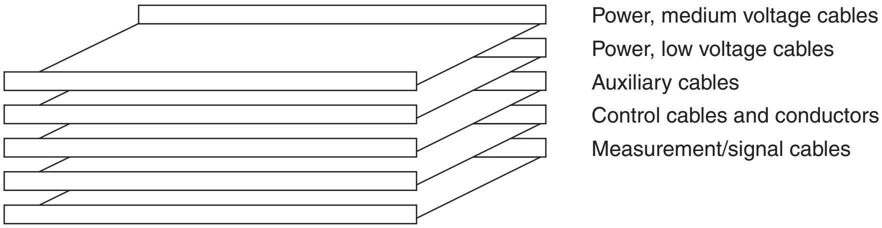

Power, medium voltage cables
Power, low voltage cables
Auxiliary cables
Control cables and conductors
Measurement/signal cables

Figure 23.15 Stacking conduits to avoid cross-talk

distance between shallow conduits for the different types of cables, for example by stacking them (*Figure 23.15*).

Earthing and bonding are key features in the EMC design of a large system or installation. Earthing is the connection of the exposed conductive parts of an installation to the main earthing terminal of that installation; bonding is an electrical connection maintaining various exposed conductive parts and extraneous conductive parts at substantially the same potential. The general requirements of an earthing network are to provide safety and to provide a low impedance path for currents to return to source.

The current preferences are for a multiple bonded earthing network which is compatible with the measures for lightning protection (*Figure 23.16*). The frame of the building is bonded to the lightning conductors to avoid flash-over with multiple connections to earth electrodes.

Internal to a building, the recommended approach is a three-dimensional network, one earthing network per floor, each network connected to one another and to the earth electrode. The mass of inter-connections is called a common bonding network (CBN). In some cases e.g. telephone exchanges an earthing mat may be installed but only connected to the building frame at one point to provide a mesh isolated bonding network (IBN). The main feature of the CBN is that each floor of a building should have a meshed ground network beneath the floor with multiple connections between the floors. The mesh network can take the form of not just the dedicated ground structure, but installed cable trays (metallic), water pipes and ducts. The meshed earthing scheme will produce an earthing system which should have low impedance, and a high probability of remaining so, even if a number of earth connections become corroded and fail.

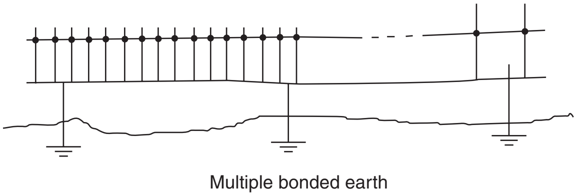


Figure 23.16 Preferred method of earthing network

23.15 Buildings

In most practical applications, electrical and electronic equipment is installed and used inside a building. Where the building is exposed to intense sources of electromagnetic disturbance, some additional protection may be required. For example, buildings on airfields are often illuminated by the incident radiation from high power radio and radar transmitters in the near vicinity, where the fields are in excess of the equipment immunity test levels of 3 and 10 Volts/metre. Measures such as shielding on the windows may be a minimum requirement but in some cases shielding of the room containing the electronic equipment may be required. This introduces the concept of protected zones which can also be developed for protection against other threats such as lightning strikes (*Figure 23.17*).

Such protection is usually considered at the design stage but can be introduced retrospectively in some cases on a small scale. The need for additional measures is usually determined by a site survey of the electromagnetic environment at the building location. Other sources that often need to be considered are overhead power lines and overhead railway a.c. electrified lines. Here the magnetic field at the power frequency can cause distortion of the display on computer monitors. The threshold for distortion is about 1 Amp/metre and provision for some measure of control, usually by re-location or local shielding for the monitors, should be made if the field is likely to exceed this level.

23.16 Conformity assessment

In Europe, products placed on the market must comply with the essential requirements of the EMC Directive. (The exceptions are military equipment, which must meet defence standards, and equipment covered by other directives which address EMC e.g. medical devices.) The manufacturer is required by law to make a declaration of conformity stating that the product complies with relevant harmonised standards referenced for application under the EMC Directive. He may then apply the CE mark to his product and will obtain the benefit of free circulation of the product within the Single European Market without experiencing any barriers to trade. It should be noted that the CE Mark denotes compliance with *all* applicable directives which for

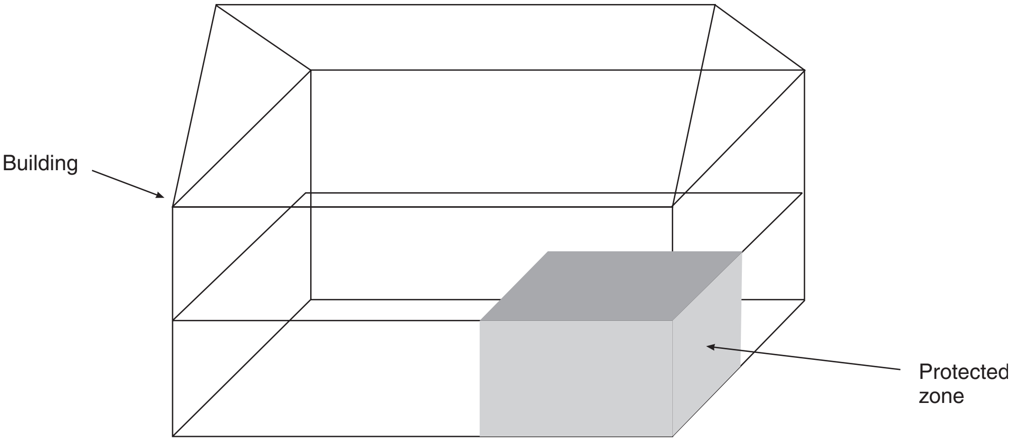


Figure 23.17 Protected zone in a building

Table 23.7 European directives and EMC standards

Product	*Directive*	*EMC Standards*
Household equipment	EMC Directive 89/336/EEC	EN 55014-1 and –2
Maritime navigation equipment	EMC Directive 89/336/EEC	EN 60945
Lighting equipment	EMC Directive 89/336/EEC	EN 55015 and EN 61547
Information Technology Equipment	EMC Directive 89/336/EEC	EN 55022 and EN 55024
Railway Equipment	EMC Directive 89/336/EEC	prEN 50121 parts 2–5
Low Voltage Switchgear	EMC Directive 89/336/EEC	EN 60947
Radio and TV receivers	EMC Directive 89/336/EEC	EN 55013 and EN 55020
Radio and radio-telecommunications equipment	RTTE Directive 99/5/EC	Applicable TBR, EN or ETS standards
Medical equipment	Medical Devices Directive 93/42 EEC	EN 60601-1-2
Motor vehicles	Automotive EMC Directive 95/54/EEC	The directive contains the technical requirements

most products, will also include the electrical safety requirements of the Low Voltage Directive 73/23/EEC.

For finished products, complex components and systems, the manufacturer has the option of either applying appropriate harmonised EMC standards i.e. those approved by the European Commission or can prepare a Technical Construction File where the harmonised standards are not available or are not applied in full. Generally most manufacturers apply the harmonised standards as the most cost-effective route and this usually means that the product must be tested to EMC standards to determine compliance.

Table 23.7 presents examples of the requirements for common types of electrical and electronic equipment.

It is the responsibility of the manufacturer to determine which directives apply to his product but information is usually generally available from standards bodies such as the BSI, from EMC test laboratories and Competent Authorities such as the UK's Department of Trade and Industry.

The European assessment requirements for fixed installations are somewhat different. The installation may comprise compliant (CE marked) equipment and/or non-compliant equipment. The installation must meet the 'essential requirement' of the EMC Directive, i.e. not to cause interference and not to be affected by electromagnetic disturbance. There is no necessity for declaring compliance nor for CE marking. However the installation designer should ensure that the equipment is installed according to good engineering practice and is advised to maintain a file containing a description of the EMC measures taken, together with any test data etc. This file should be available for inspection by the authorities if challenged at some later date. The key issue is the acceptance by the designer or operator of non-compliant equipment. This is permissible but becomes the responsibility of the operator if interference results and therefore the operator should ensure that this equipment is compatible with its environment either by imposing some contractual requirements for EMC characterisation on the supplier or by in-situ confidence testing.

23.17 EMC testing and measurements

For the vast majority of electrical and electronic equipment, compliance with the relevant technical specification is achieved by testing to EMC standards. Emissions measurements can comprise two types, a) tests for low frequency phenomena such as power frequency harmonics and voltage fluctuations to protect the power distribution network, and b) high frequency emissions to protect the radio spectrum. For commercial and industrial products, high frequency emission measurements are normally made over the frequency range 150 kHz to 1 GHz but some standards call for measurements down to 9 kHz. It should be noted that it is becoming increasingly necessary to measure emissions above 1 GHz to protect cellular and other radio systems.

The low frequency phenomena are usually measured with proprietary test instrumentation dedicated for that purpose comprising a harmonics analyser and flicker meter. Most high frequency tests comprise measurements of conducted disturbance over the frequency range 150 kHz to 30 MHz and radiated disturbance at higher frequencies, 30 MHz to 1000 MHz. The different test methods effectively reflect the propagation mechanisms that dominate in practice.

Conducted disturbance measurements on the mains power input comprise a measurement of rf voltage across a passive network (50 uH/50 ohms) having an input impedance representative of the rf impedance of the mains at high frequencies. The voltage is measured using a calibrated EMC measuring receiver or spectrum analyser. Most modern instrumentation systems are computer controlled for maximum efficiency.

Radiated measurements are performed on a test range meeting particular requirements for path loss, or in a facility giving results which can be correlated to those on the test range. The standard test range comprises on open area site which is flat and free from reflecting objects and typically having the dimensions shown in *Figure 23.18*. A significant proportion of the area should be covered by a conducting ground plane in order to achieve the site attenuation calibration requirements.

Shielded enclosures lined with radio-wave absorbing material or ferrite tiles on all internal sides except the floor can be constructed to achieve site attenuation performance characteristics comparable with an open area test site. Due to cost limitations, most chambers are built to accommodate a 3 metre range and some additional calibration work may be required to relate the results to the 10 metre range, particularly for physically large items of equipment.

Measurements of field strength are made using a receiver or spectrum analyser and a calibrated antenna situated at a fixed distance from the equipment under test. A key factor

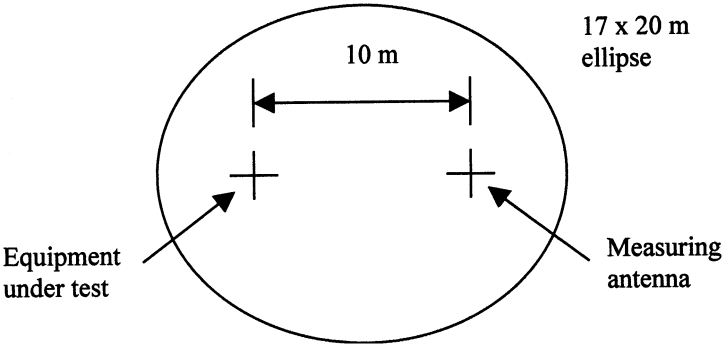


Figure 23.18 Dimensions of the CISPR open area test for radiated emissions

Table 23.8 Test equipment

Test	*Equipment*
ESD	ESD Gun
Fast transients	Transient generator and coupling clamp
Surges	Surge generator
Dips and interruptions	Dips generator
Rf fields	Signal generator, power amplifiers, antennas
Rf voltages	Signal generator, power amplifiers, coupling networks

in emissions measurements is to configure and operate the product in a manner which is both representative of practical use, yet maximises the emissions from the product. For products and small systems which contain cables some pre-testing is often required to determine the optimum configuration for test, particularly for radiated emissions tests.

Immunity tests are performed by subjecting the equipment to electromagnetic disturbances of the type and maximum level that the product is likely to encounter in its intended operating environment. For many of the tests, specialised test instrumentation is required, providing the well-defined disturbance characteristics described in the standards. The instrumentation is commonly available from a number of specialist suppliers and the tests which involve the direct application of disturbance voltages are relatively straightforward.

The facilities for subjecting the product to disturbance fields are more complex and costly. The rf field immunity test is performed inside a shielded room lined with radiowave absorbing material or ferrite tiles and the costs of purchasing and installing the facility are often quite high. The test method requires that the test volume is pre-calibrated to achieve the desired test level. There is an additional field uniformity requirement for the maximum level of field over 75% of a 1.5 m square area to be no greater than twice the required test level. Rf power over the frequency range 80–1000 MHz is fed to a dedicated radiating antenna in the room from a signal generator and high power amplifier. The system is usually computer controlled to achieve the calibrated level in an efficient and reproducible manner.

To simulate coupling with rf fields at frequencies below 80 MHz, a common mode voltage test is applied to the cables attached to the product, i.e. the mains leads and signal and interface cables using coupling/decoupling networks (CDNs). The measurements are usually made inside the shielded enclosure for convenience and to avoid causing interference to radio services.

Immunity to power frequency magnetic fields is performed by setting up current flow in a loop around the product or preferably by placing the product within the confines of a Helmholtz coil.

In immunity testing, the configuration and operating conditions of the equipment under test can be very important. These conditions are usually specified in the EMC standard applicable to the product but in the absence of specific information the test engineer must attempt to adjust the conditions for maximum sensitivity to the applied disturbance. This can be difficult because immunity testing does not give a resultant variable, only an attribute of compliance or non-compliance, and care is required to follow the specific instructions of the standard and the advice and guidance of the manufacturer. Monitoring of the product's performance should be performed in a non-invasive method that does not influence the coupling of the applied disturbance. In addition the criteria for assessing the product's performance need to be clearly established prior to testing. For example for an analogue instrument such as a temperature sensor, the margin of acceptable error, say 1%, needs to be stated in the User Manual.

23.18 Management plans

For all product, system or installation developments, compliance with EMC requirements can be made much easier by the adoption and adherence to an EMC management plan. This would probably be integrated within a broader conformity assessment programme which would include

Table 23.9. Product development process

Product development stage	*EMC input*
Design Concept	Apply EMC control principles
Design Process	EMC design measures
Prototype	Pre-compliance EMC tests
Production Prototype	EMC certification
Production	Conformity in production tests
Upgrades/modifications	EMC re-assessment

safety and other approvals. The plan would be prepared at the inception of the product development with an intention of achieving complete compliance at the lowest cost and in the shortest timeframe both at the initial certification stage and also through the lifetime of the product. The plan would include Control Plans which would specify the intended operating electromagnetic environment, any particular EMC elements in the Technical Specification, a Compatibility Matrix to identify couplings with other systems, EMC Test Plans, and applicable codes of practice to aid the design team. *Table 23.9* lists the EMC inputs required at the various stages of a product development.

Procedures can be set up for regular EMC design reviews and formal acceptance of supplied sub-assemblies or equipment from sub-contractors. In some cases, particularly for large systems and installations, it may be necessary to establish procedures for dealing with concession requests and dispensations where supplied items are non-compliant.

References

1 PAUL, C. R., *Introduction to Electromagnetic Compatibility*, John Wiley and Sons Inc (1992)

2 GREEN, M., *Optimised and Superscreened cables*, Raychem Ltd

3 *Design for EMC. ERA Report 90-006,* ERA Technology Ltd (1990)

4 DEGAUQUE, P. and HAMELIN, J., *Electromagnetic Compatibility*, Oxford University Press (1993)

5 OTT, H. W., *Noise Reduction Techniques in Electronic Systems*, John Wiley Interscience, NY (1988)

6 GOEDBLOED, J. J., *Electromagnetic Compatibility*, Prentice Hall (1992)

7 *Cabling Installations: User Friendly Guide*, ERA Report No 98-0668

8 *Information Technology—Cabling Installation,* Part 2 Installation Planning and Practices inside Buildings, prEN 50174-2 June 1998

24

Health and Safety

D A Dolbey Jones BEng, FIEE, MIOSH
Former HM Senior Electrical Inspector of Factories

Contents

24.1 The scope of electrical safety considerations 24/3
  24.1.1 General overview 24/3
  24.1.2 Recent developments 24/5
  24.1.3 Legal and administrative 24/5

24.2 The nature of electrical injuries 24/6
  24.2.1 Types of injury 24/6
  24.2.2 Electric shock 24/6
  24.2.3 Other injuries 24/7
  24.2.4 Protection against electrical injuries 24/8
  24.2.5 Toxic hazards 24/9
  24.2.6 Conclusions 24/9

24.3 Failure of electrical equipment 24/9
  24.3.1 Causes 24/9
  24.3.2 Particular equipment 24/9
  24.3.3 Protection against failure 24/11
  24.3.4 Fires and explosions 24/11

24.1 The scope of electrical safety considerations

24.1.1 General overview

It has often been said that electricity poses hazards quite unlike any other hazard which commonly presents itself to persons in the workplace and elsewhere, including in the home. Electricity has the capacity to kill and injure through electric shock, through burns and through fires started electrically. Other dangers involving the use of electricity may be found in the aberrant and dangerous actions of electrically controlled machinery and the maloperation of plant control systems based on relay and/or computer logic circuitry.

Electrical accidents, unlike most other industrial accidents, are more likely to involve professional and supervisory staff. In some situations they may be at greater risk than most other staff particularly where electrical power distribution circuits are being switched or maintained. In a typical year[1] 47% of electrical accidents involved electrically skilled persons and out of a total of 805 electrical accidents 57 were to supervisory and testing staff. Electricity has a stealth and power which are often disregarded even by electrical engineers and other electrically trained persons.

All those who are involved in potentially hazardous work with electrical equipment, such as the testing of high-voltage apparatus or supervising the alterations to circuits in an important substation, have statutory legal duties of which they must be aware. They must conduct their work in accordance with the statutory requirements and ensure that those under their control also conduct their work properly. In the UK these are criminal law obligations, quite distinct from liability for negligence in civil law which may also arise if matters go wrong. All engineers should make themselves aware of the statutory requirements and of the potential liabilities which they run as persons with responsibilities for the health and safety of themselves and others.

Nearly all electrical accidents, even those at low voltages (e.g. 230 V a.c.) are potentially fatal and the causes and remedies are now almost completely understood. There is no shortage of guidance on the various hazards presented by electricity and on the statutory requirements relating to these. In Great Britain much of this guidance is published by the Health and Safety Executive (HSE) which has the main responsibilities for safety and health legislation arising from 'work' activities and the enforcement of this law.

Other considerations in the context of health and safety of which electrical professionals should be aware are the various directives of the European Union (EU), such as the 'Low Voltage Directive' 73/23EEC, the 'Machinery Directive' 98/37/EC and the 'ATEX Directive' 94/9/EC which have a safety content. A raft of EU directives and domestic implementing legislation covers the generality of health and safety on hazards other than from electricity but these are outside the scope of this discussion. For further details on these refer to the HSE, DTI and other sources listed below, some of which are available on-line through their Internet websites.

24.1.1.1 Control of staff and permits to work

In the context of persons 'at work', the emphasis in all work on electrical equipment must be that this normally be conducted with the equipment de-energised and made safe, i.e. 'dead' working. 'Live' working is a specialised activity and is permitted only in exceptional circumstances. No one should be allowed to do anything which carries electrical risks without the provision of adequate precautions to reduce the risk of injury to a minimum. The persons must also have the necessary skill, experience, technical knowledge and other resources to do the work safely. Managers and supervisors must therefore satisfy themselves that no one is asked to do any work for which they are not competent and specifically authorised by the management. Such authorisation is usually through a formal process of instruction and examination in company procedures.

Particular care must be taken with students, trainees and apprentices. Apprentices are often expected to learn to carry out potentially dangerous jobs during their apprenticeship but they must be carefully supervised.

When work must be done where electrical danger may be present, and some electrical work (particularly in testing and maintenance) cannot be made absolutely safe, instructions must be precise and unambiguous. Procedures must be observed correctly. All necessary and relevant technical manuals and drawings etc. should be available.

Much 'dead' work is carried out under 'permit to work', particularly in the electricity generation, transmission and distribution industries which have an established and universal system of safety rules (with local variations for each electricity company). Permits usually relate to equipment which has been made dead. Permits must be issued and cancelled in an orderly and clearly defined manner. A full record must be kept so that it is possible, at any time, to find out what is going on, who is involved or at risk, and what precautions have been taken.

The permit must state clearly and fully to whom it is issued (this person should be present at all times and is responsible for what happens), name those persons who may be present in the working area, and state what special precautions have been taken to prevent danger. The safe and unsafe areas must be stated on the permit, preferably supported by means of a diagram, and clearly indicated on the actual work site by means of flags and barriers etc. The work to be done must be clearly described and no other work may be carried out because this could entail risks not contemplated by the person issuing the permit, who may not have taken the extra precautions necessary.

At the end of the work there must be a clearly defined procedure for handing over. A check must be made that all persons have been withdrawn from the work site and the result recorded. Before the permit is cancelled, a statement must be recorded (preferably on the back of the cancelled permit and also in the log book, where one has been kept) of what work has been done, what has not been done, and what steps have been taken to make the site ready for resumption of normal operations. Until a permit has been cancelled the person to whom it was issued remains responsible for the conduct of any work on that equipment.

If the work lasts for more than one shift there must be an appropriate method of hand over to ensure that the new shift supervisor is familiar with the state of the work and the terms of the permit. It is often preferable to cancel the first permit and to issue a new one. Sometimes the person with the authority to issue permits takes charge of the work, in which case they should issue a permit to themselves.

Experience has shown that all this detailed procedure is essential. The routine not only ensures that there is a record which will help to identify the cause of any mistake, but the action of writing down all the details helps to prevent anything being overlooked. Since all key persons involved must sign all of the records and statements the routine reinforces the understanding of the various instructions and helps to ensure that these instructions have been read and understood.

Some testing and research work presents its own hazards. As the conditions may vary greatly it is difficult to lay down general rules and safety depends largely on the skill of the staff. For certain work earth-free areas with power supplies isolated from earth are provided and used in conjunction with other precautions such as the use of unearthed tools (such as very low voltage soldering irons).

Routine high-voltage testing is normally carried out in enclosures with interlocked doors and provision for supervision from the outside. Where unskilled or semiskilled persons do routine testing on a production line arrangements must be made, by guards and interlocks for example, to ensure that they are not exposed to live conductors operating above 25 volts a.c. or 60 volts (smooth) d.c. thus eliminating likely shock risks.

24.1.1.2 Non-electrical causes

Many electrical accidents are the result of mechanical and other causes. Examples of these causes are mechanical stresses, thermal shock on insulators, resonant vibrations of conductors leading to fractures, low-temperature brittleness or corrosion fatigue. Too often however the causes are banal, such as a badly constructed joint or a loose connection. To deal with such troubles it is necessary to have more than a narrow interest in electrical matters. Much of electrical safety engineering requires a good understanding of mechanical engineering and the strength of materials because many accidents involving electricity are due to a poor understanding of the properties of materials. However, electrical accidents are also largely due to the uses and abuses to which otherwise sound equipment is subjected.

A further dimension in electrical accidents is that the stealthy power of electricity is inadequately understood. Even many electrical engineers, perhaps because their training and experience may not have exposed them to the destructive capabilities of electricity, can be disrespectful of this power. For example, very few engineers actually get to witness a substantial power arcing fault and few get to examine the after effects of a serious electrical fire, explosion, arcing fault or other dangerous occurrence. However, those that do usually find themselves surprised and acquire a deeper respect for electricity.

The official report on the enquiry into the disastrous explosion and fire at Flixborough in 1974 (which featured the mechanical failure of a temporary modification to the naphtha cracker) stated that engineers should have academic and practical training in all branches of engineering outside their speciality which may affect their work. This might seem to imply the need for training in the many fields of expertise of only tangential interest to one's own discipline but in reality one person can seldom expect to understand every aspect of a particular plant, apparatus or process on which they have to work. There is often a need to draw upon shared expertise with colleagues or even to seek outside help in order to deal properly with certain problems. The key point is that the engineer should recognise his or her limitations. If the competence, experience or expertise are inadequate to the circumstances then help should be sought.

24.1.1.3 Design standards

The integrity of much electrical equipment is now greatly assured through a comprehensive range of published Standards which have been established nationally and internationally through the various consensus seeking Standards organisations such as the British Standards Institution (BSI), the European Electrical Standards body, CENELEC, and the International Electrical Standards body, IEC. The work of these bodies has contributed hugely to the safety and integrity of much electrical equipment over many decades. However, safe use of equipment also relies on the correct interpretation of the scope of the Standards applying to that equipment. A Standard is only one link in a chain.

It is worth remembering too that with few exceptions Standards are written in a consensus seeking forum comprising representatives from the various interested groups. These groups usually comprise manufacturers, or their trade associations, government bodies, test houses and inspection agencies and, to a lesser extent, representatives of the actual users of the equipment.

British Standards have fulfilled a useful role over decades but it must now be recognised that almost all Standards work is conducted in an international forum with the needs, hopes, customs and expectations of domestic users being subject to a large measure of international influence. This internationalisation is not always welcomed by some but it is inevitable.

Traditionally British Standards have been drawn up in a manner which recognise some basic principles of electrical engineering safety. Some of these are:

- The insulation of conductors should be unable to come into contact with moving parts.
- Earthing terminals should be adequately locked against loosening. These terminals shall not serve for any other purpose, e.g. for securing parts of the case.
- Electrical connections should be so designed that the contact pressure is not transmitted through insulating material other than ceramic or other materials not subject to shrinkage or deterioration.
- Soldered connections should be so designed that they keep the conductors in position if the conductor breaks at the point of connection

In many situations it is important that fingers, steel rules etc. cannot touch live or moving parts of electrical equipment and a number of probes have been devised to test and prevent this, including a standard test finger which is hinged to mimic the reach of the fingers on a human hand.

24.1.1.4 Investigations

Most engineers will, at some time, have to investigate an accident or plant failure. The purpose of an investigation is to ascertain the facts and initiate any necessary action. The first requirement is to make sure that all the relevant information has been obtained and that it is correct. Persons who have witnessed a severe accident are often shocked and emotionally distressed. They may be quite unable to distinguish between what they have seen and what they think they ought to have seen, or perhaps what they have imagined when trying to rationalise their confused memories. Some people may also have good reason for wanting to mislead while others quite genuinely lose all memory of events immediately leading up to the accident. Sometimes however, the person injured is less upset and a better witness than the on-lookers.

It is important to remember that the impossible does not happen, and the improbable only happens occasionally. Accidents are usually the result of several contributory factors coinciding whereas a machinery design or a procedure is likely to have been based on only one factor occurring at a time.

It is important to examine the debris and all equipment very carefully after a failure and be critical of stock wiring diagrams; they frequently contain mistakes or refer to the wrong apparatus. It is common for modifications not to have been recorded properly.

Having determined how an accident has happened, it is important to find out why. This usually involves the highly complex issues of interfaces between humans and machines and the management of staff. Temperament is important in some jobs and boredom, stress, overwork, rivalries, physical and mental health and all human emotions play a role in the workplace. An example of the importance of temperament matching the task is that of a plant control engineer who may have long periods of dull routine punctuated by occasional emergencies when quick and correct decisions are necessary. It is a job which suits only certain types and where the consequences of error can be extremely serious.

24.1.1.5 Written reports of accidents

If an accident report is muddled or unconvincing, the time spent on the investigation will have been wasted. A report should err slightly on the long side rather than be obscure but it is preferable that it conveys the main points lucidly and concisely.

Reports that recommend action by managers or directors will fail if the reports are incomprehensible to these people. Good report writing is an art which usually requires much concentration and practice. Some pointers are:

- Careful preparation pays dividends. Well chosen headings and a planned layout are essential. The argument needs to be thought out clearly and arranged in a logical sequence so that, as far as possible, each paragraph follows naturally from the one before. If this thinking is done thoroughly the actual writing of the report will take on a natural flow which will be limited only by the speed at which one can write or type.
- Each sentence should carry the argument forward starting in such a way that the reader half expects what is to follow.
- Care is needed with punctuation. For example, commas should not be used excessively or in order to divide up a sentence like the brackets in an algebraic equation.
- Avoid long and involved sentences. Economy of words is a virtue. Short words are preferable to long ones, common words preferable to obscure ones. Words should express exactly the concept or message which must be conveyed. A Thesaurus can help identify the best words for the purpose. All doubted meanings should be checked with a good dictionary; it is really astonishing the diversity of understanding which even the commonest words have among most people (e.g. acute/chronic). Avoid technical jargon as far as reasonably possible. It often has surprisingly limited currency.
- Put all the supporting documentation into appendices in order to keep the main text succinct. The main text of the report should ‘stand alone’, i.e. it should be a readable narrative with a clear and logical progression of its own without the need to refer to the appendices.
- Check spelling. (Computers can now do much of this automatically, although one needs to be aware of international differences.)

Some relevant publications on this topic are given below.[3]

24.1.2 Recent developments

24.1.2.1 Technical

An important development in recent years has been in the application of ‘solid-state’ devices and computers to electrical control. Even in the traditional areas of electrical power generation, transmission and distribution the use of microelectronics is now commonplace throughout all modern networks, from control rooms to protection relays. The reliability of most conventional hardware is now well established and it is possible to compute the probable ‘life before failure’ of most items of equipment. This, however, does not help to trace those defective components which operate only in emergency and could have remained unused and defective but undetected for years. It is virtually impossible to detect all weak links with certainty.

The weaknesses of the software in a computer system are even more difficult to detect. The embedded nature of software faults puts this type of fault into a special and sinister class of its own. However, the usefulness of computers makes it imperative that this weakness be understood and accommodated where computers are used in safety related applications. An enormous quantity of work has been done to get to grips with this problem and is the subject of comprehensive guidance and Standards.[4]

An associated development has been the increasing use of fibre optics for the transmission of information and instructions and the subsequent development of ‘optical’ switches and relays. These reduce fire and explosion hazards. They also eliminate the disturbance of control systems and telecommunications by electrostatic and magnetic induction and by gradients in the ground and structures caused by power faults and lightning discharges.

Another important development has been the realisation of the seriousness of the toxic hazards arising from the use of polychlorinated biphenols either through leakage into the environment or through the production of toxins arising from fires. These coolants were used increasingly from the mid-1950s to about 1975 to replace mineral oil in power transformers and capacitors to reduce the danger from fire and explosions. Oil-less, otherwise referred to as ‘dry’ or sometimes as ‘air cooled’, transformers have been used to some extent but, even if not immersed in an insulating liquid, conventional solid insulation is itself a fire risk. In addition, these transformers are prone to failure caused by moisture absorbed from the air and from surface contamination and tracking in industrial situations. Fire risks due to oil escape are an important industrial hazard and the losses have at times amounted to several million pounds per incident. Bulk oil type high voltage power switchgear has also been involved in many accidents, fires and explosions involving the combustion of the oil and its arc breakdown gases. There is now an understandable trend away from this type of switchgear equipment towards sulphur hexafluoride (SF_6) and vacuum technology at the lower power distribution voltages.

There have also been important developments in understanding the difficulties caused in control, telecommunication and instrumentation systems by electrostatic and electromagnetic radiation. Microelectronic ‘chips’ can be destroyed by the tiny energy levels in static electricity charges normally carried by the human body and special precautions are necessary in their handling. Electromagnetic radiation by electrical equipment and its susceptibility to such radiation is the subject of the Electromagnetic Compatibility (EMC) Directive 89/336/EEC.

24.1.3 Legal and administrative

The principal criminal legal provisions in Great Britain covering the safety of persons at work are contained in the Health and Safety at Work etc. Act 1974. Under this Act are gathered numerous other Acts of Parliament dealing with safety of the workplace and other, secondary, legislation such as Regulations. Of principal interest to engineers will be the Electricity at Work Regulations 1989 which apply to all electrical equipment used at work and all 'at work' activities. Other important regulations are; the Management of Health and Safety at Work Regulations 1999, the Provisions and Use of Work Equipment Regulations 1998, Personal Protective Equipment at Work Regulations 1992 and the Managing construction for health and safety, Construction, Design and Management Regulations 1994.

A breach of these regulations becomes a criminal matter which may result in prosecution of the offender (employer, employee or self-employed person) in the criminal courts. The maximum penalties are several thousand pounds in the lower courts, whereas in the higher courts, depending on the circumstances of the offence, the maximum penalties are up to 2 years imprisonment and/or an unlimited fine. Most cases are dealt with in the summary courts, e.g. Magistrates Courts. The enforcement authorities are the Health and Safety Executive (HSE) and the Local Authorities. Information on recent prosecutions and fines may be seen on the HSE website.

Electrical safety in situations other than those involving persons 'at work' can be subject to other legislation, for example the Consumer Protection Act 1987 and the Electrical Equipment (Safety) Regulations 1994 which were made under that Act. Those regulations govern the safety at point of sale of all domestic appliances which are sold to members of the public.

The safety of the public from the electricity distribution system in the UK is a matter for the Electricity Supply Regulations 1988, as amended. Those regulations relate to the Electricity Act 1989 and they place duties on the electricity power supply distribution companies in respect of electric overhead lines, cables and other apparatus. They are enforced by the Engineering Inspectorate of the Department of Trade and Industry.

The well-known Regulations for Electrical Installations published by the Institution of Electrical Engineers (the Wiring Regulations) are a code of good practice and are published as a British Standard, BS 7671:2001. These are not statutory regulations and are therefore not enforced through the criminal courts. Mostly they are used as a benchmark of safety in contracts for new installations of electrical fixed wiring. The sixteenth edition of these Wiring Regulations was first published in 1991 and the Regulations have been amended several times since with several Codes of Practice published in support.[5]

24.2 The nature of electrical injuries

24.2.1 Types of injury

Electrical injuries are of three main types: electric shock, burns, and falls caused by electric shock. There is a fourth category of very temporary discomfort or incapacity which is not serious but very painful while it lasts. This is conjunctivitis (or arc eye) which may be associated with shock and burn accidents, but is largely confined to electric arc welding.

24.2.2 Electric shock

Serious electric shock is almost entirely associated with alternating currents and is rare when low or medium voltage direct currents are concerned. Shock is not, however, a single phenomenon but is a general term for the excitation or disturbance of the function of nerves or muscles caused by the passage of an electric current. It is usually painful but is not necessarily associated with actual damage to the tissues of the body. The most common feature is severe stabbing and numbing pain at the points of entry and exit and sometimes along the path of the current through the body. This is frequently accompanied by involuntary contraction of muscles associated with the path of the current. It may even lead to torn muscles.

As a direct result of a moderately severe shock hand muscles may contract so tightly as to grip the conductor from which the shock current is being received leaving the victim quite unable to release their grip by voluntary action. This is an extremely dangerous situation and has resulted in many fatalities. Another possible result of muscular contraction is that the muscles of the chest, diaphragm and glottis may contract strongly and thus prevent breathing. This can lead to death by suffocation. Death may also be caused when breathing is stopped due to electric current passing through the respiratory control centres of the central nervous system.

24.2.2.1 Hold-on current and permissible leakage

A typical characteristic showing the relationship of 'hold on current' to the frequency of the a.c. supply (*Figure 24.1*) indicates that the lowest values of current occur in the band of between 10 to 200 Hz which includes normal mains operating frequency.

24.2.2.2 Ventricular fibrillation

It is generally believed that the great majority of fatal electrical accidents are caused by ventricular fibrillation, the muscles of the heart going into spasms which prevents the heart acting as an effective pump. Death follows quickly in these circumstances due to lack of oxygen supply to the brain.

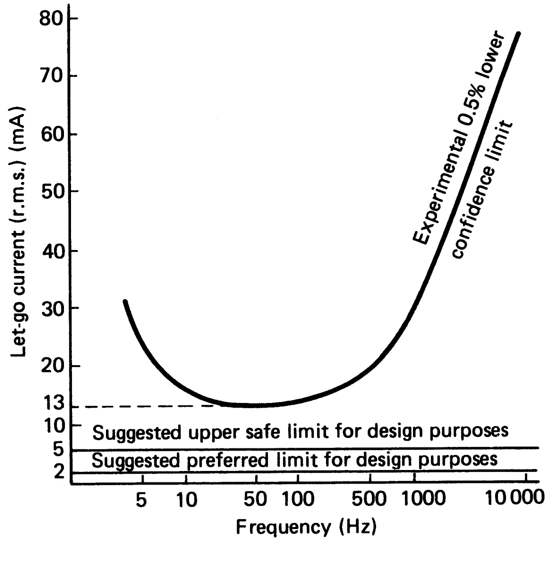


Figure 24.1 Relationship between the frequency and the let-go current based on experimental results. Such curves are of qualitative rather than quantitative value

A great deal of experimental work to investigate such effects has been undertaken on animals and this has been central to the production of international Standards documents on the subject of electric shock and its effects on humans and animals.[6]

24.2.2.3 *Limitations of experimental results*

Because of the practical and ethical difficulties in experimenting on humans and because there are substantial variations between humans in their susceptibility to electric shock, the results from such research must be interpreted carefully so that adequate margins of safety are built into equipment and work practices which depend on this knowledge. A 0.5% risk of death is too high to be acceptable (*Table 24.1*). There is an essential difficulty in extrapolating from animal to human subjects and it is also the view that the susceptibility of human subjects can be considerably affected by subjective considerations, there is evidence that the effect of an electric shock is greater when either the shock is unexpected or the person is abnormally afraid of electricity.

24.2.2.4 *Body resistance*

The impedance of the human body from hand to hand or hand to foot is variable. It depends on the area of contact and is affected by whether the hands (or feet) are dry, moist or wet, the condition of the skin and on the frequency and voltage of the current. Values are thought to range from 1–10 kΩ. These variations are discussed at some length in the international Standards on electric shock.

Table 24.1 Current (r.m.s.) (mA) to give rise to various physiological sensations with a.c. 50 Hz*

Physiological sensation	*Percentage of test subjects*		
	5	*50*	*95*
Current just perceptible in palms	at mA 0.7	1.2	1.7
Slight prickle in palms as if hands had become numb	at mA 1.0	2.0	3.0
Prickle also perceptible in the wrists	at mA 1.5	2.5	3.5
Slight vibrating of hands, pressure in wrists	at mA 2.0	3.2	4.4
Slight spasm in the forearm as if wrists were squeezed	at mA 2.5	4.0	5.5
Slight spasm in upper arm	at mA 3.2	5.2	7.2
Hands become stiff and clenched; letting go live parts is still possible; slight pain is already caused	at mA 4.2	6.2	8.2
Spasm in upper arm, hands become heavy and numb; prickle all over arm surface	at mA 4.3	6.6	8.9
General spasm of arm muscles up to the shoulders; letting go of live parts just about possible (let-go current)	at mA 7.0	11.0	15.0

* Current path: hand–body–hand.
From Friesleben and Fitzgerald.

24.2.2.5 *The limits of safety*

Recommendations about the safe operating limits are more useful for practical situations if stated in terms of supply voltages rather than 'hold on' or fatal currents, which are difficult to determine for real situations. *Table 24.2* gives information which is in accordance with experience and practice. There can be no exact determination of these limits, but they are given as a guide.

24.2.2.6 *Effect of frequency*

The results shown in *Figure 24.1* indicate that 50 or 60 Hz is almost exactly right to produce the maximum excitation of a nerve ending, but that the nerves could not respond to substantially higher frequencies. Radio frequency burns may be serious however and there are possible situations where such oversimplified rules might not apply.

24.2.2.7 *Respiratory arrest*

Experience of electro convulsive therapy indicates that, in the absence of severe damage to the nervous system (which is rare), respiratory arrest from a shock involving only the head is unlikely to persist, unless presumably the shock has lasted long enough to cause a dangerous reduction of oxygen in the blood (anoxaemia). Since head shocks are very infrequent this is not of great importance. Even the possibility of asphyxia in persons unconscious from electric shock appears to be uncommon. However, there is experience that artificial resuscitation does aid recovery, although the reasons for this are obscure. It may simply be that some external massaging of the heart is sufficient to transfer enough blood around the brain to sustain life even if the blood is depleted in oxygen. The recommendation has traditionally been that when a person is unconscious and not breathing artificial resuscitation should be immediately commenced. Both works 'first-aiders' and electricians should normally be instructed how to do this since a medical diagnosis is rarely quickly available and prompt action is essential. Artificial resuscitation should be continued until breathing is resumed or for a minimum of 1 hour. (The operator should then stand by to give further help if breathing falters). Knowledge of the effectiveness of artificial resuscitation in practice is necessarily based largely on records of such action. There is an 85% chance of recovery provided that resuscitation has not been abandoned after a period of less than 1 hour. The patient should be under close observation for some time after recovery, as there is some danger of relapse.

24.2.3 Other injuries

24.2.3.1 *Acoustic shock*

Accounts of the after-effects of lightning strikes frequently refer to temporary or permanent impairment of hearing and sometimes ruptured eardrums. This is in most cases almost

Table 24.2 Approximate threshold shock voltages at 50 Hz a.c.

Minimum threshold of feeling	10–12 r.m.s.
Minimum threshold of pain	15 r.m.s.
Minimum threshold of severe pain	20 r.m.s.
Minimum threshold hold-on volts	20–25 r.m.s.
Minimum threshold of death	40–50 r.m.s.
Range for fibrillation	50 or 60–2000 r.m.s.

certainly caused by the intense acoustic shock wave sent out when a column of air (the lightning channel) rapidly expands on being suddenly heated to about 15 000°C by the passage of the enormous current in a lightning discharge. The range of danger is probably limited to a few feet.

24.2.3.2 Arc eye

This is a painful condition resembling pepper in the eyes and develops some hours after exposure (even momentary) to an intense source of ultraviolet light. It occurs mainly when a person works near arc welding and looks directly into the brilliant light given out by the electric arc.

Fortunately 'arc eye' lasts only a short time, no more than a day or two, and although painful leaves no permanent injury. Treatment is by the application of a soothing lotion.

This complaint is easily prevented by the use of protective goggles with side protection, as worn by welders. The usual victims are assistants and bystanders without goggles, although ordinary glass cuts out much of the ultraviolet light.

Theoretically there is a risk of a form of cataract caused by prolonged exposure to infrared light which affects persons who work for long periods on glass furnaces, etc.

24.2.3.3 Fractures and torn muscles

Strains and fractures may arise from falls following an electric shock. e.g. from a crane or ladder. It may not be apparent at the time that the victim has suffered an electric shock and that the heart may be in fibrillation. If they are unconscious artificial resuscitation would be the advised first aid treatment.

24.2.3.4 Burns and side effects

Burns are probably the most serious after-effect of electrical accidents. They are the principal danger with direct currents or at very low voltages (below about 80 V). With low alternating voltages shock is the typical injury although there may also be severe burning. At extra-high voltages shock may not be as important it being the actual current and flash burns which tend to be severe with large areas of the body affected. Severe electrical burns have led to many deaths, usually after several days or even weeks of painful suffering. Burns may be of several types:

Contact burns These occur when the person has touched a live conductor. They may be local and very deep reaching to the bone, or very small, being just an area of 'white' skin which may be easily overlooked at a post-mortem examination. The position of such small burns may be important in reconstructing an accident and should be recorded.

Arc burns These may be extensive, and of any degree, particularly when there has been a high-voltage flashover. Provided that the person survives the initial wound and surgical shock, and the surface area involved is not too large, they are likely to make a good recovery because the injury should be largely sterile. They may, however, be badly scarred or even lose a limb. The large fault current levels which now exist on many low voltage distribution circuits poses a serious risk of arc burns if a flashover is caused. There have been a considerable number of fatalities to electrical staff due to this cause, usually when they have attempted to work on live low voltage (230/400 volt) busbars and switchboards without adequate training and proper insulated tools.

Radiation burns These burns arise from short-circuit arcing and are, in effect, a severe form of sunburn. Some radio frequency (RF) equipment can also impart burns which can be deep. RF burns usually occur due to contact with the charged conductor but this will depend on the power output of the equipment and the frequency.

Vaporised metal When an open fuse or small conductor fuses some copper (silver or tin) is vaporised and at close quarters this may burn or impregnate the face or hands. This is usually harmless unless it enters the eyes in which case the result is potentially serious.

Deep burns and necrosis There is the potential danger of deep burns destroying tissues below the skin even though superficially there is only a small injury. Thus electrical burns, and in particular high-voltage contact burns, must be taken seriously and the person kept under medical supervision. However, such burns are rare.

Metal fume fever This is caused by inhaling metal or metallic oxide fumes, e.g. by a welder working in an enclosed space.

24.2.4 Protection against electrical injuries

The obvious remedies for protecting personnel from electric shock, burns or radiation effects are to make live metal inaccessible and to keep persons separated from the dangerous plant, machinery or process. Another approach is to use protective equipment and defined work practices.

The Health and Safety Executive recommends many preventive measures in its various guidance documents but in particular it recommends the following through its Memorandum on the Electricity at Work Regulations 1989:[7]

- Using low (and safe) voltage.
- Insulating and/or enclosing live parts.
- Preventing conducting parts not normally live from becoming live by one or more of the following:
  - (a) earthing and automatic disconnection of the supply,
  - (b) double insulation,
  - (c) separating the supply from earth, and
  - (d) limiting electrical energy.
- Selecting equipment suitable for the environment in which it is to be used.
- Using equipment as defined in the maker's instructions.
- Ensuring that electrical equipment is adequately maintained.
- Avoiding the use of electricity altogether where its use would be dangerous.

The following aspects are also worthy of note:

- Metalwork in the vicinity of functional conductors may become live by electromagnetic or electrostatic induction. Serious or fatal shock is unlikely on most circuits although some serious accidents have occurred due to mutual induction on parallel running power lines.
- When work is to be carried out on (dead and isolated) high voltage power equipment conductors these conductors should be earthed. Potentially dangerous electric charges may otherwise accumulate on unearthed conductors either from the slow release of charge from the high

voltage insulation (dielectric relaxation) or from induction from adjacent circuits.

- It is possible to receive severe shocks from leakage over the surface of insulation at quite low voltages, e.g. 230 volt. Contamination of the surface of insulation by moisture and conductive salts can give rise to dangerous levels of surface leakage currents.

24.2.5 Toxic hazards

Toxic hazards arise as indirect results of electrical accidents. In this category are the hazards associated with the use of polychlorinated biphenols, which have been used in place of mineral oil in transformers, and those associated with carbon monoxide poisoning from incomplete combustion of insulation or oil.

Care is also needed in the handling of sulphur hexafluoride (SF_6) which has been degraded by arcing since toxic by-products may have been produced.

24.2.6 Conclusions

Electric shock is not a single simple phenomenon and is not perfectly understood. The majority of electrical accidents occur at the common domestic and commercial electricity supply voltages. i.e. 230 and 400 volts, and a high proportion of electrical accidents are serious or fatal. Artificial resuscitation for persons unconscious after a shock is effective and should be continued for at least one hour or until hospitalisation.

Electrical burns may cause death in extreme cases where they are extensive. In other cases however, although they may be deep they are largely sterile and therefore tend to heal quickly and well, although they may leave scars. Damage to muscles may be serious and amputation may be necessary in very bad cases. Attention should be paid to small burns caused by contact with high-voltage conductors because there may be serious deep-seated damage (necrosis) which is not visible.

24.3 Failure of electrical equipment

24.3.1 Causes

There are two fundamental causes of failure of electrical equipment, mechanical failure or electrical failure of insulation.

24.3.1.1 Mechanical causes

The safety of electrical equipment depends to a large extent on sound mechanical design. The majority of circuit-breaker failures are mechanical rather than electrical in nature. Typical faults are loose joints leading to overheating or arcing and the existence of voids and contamination in insulation causing arcing and breakdown products. Where the insulation is bulk oil the products of arcing are themselves highly flammable (acetylene for example) and have often led to explosions.

Fractures may be caused by resonant vibrations of current carrying conductors either from purely mechanical movement or from electromagnetic forces leading to fatigue hardening and subsequent breakage. Where metallic elements are stressed in a corrosive atmosphere (e.g. damp or polluted atmospheres) along with alternating forces, failure may occur at comparatively low stress. Some steels, which under normal conditions exhibit considerable ductility, will fail at low temperatures by brittle fractures with no ductile deformation.

Mechanical failure of insulators may displace conductors and cause short circuits. Ceramic insulators are brittle but have great strength in compression. However ceramic insulators are vulnerable where they are used in tension or shearing situations. They are now largely confined to outdoor overhead lines and switchgear where their robust construction makes them less susceptible to mechanical failures although they are then vulnerable to vandalism.

24.3.1.2 Breakdown of insulating materials

The electrical breakdown of insulating materials may also occur as follows:

- Mechanically, as by friction or tearing.
- As a result of excessive electrical stress.
- As a result of excessive temperature (and occasionally very low temperature) or temperature cycling. The latter may cause mechanical stresses as a result of differential expansion or contraction.
- Chemical and physical reaction with other materials, e.g. oxidation, contamination or the leaching out of important ingredients which may lead to de-plasticisation, i.e. they become brittle. The ingress of water is a very common contamination leading to 'treeing' and eventual electrical breakdown.

Failure is rarely the result of inadequate electrical breakdown strength where reasonably pure materials are used. In practice, insulation is rarely designed to be stressed to more than 10% of its strength as determined by laboratory tests. It fails because of impurities, lack of homogeneity, the unavoidable variations in commercially available materials as well as in those natural products such as paper, wood and petroleum products. The insulation performance of most commercially used materials is now well documented and standard testing procedures have been established.

24.3.2 Particular equipment

24.3.2.1 Switchgear

Hazards associated with switchgear failure include fire, explosion and electric shock. There is a particular risk to the technical staff who require to operate high voltage switchgear and to conduct testing and maintenance on this equipment. Unfortunately oil circuit breakers and fuse switches have been the cause of many serious accidents and fatalities have not been rare. Some of the most expensive fire losses have also been caused by switchgear.

An important proportion of these accidents and losses are caused by failure of circuit breakers to operate correctly but many accidents are caused by mechanical problems with auxiliary equipment such as isolators and from failure of routine and emergency operating procedures. The causes of high-voltage switchgear failures are typically as follows.

- Poor maintenance leading to contamination of insulation and loose connections etc.
- Incorrect or inappropriate use of test equipment on live equipment.
- Unrestricted repeated operations of oil circuit breakers (OCBs) leading to breakdown of the oil insulation and/or to contact collapses.

- Hesitant operation of manually closed switches onto faulted circuits leading to panic opening on fault and the consequent fault arc drawn at the contacts of a non-rated switch.
- Failures of PTs (potential transformers, often referred to as VTs or voltage transformers) and failures of CTs (current transformers).

24.3.2.2 Transformers

Fires and explosions in oil-immersed power transformers have been less common than in oil-immersed switchgear and the immediate results are, on the whole, less damaging. However, because of the large amount of oil which may be released and ignited, oil-immersed transformers are potentially at least as dangerous and the consequences in terms of loss of power availability and pollution can be considerable.

Most discussions of transformer failure relate to interturn faults and their causes, but these cause a comparatively small number of fires, etc. Many incidents are generally associated with automatic tap-changing gear, while bushing failures, flashover and arcing at the transformer top, etc. (including some lightning damage) cause about the same number of incidents. Major transformer failures are comparatively rare events.

The main dangers are the spread of fire by the release of a very large volume of burning oil and the emission of clouds of black smoke but there may be secondary hazards such as plant being stopped by damage to cables and protection, control and alarm equipment even when the fire itself was not initially very large.

Typical failures and their causes are described below.

Lightning and voltage transients These may cause external flashover of bushing insulations which may shatter or split oil-filled insulators, causing fires. If the voltage surge reaches a transformer winding it may cause a flashover above the oil and damage the winding insulation.

Internal flashover above oil This can be caused by voltage transients as described above but there are other causes which are more easily eliminated. These are arcing caused by fractured conductors, sparking at loose connections, and badly made joints which provide ionisation above the oil level. Hot joints and connections are also suspected causes.

Interturn faults Whatever the cause such faults are likely to develop slowly. The volts per turn are normally quite low (say 5–20 V for transformers below 2 MVA) and will not sustain continuous arcing until a number of turns become involved but this may not apply where, in modern transformers, coils and turns are interleaved without spacers to improve surge voltage distribution. Overheating can be caused by short-circuited turns. Failures may go undetected by conventional protective relays as there is likely to be no significant change in through current. However some interturn faults can be detected by certain forms of monitoring and testing.

Failure of the insulation of the magnetic circuit Such failure, e.g. of core frame, or clamping bolts, or between laminations, can allow parasitic eddy currents and local heating. The former may cause sparking and the evolution of arc gases. Such faults do not lead to immediate danger but contribute to the deterioration of insulation and oil. Also an accumulation of arc gas (particularly H_2 and C_2H_2) in the airspace above or dissolved in the oil is a concealed explosion risk. Other faults which occur from time to time are due to poor hygiene and untidiness during manufacture and maintenance. Nuts, bolts, cut-off ends of wire and even spanners have been found wedged between windings. These can cause local stress concentration and heating with possible mechanical damage to insulation.

Overloading and 'through' faults These will overheat the windings and cause cumulative damage to the insulation. They may not cause immediate danger unless they persist but the life of the transformer will be shortened.

There are advantages in using sealed transformer tanks. Experience indicates that in the absence of air (oxygen) the oil deteriorates much more slowly and less maintenance is necessary. There are drawbacks however. Unless there is an adequate space above the oil, which can be filled with inert gas, e.g. nitrogen, special means must be provided to allow for the expansion of the oil with rise of temperature, otherwise the tank may fail under the hydrostatic pressure developed. For this purpose internal 'bellows' or flexible diaphragms are sometimes provided.

Also, a sealed tank, unless it has a separate expansion chamber or conservator at the top, cannot be fitted with a Buchholz relay and it may be impossible to give full protection against slowly developing faults. A sealed tank should therefore be provided with a bursting disc or similar device as a major fault between phases or to earth, caused by lightning, fractured conductors, loose connections, etc., may otherwise cause the tank to rupture or the lid to be sprung before differential or overcurrent relays have time to operate the circuit-breakers.

Such faults are rare, and it may be that there are situations where the risk can be accepted, but this would not be the case for very large transformers, transformers in occupied buildings, or high fire-risk situations.

24.3.2.3 Cables and installations

Traditionally, high voltage distribution cables, if they were buried in the ground, were impregnated paper (tape) insulated, lead covered, steel wire armoured and served (i.e. protected with a sheath of material resembling bitumen impregnated sacking) and frequently further protected by ceramic tiles a few inches above the cable. Modern cables comprise plastic materials with metal extruded or wire armouring. Cables often have a life measured in many decades and many cables of early style construction may be found in use, particularly on public electricity distribution systems.

Medium voltage (400/230 V) distribution cables were steel tape armoured but sometimes tough rubber or similar sheaths were used in ducts. Such cables pose a very low fire risk. Modern cables mainly use plastic insulation. The main hazards are:

- Burst cable boxes at transformers, switchgear and other joints. Link, or street mains disconnecting, boxes which use the traditional bitumen insulation have been the subject of many incidents of failure, mainly due to voids in the bitumen, water ingress and to loose connections.
- Puncture by pneumatic road drills which causes short circuits and a shower of sparks, but which is often more alarming than dangerous because the sheath and armouring provide an efficient earth-return path. There have been incidents of serious burn injuries too however, the severity of the arcing being dependent on the power fault level of the transformers feeding onto that particular circuit. In cities these fault levels are often very high.
- The spread of fire in ducts or tunnels.

The evolution of clouds of black smoke from cable oil is a serious hazard because it hampers effective fire extinguishing operations, contaminates buildings and may result in the need for extensive cleaning and renovation work.

Although plastic insulation and sheaths such as polyvinyl chloride (PVC) are now common, and difficult to ignite, they will nonetheless burn if in adequate bulk and sufficiently preheated locally. However, when ignited PVC produces clouds of dense black smoke, as do some 'rubbers'. There is then a substantial liberation of hydrogen chloride, which may corrode sensitive electrical equipment, e.g. in telephone exchanges and computer departments, and steel reinforcement bars in concrete.

Important cable tunnels should be provided with fire detection and monitoring equipment and possibly fixed fire extinguishing installations.

The relative positions of power, telecommunication, control, and alarm and safety circuits must be considered, to ensure that power short circuits and fires do not damage safety circuits.

Cable tunnels are perhaps more vulnerable to fortuitous 'external' fires than to short circuits, etc. The experience of the London Underground system shows that cable fires along the tunnels are likely to result from accumulation of rubbish ignited probably by sparks from the collector shoes of passing trains.

Inside industrial buildings, cables are usually drawn into conduit or laid in ducts or on trays. The possible spread of fire must again be considered. Attention must also be paid to the possibility of damage by vehicles such as forklift trucks and cranes.

Supply to 'island sites' in a work area presents difficulties. Under-floor ducts are subject to flooding and the entry of rubbish, flammable liquids and vapours. Overhead distribution busbars along a line of machines are frequently used, particularly where the position of machines may be changed at short notice when there is a change of product or method. The conductors are usually, but not invariably, covered with insulation, except at fixed points where provision is made for tapping off, and the whole assembly of busbars is enclosed in suspended metal trunking with provision for 'plug and socket' attachments or spur switch and fuse combinations. The construction must be considered in relation to possible damage by cranes, forklift trucks, etc., and the entry of water from a leaking roof or sprinkler systems.

It is preferable to make the connections to individual machines by cables with a flexible outer metallic screen, pliable armoured cable or by flexible conduit.

24.3.3 Protection against failure

24.3.3.1 Earthing principles

Earthing is a complex subject. Its complexity should be apparent from a cursory reference to the current British Standard on the Code of practice for earthing which runs to 84 pages.[8] In the context of the safety of most electrical installations however there are three main points to be made.

- It prevents the outer conductive casing of apparatus and conductors from assuming a potential which is dangerously different from the surroundings for any appreciable length of time although if a fault does occur inside the equipment the outer casing may actually be raised to a dangerous potential momentarily.
- The earthing connections must be sufficiently robust and conductive to allow sufficient current to pass safely in order to operate overcurrent or other protective devices promptly, including those for earth-leakage protection. Protective devices must be suitably rated for the fault currents which might be expected.
- Care must be taken against the occurrence of dangerous earth-potential gradients due to fault currents flowing in the terminations made to the general mass of earth. The 'resistance areas' of the earth electrodes of different systems should not be permitted to overlap to any significant degree. The purpose of this is to prevent faults on one system imposing dangerous voltages onto the protective (earthing) conductors of other systems.

Earthing is also used for functional purposes, for signalling and draining leakage currents. The needs of these uses can sometimes conflict with the safety objectives of earthing.

24.3.3.2 Interlocks and guards

Different classes of guards are described below.

Fixed guards By virtue of their position, these prevent access to dangerous parts.

Automatic guards These move in advance of each operation or stroke of a machine and sweep or push the operator's arm or person out of the way before the stroke is made, or automatically take up a position similar to a fixed guard before the danger can arise. They are sometimes used where material or blanks are fed by hand into the machine. They are best combined with trip guards, but are largely being replaced by interlocked guards.

Interlocked guards These guards are moveable and whose movement is interconnected with the power or control system of the machine. They must be correctly adjusted before any potentially dangerous operation can commence.

Trip guards These stop the machine or automatically cause other appropriate action to be taken immediately danger arises. They can often be combined to advantage with automatic guards or with positional and distance guards.

Positional and distance guards These are often a rather elementary form of fixed guard, so placed as to keep an operator or other person at a safe distance from a machine which cannot be enclosed because of the nature of the material being handled or the work being done.

Electrosensitive safety devices 'Curtains' of interlaced light beams are used extensively on machines such as power presses, bending brakes and guillotines, where the size and shape of the workpieces prevent complete enclosure with fixed guards. The risk of failure with such machines can lead to serious injuries of amputation and crushing so this demands that the design of these guards be a specialised matter with check circuitry and timing requirements to prevent failures to danger. There are international Standards which specify the safety integrity for these devices.[9]

24.3.4 Fires and explosions

Electrical installations are often cited as the cause of fires but this is probably due in large part to the difficulty in determining the actual cause of many fires. In effect electricity is an easy candidate for blame in cases of doubt when the more likely causes probably lie with carelessly discarded smoking materials and other banal causes. However electricity can cause fire in a number of ways.

Serious overloading is necessarily a fire risk, since most insulating materials are in some degree flammable (ceramics, minerals such as asbestos or alumina, and mica being the chief exceptions). Overloading or resistive heating at poor electrical joints and connections may also ignite other flammable material in the immediate neighbourhood, for example wooden and similar panelling over cable runs has caused disastrous fires in shops, theatres and other entertainment centres. A common cause of fires however arises through the inappropriate use of electrical heating appliances which simply radiate heat into materials until they ignite. Radiant electric fires, convector heaters, electric irons, cookers and incandescent lamps are involved.

Fires are also caused through arcing or sparking release of electrical energy. The use of electrical equipment in areas where flammable or explosive atmospheres may exist is the subject of extensive expertise and availability of specialist apparatus. Coal mines (which have firedamp, i.e. methane), oil refineries and chemical plants, are examples where this specialised apparatus is required. In the EU it is also the subject of the ATEX directive 94/9/EC.

Further reading and sources:

HSE and HSC publications on electrical safety

HSR 25 *Memorandum of guidance on the Electricity at Work Regulations 1989*. HMSO
GS 6 *Avoidance of danger from overhead electric lines.*
GS 38 *Electrical test equipment for use by electricians.*
HSG 47 *Avoiding danger from underground services.*
HSG 85 *Electricity at work—safe working practices.*
HSG 141 *Electrical safety on construction sites.*
HSG 180 *The application of electrosensitive protective equipment using light curtains and light beam devices to machinery.*
HSG 204 *Health and safety in arc welding*

Websites

Health and Safety Executive. http://www.hse.gov.uk
Department of Trade and Industry. http://www.dti.gov.uk
British Standards Institution. http://www.bsi.org.uk
European Commission. http://www.europa.eu.int
European Agency for Safety and Health at Work. http://europe.osha.eu.int
HM Stationery Office. http://www.hmso.gov.uk
D A Dolbey Jones. http://www.ukelectricalsafety.com

References

1 FORDHAM-COOPER *Electrical Safety Engineering*, 3rd edition, Butterworth-Heinemann, 1993, Chapter 1 Revised by Dolbey Jones, D. A.
2 GOWERS, E. (Ed.), *Fowler's Modern English Usage*, Oxford University Press, Oxford (1965)
3 GOWERS, E., *Plain Words*, Penguin, Harmondsworth
4 BRITISH STANDARDS INSTITUTION BS IEC 61508 *Functional safety of electrical/electronic/programmable electronic safety-related systems*
5 Requirements for Electrical Installations, BS 7671:2001 (IEE Wiring Regulations Sixteenth Edition) published jointly by the Institution of Electrical Engineers and the British Standards Institution
6 Guide to effects of current on human beings and livestock. IEC 60479:1998 (In three parts). Published by the BSI as PD 6519
7 HEALTH AND SAFETY EXECUTIVE, Memorandum of Guidance on the Electricity at Work Regulations 1989, Publication HS (R) 25, HMSO, London
8 BRITISH STANDARDS INSTITUTION BS 7430:1998, *Code of practice for earthing*
9 BRITISH STANDARDS INSTITUTION BS 60204-1:1993 (and 1998) *Safety of machinery. Electrical equipment of machines*, PD 5304:2000 *Safe use of machinery* and BS EN 954-1, *Safety of machinery—safety related parts of control systems*

25

Hazardous Area Technology

B Parkinson

Contents

25.1 A brief UK history 25/3

25.2 General certification requirements 25/4

25.3 Gas group and temperature class 25/5

25.4 Explosion protection concepts 25/5
- 25.4.1 Intrinsic safety ‘i’ 25/5
- 25.4.2 Flameproof enclosure ‘d’ 25/6
- 25.4.3 Increased safety ‘e’ 25/7
- 25.4.4 Purged and pressurised ‘p’ 25/7
- 25.4.5 Quartz filled ‘q’ 25/7
- 25.4.6 Oil Immersion ‘o’ 25/7
- 25.4.7 Encapsulation ‘m’ 25/7
- 25.4.8 Non sparking ‘n’ 25/7
- 25.4.9 Special protection ‘s’ 25/8
- 25.4.10 Mixed concepts 25/8
- 25.4.11 Dust protection 25/10

25.5 ATEX certification 25/10
- 25.5.1 Example of an ATEX label 25/11

25.6 Global view 25/11

25.7 Useful websites 25/12

25.1 A brief UK history

In the early days of coal mining (with candles or oil lamps as the only form of lighting) it was soon learnt that explosions could be caused if an accumulation of gas occurred when working a coal seam. It was also soon realised that the gas tended to collect along the top of the tunnels, so someone (the original fireman!) had the unenviable job of crawling along the tunnels in advance of the miners, carrying a candle on the end of a pole (*Figure 25.1*). They could reach to the roof and if gas was present at that level, it would hopefully be ignited and burnt off without major problem, thus rendering the area safe to work in. This system was by no means foolproof, and on occasions when a large concentration of gas was encountered there could still be a major fire or explosion. The Davey lamp was the first major step forward in this situation. Its design was such that the flame within the lamp would not propagate through the metal gauze surrounding it so it would not ignite the surrounding atmosphere. At the same time, the colour of the flame showed when the explosive gas was present, so giving a good warning to take extra precautions.

With the advent of electricity and resultant mechanisation, more thought was needed in guarding against the ignition risks from electrical or frictional sparks or hot surfaces, and a major mining disaster around 1913 served to focus attention even more on the risks. Around this time the safety in mines research establishment (SMRE) was formed, and located near the small village of Eskmeals on the Cumbrian coast, but hopefully far enough away so as not to disturb people with any test explosions which occurred.

In the mid 1920s SMRE moved its testing facilities to Buxton in Derbyshire, again placed on a hill out of town so that any explosions would not worry the neighbours. In mining, both flameproofing and intrinsic safety gained ground as accepted protection techniques, in combination with forced ventilation to keep the gas concentration below explosive levels whenever possible. At the same time as SMRE assisted with the safety of equipment in the mines, the board of trade (BoT) looked after the safety of surface industries such as chemical works and oil refineries. As electrical installations became more complex, a system of issuing certificates for both apparatus and intrinsically safe systems was developed, to ease the job of inspectors checking an installation. The first standards used were some early British Standards such as BS 229 and BS 1259, which were very brief and general in their wording, so that they almost let the certifying officer make up his own rules for certification. At this time (the 50s and 60s) the certification was still carried out by the factory inspectorate, and gradually the proliferation of electrical installations led to increasing delays in inspecting sites. In 1967 the British Approvals Service for Electrical Equipment in Flammable Atmospheres (BASEEFA) was formed to take over the basic certification of apparatus from the factory inspectorate and so ease their load.

In the early 70s BASEEFA produced its own more detailed standards, in particular for intrinsic safety, special protection, and gas detectors using pellistors (both for explosion protection and for performance). Then in the mid 70s, the major explosion and fire at Flixborough focused attention even more on the need for safety in electrical equipment, the influence of the European Union was gaining ground, and a series of harmonised standards (the EN 50014 to EN 50020 series plus one or two others) was evolved for use by all the member states. These standards were published in 1977 in the UK by BSI, as BS 5501 Parts 1 to 7 and others.

When BASEEFA was first set up the certification went hand in hand with a licence to reproduce the BASEEFA mark, which in turn was conditional on the manufacturer having an acceptable quality control system in place to ensure the certified equipment was made correctly. Also with the advent of the EN 50014 series of standards, there was a general comment in them which placed a responsibility on manufacturers (and arguably on certification bodies too)

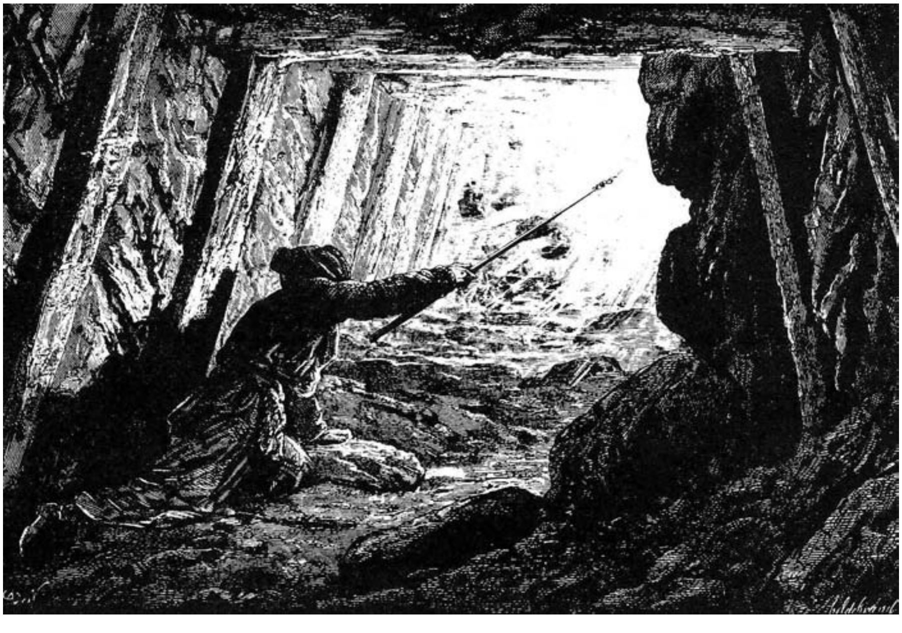


Figure 25.1 Fireman igniting firedamp. Illustration courtesy of HSE

to ensure product was made correctly. The BASEEFA licence was issued in conjunction with a surveillance visit (or audit) as a third party check on production quality. Whilst such audits were common with regard to American organisations such as Factory Mutual (FMRC) or Underwriters Laboratories (UL), they tended to be the exception in Europe, with BASEEFA being possibly the only certification body carrying out such visits, and pressure mounted from some quarters for these audits to be drastically reduced if not phased out completely. However, at least one of the large user companies became very dissatisfied with the acceptability of goods it was receiving from various companies both in the UK and the rest of Europe, and so led a push to reinstate and reinforce the quality audit situation. The problem was not poor quality of equipment, but that well intentioned changes to the design to improve functionality or ease manufacture sometimes resulted in non-compliance with the strict interpretation of the standards and hence invalidation of the certificate. The worry was that sooner or later such a problem might go beyond a technical non-compliance and actually result in an unsafe piece of equipment being installed somewhere. This resulted in BASEEFA launching its conformity assurance programme for quality auditing. Now as we enter the 21st century quality control in general has achieved prominence, and the new ATEX directive legislation has introduced a clear quality module into the certification process, so that quality requirements will be recognised and enforced throughout Europe. This is mentioned again later on.

For many years various testing and certification bodies throughout the world have met at regular intervals to exchange knowledge and experiences, and this has led to a gradual convergence of the various explosion protection standards. The latest editions of the EN 50014 series of standards (now numbered by the British Standards Institution as BSEN 50014, etc.) are the basis for ATEX directive certification, and greater alignment is also being achieved with the IEC 60079 series of standards which will eventually open the way for global certification with the IECEx scheme. One immediate advantage to users of certified equipment is that the marking of apparatus will become very nearly the same anywhere in the world. These will be mentioned in more detail later in this chapter.

25.2 General certification requirements

Most countries in the world have government regulations, or general requirements laid down by insurance companies, or both, requiring workplace managers or owners to take reasonable precautions to ensure the safety of their employees and the general public. For example in UK there is the health and safety at work act 1974 plus the working regulations 1996 act, or in the USA, OSHA (the Occupational Safety and Health Administration) which is one of the US Government Agencies.

The need for using certified apparatus in a particular situation depends on several factors, such as the quantity of flammable material present, how the material is handled and contained, what ventilation exists, and how large the work area is in comparison to the quantity of flammable material.

The explosion triangle shows the elements needed to create an explosion (*Figure 25.2*). The first obvious means of explosion protection is to keep ignition sources away from the flammable material. The next can be to ensure that there is plenty of ventilation, so as to keep the air fuel mixture below the lower explosive limit. Some people would suggest that ensuring air is totally excluded so that there is always 100% fuel is also a 'safe' situation to aim for, which in turn would allow invasive probes to be included directly into the process system without any need to use certified types. This is debatable; as although some countries' test authorities will accept it as a safe situation, others would argue that you could well at some time run into a potentially explosive situation if there was a problem with the process and air entered during the repair period. Then great care would be needed when refilling the system to ensure all air was flushed out before re-energising any non-certified invasive probes.

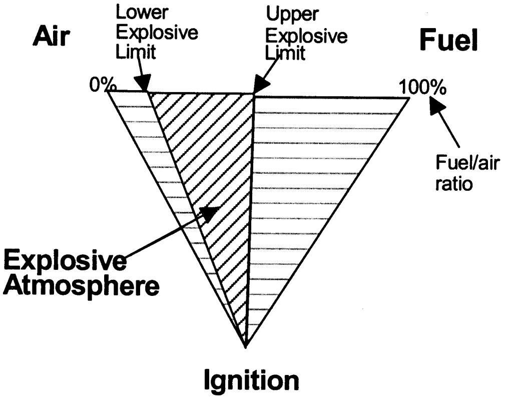


Figure 25.2 The explosion triangle

As mentioned earlier, in mining, forced ventilation is used to dilute any gas and keep the mixture well below the explosive limit. In surface industries, natural ventilation may be sufficient much of the time, especially if the work area is large compared to the amount of flammable material being handled. Another factor in siting equipment can be the density of the flammable mixture, which would determine where the mixture might concentrate most (floor or ceiling). For further guidance on assessing the disposition of hazardous areas, see BSEN 60079-10.

The probability of an explosive atmosphere being present is based on a system of zoning.

Zone 0 is where a hazardous (or explosive) atmosphere may be present continuously or for long periods, for example within or immediately around a vessel containing a flammable liquid, or within the pipework of some process.

Zone 1 is where the explosive atmosphere can be present in normal operation, but for a more limited period of time. This could be where (normally sealed) containers of flammable liquid are kept and at regular intervals the liquid has to be poured, thus releasing vapour into the surrounding atmosphere. A common rule of thumb is to say if the vapour is likely to be present for less than around 100 to 1000 hours per year then this constitutes a zone 1 area as opposed to zone 0.

Zone 2 is where an explosive atmosphere is not normally present but could occur under fault conditions. This could be a storage area where the flammable material is kept in sealed containers, but spillage could occur if a container was accidentally dropped and ruptured.

In America and Canada, the area classification up to now has been by division. In this system division 2 is very similar to zone 2, but division 1 encompasses both zone 0 and zone 1.

The zoning system is beginning to be used with the advent of the latest standards discussed later, but it remains to be seen how long it takes to gain general acceptance in those countries.

25.3 Gas group and temperature class

The next factor which comes into play is the sensitivity of the explosive atmosphere to possible ignition sources. Materials are classified by minimum ignition energy (m.i.e.) based on how easily they will ignite due to a spark, maximum experimental safe gap (m.e.s.g.) based on how easily they will transmit a flame through a small gap, and auto ignition temperature based on how easily a hot surface can ignite them. The m.i.e. and m.e.s.g. tend to go fairly closely together. For example hydrogen is the most onerous gas in this respect, and both the permitted energy in intrinsically safe equipment and the permitted gaps in flameproof equipment are the smallest. The auto ignition temperature, however, tends to work independently. Using the hydrogen example, its ignition temperature is around 560 °C, so hot surfaces are not a particular problem, whereas many other explosive mixtures could be ignited by temperatures around say 250°C or 300°C.

The gases, based on m.i.e. and m.e.s.g. are classified into subdivisions which align with the apparatus groups shown in *Table 25.1*. BSEN 50014 and BSEN 60079-0 both contain information on the subdivisions of many compounds, and includes IEC/TR3 60079-20 also include auto ignition temperature information for a large range of compounds. The auto ignition temperature is used to determine the suitability of apparatus for use with a particular gas, by reference to the apparatus temperature class. For compounds not listed, the health and safety laboratory or other specialist laboratories may be able to provide advice or testing to allocate the material to the appropriate group or temperature class.

In the USA and Canada the group marking usually lists all the group letters applicable, such as ‘for group D’ or ‘for groups A, B, C and D’, whereas in the group marking for european or IEC, a mark of IIC automatically implies that the apparatus is also suitable for IIB and IIA, and IIB implies suitable for IIB and IIA.

Note that acetylene is listed separately in the *Table 25.1* the reason for this will be explained when discussing flameproof equipment later in this chapter.

Temperature class sometimes causes confusion, so hopefully this explanation will help to clarify things. The temperature class marked on a piece of apparatus gives indication of the likely maximum temperature which will be reached in service (either in normal operation or under certain fault conditions). In the absence of anything more than the temperature class, the implication is that this will be the maximum temperature of the apparatus when located in a surrounding ambient not exceeding 40 °C. For other ambient temperature ranges and extra marking such as $(-20\,^\circ\mathrm{C} < \varsigma T_a < \varsigma + 60\,^\circ\mathrm{C})$ will be included. Now the gas mixture of concern will have either a suitable temperature class or an auto ignition temperature (or both) shown, and provided the temperature class of the apparatus is equal to or ‘better than’ that for the gas mixture, then the apparatus is suitable. For example, if a gas mixture was shown as requiring T_2, then apparatus of T_2 up to T_6 would be suitable, but *not* T_1, or if the mixture was quoted as having an auto ignition temperature of 180 °C then apparatus of T_4, T_5 or T_6 would be suitable, but *not* T_1 to T_3.

Table 25.1

Representative gas	*Group (European or IEC)*	*Group (USA and Canada)*
Acetylene	IIC	A
Hydrogen	IIC or IIB + H2	B
Ethylene	IIB	C
Propane	IIA	D
Methane	I	

Table 25.2

Temperature class	*Temperature of apparatus* (°C)
T_1	450
T_2	300
T_3	200
T_4	135
T_5	100
T_6	85

In America and Canada, when working to their earlier standards, some temperature classes may be further subdivided with a suffix, such as T_{2A} to T_{2D} which refer to intermediate temperature steps. They are omitted from this description for simplicity, but full details can be found if required on the appropriate websites.

25.4 Explosion protection concepts

Table 25.3 repeats the zoning mentioned earlier and shows the types protection permitted in each zone.

25.4.1 Intrinsic safety ‘i’ (which may be subdivided into ‘ia’ or ‘ib’)

This technique is used mainly for electronic equipment such as level, density or temperature transmitters, toxic or combustible gas detectors, or proximity detectors, or portable radio transceivers. The power levels involved are relatively small, usually less than 1.3 W, but in some cases up to around 5 or 6 W. For fixed apparatus fed from a separate power source, both the hazardous area apparatus and the

Table 25.3

Zone	*Division*	*Suitable type of protection*	*ATEX category*
0	1	‘ia’ ‘s’ when specified as being suitable for zone 0, mixed concepts covered by BSEN 50284 (only used in ATEX certificates)	1
1		‘ib’, ‘d’, ‘e’, ‘p’, ‘q’, ‘m’	2
2	2	‘n’, ‘N’	3

power source itself are designed so as to ensure that under fault conditions, any spark energy or surface temperature is insufficient to cause an ignition of a surrounding explosive atmosphere. The spark energy has an extra safety factor applied to it, and the temperature is assessed under worst case fault conditions to provide a safety factor there also. The power source may be a custom designed power supply or control unit, or a general purpose zener barrier or repeater power supply. 'ia' relies on a two fault analysis and is suitable for use in zone 0, and 'ib' relies on a one fault analysis and is suitable for zone 1. This type of equipment may also be self contained battery powered (*Figure 25.3*).

For the fixed installation with separate power supply, the cabling linking the two is specified in terms of capacitance, inductance and inductance to resistance ratio, but otherwise can be almost any lightweight cable, which can save on installation costs. Also the equipment tends to be very light and manageable in comparison with other protection concepts, because the enclosure is basically only required to keep the circuitry clean rather than being a vital part of the protection concept. Another advantage is that 'live maintenance' is possible, whereas with other protection concepts the equipment usually must be de-energised before any work is carried out. The consideration to be offset against this is that both the power source and the hazardous area equipment must be certified, and in general only equipment operating at 30 V or lower and 5 W or lower is available.

This equipment is sensitive to gas group, and may be certified for IIC, IIB, IIA or I. As hot surfaces may be present, the temperature class also needs checking against the gases present where it is used (*Figure 25.4*).

25.4.2 Flameproof enclosure 'd'

This technique is often used for such things as luminaires, valve actuators, high powered control units and heavy duty motors and pumps. It is also used for level, density or temperature transmitters of similar size to those certified 'i', but in situations where for various reasons the user happens to prefer the 'd' concept. In this technique the enclosure is designed to be strong enough to contain an internal explosion, and with covers, cable glands, etc. arranged such that any gaps (or flamepaths) to the outside world are small enough to ensure that the internal explosion will not be transmitted to the surrounding atmosphere. Testing for non-transmission and the pressure test both have an extra safety factor applied compared to normal operation, to give the required integrity. The obvious advantage compared to 'i' is that high powered equipment operating up to several hundred volts can be protected 'd'. To be offset against this is that even small equipment can be relatively bulky and heavy, and the cabling always needs to be well protected, either by use of armoured cable or by use of conduit or other mechanical protection (*Figure 25.5*).

Acetylene is dealt with as a separate item in the flameproof standards because it has a unique characteristic compared with other flammable or explosive mixtures which causes problems with this equipment. If the acetylene mix is very rich, it produces carbon particles, and whilst a properly dimensioned plain flange will stop normal flame propagation, it does not prevent incandescent particles being ejected. Thus acetylene requires the use of such things as spigot joints or threads to provide added integrity.

This equipment is also sensitive to gas group, and may be certified for IIC, IIB, IIA or I, and again the temperature class needs checking for suitability. With this type of equipment, the temperature of internal components may be much higher than the external surface of the enclosure, which is permissible in normal use, but can result in extra conditions being imposed when it is necessary to open the equipment, such as de-energising and waiting a specified minimum time before opening.

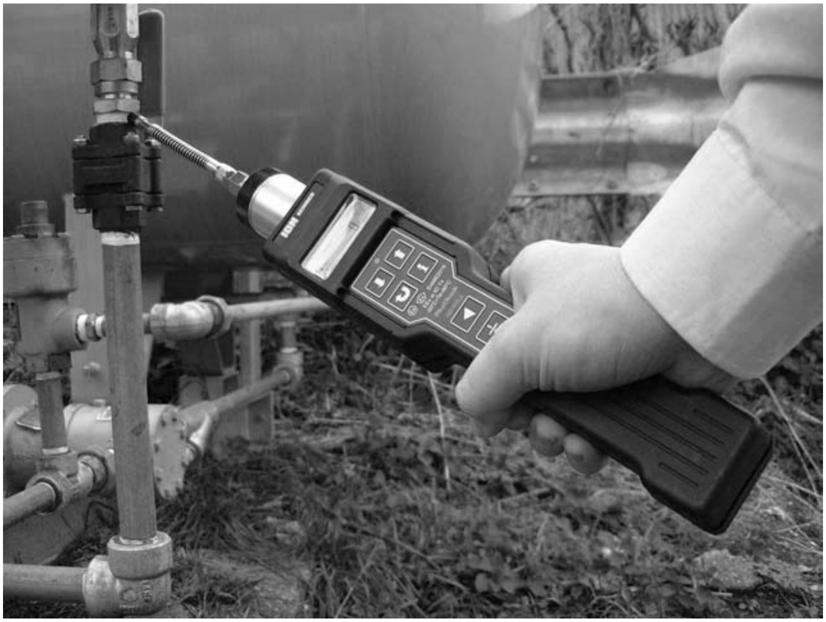


Figure 25.3 An example of a battery powered hand held intrinsically safe instrument. Photograph courtesy of Ion Science Ltd

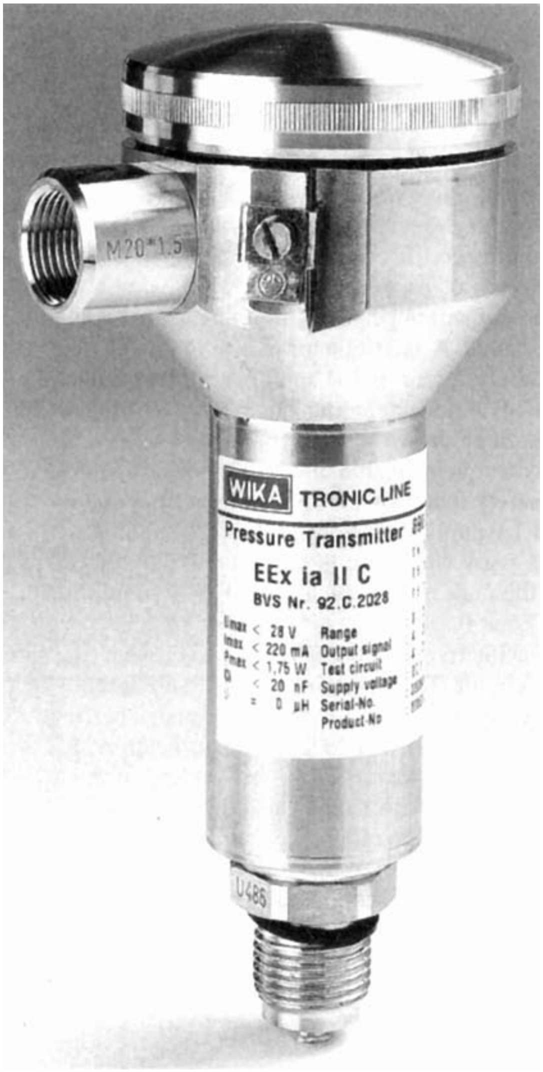


Figure 25.4 EEx 'ia' field-cased pressure transmitter (courtesy of WIKA Instruments Ltd)

25.4.3 Increased safety 'e'

Typical equipment here includes luminaires, induction motors, pumps, etc. This technique relies on ensuring that all components are run well within their capabilities so as to avoid hot surfaces, and connections are arranged to be of high reliability so that they will not produce sparks. Again the advantage here is that high powered equipment can be protected this way, the weight and bulk tends to be somewhat less than 'd' equipment, but still the cabling needs to be well protected.

This equipment is not sensitive to gas group as it is essentially non-sparking, so the group marking is likely to be just II implying any gas group, however the temperature class may still vary and so need checking against the gases present in the location where it is used (*Figure 25.6*).

25.4.4 Purged and pressurised 'p'

This technique relies on keeping the explosive atmosphere away from the live electrical circuits, by dilution with air or inert gas. Depending on the nature of the equipment being protected, either the unit may be a sealed enclosure (with probably an air or inert gas bottle attached) with a very low leakage rate, or it may be deliberately ventilated by means of a fan bringing in air from a 'clean' location. In each case the technique relies on monitoring the internal pressure of the unit, or the flow rate, or both, and cutting off the electrical supply to the unit if the pressurisation fails. When starting up the system it is necessary to monitor the pressure and/or flow and ensure the enclosure is up to pressure and has been flushed 'clean' (or purged) before re-energising the electrical circuits. The control equipment is often used in conjunction with some 'i' or maybe 'd' circuits, to enable fully automatic start-up.

Basically while the pressurisation system is operating, neither the gas group nor the temperature class are of any consequence in respect of the items within the enclosure, but clearly any residual high temperatures will be of concern if the system fails, and any associated 'i' or 'd' equipment will need to be chosen to suit the gasses involved (*Figure 25.7*).

25.4.5 Quartz filled 'q'

This technique uses quartz filling around the hot or sparking components to prevent spark emerging into the surrounding atmosphere and to keep external temperatures to a suitable level. This seems to have found some favour in Germany, but has not particularly been used in the UK so far.

Again this technique is not gas group sensitive, but the temperature class needs consideration.

25.4.6 Oil immersion 'o'

With oil immersion, the sparking or hot parts are totally covered in an oil bath. Again this technique is not as popular as other methods of explosion protection, although there are some situations where it is the only practical alternative.

Again this technique is not gas group sensitive, but the temperature class might need consideration. One possible disadvantage is that the oil itself is flammable, and there have been incidents where the oil has ignited under severe fault conditions.

25.4.7 Encapsulation 'm'

This relies on the circuits being covered with a certain depth of encapsulating material. For this technique the encapsulant has to survive various tests to prove its durability, and ability to survive the possible temperatures in the encapsulated circuit.

This technique is not sensitive to gas group but always carries a temperature class.

25.4.8 Non sparking 'n'

This technique is suitable for zone 2 only and uses a combination of suitable enclosure, adequate rating of components, temperature limitation, and non-sparking circuits (such as sealed relays) and if necessary energy limited circuits where any spark is non-incendive, or enclosed break components (which are similar to flameproof, but accepted on the basis of passing a test for non-transmission

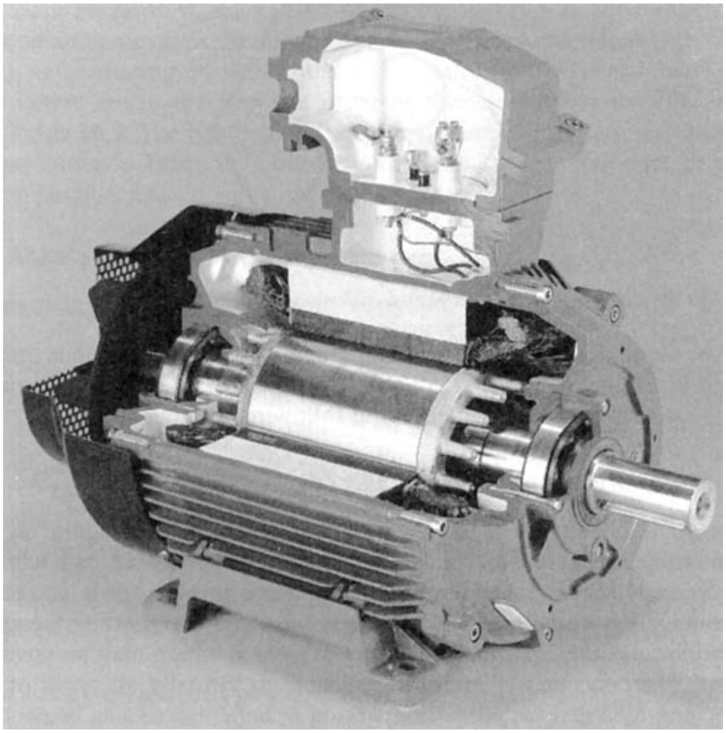


Figure 25.5 Sectional illustration of an EEx 'd' induction motor (courtesy of Invensys Brook Crompton)

of an ignition, but without any specific minimum mechanical requirements). The philosophy is that once the requirements of the standard are complied with, then the chances of a fault occurring in the apparatus at the same time as the explosive atmosphere is present are so small as to be ignored.

The gas group marking may be II or IIA, IIB or IIC depending on whether or not the equipment incorporates energy limited circuits or enclosed break components, and there will also be a temperature class.

25.4.9 Special protection 's'

This standard is unlikely to be used in future. The UK version was a BASEEFA standard which contained various extra tests and guidance to enable certification of apparatus which did not exactly fit the other standards in existence at the time, but could be shown to be of 'equivalent' safety. It was a national standard, with each European test house having its own similar version, but entailed re-certification in each country where the apparatus was used. It was introduced in 1972 in the UK and updated in 1985, but since then some of the principles in it have been incorporated into later versions of the flameproof Standard BSEN 50018, and into EN 50028 for encapsulation 'm'. Also the introduction of BSEN 50284 along with ATEX has come along, so that the standard is now very close to becoming redundant, as most if not all equipment may now be covered for use throughout Europe by means of the newer harmonised standards, or directly against the EHSRs of the ATEX directive.

25.4.10 Mixed concepts

The are some situations where, for various reasons, a piece of equipment is protected by more than one technique. The most likely combinations which will be seen in practice are 'di' or 'id', 'd[i]', 'emi', 'de' and 'n[i]'.

The [i] combinations are often used to enable intrinsically safe barriers or repeaters to be mounted in a 'd' or 'n' enclosure. The interior of that enclosure then counts as a safe area for the barriers and the enclosure can then be located nearer to the instrumentation which the barriers feed than might be possible by relying purely on mounting them in the main safe area.

A common practice is to fit a 'd' motor with an 'e' terminal chamber, to save a little weight. In this situation, the motor and terminal chamber might each carry their own certification, or there might be just one certificate for both, coded 'de'.

The other combinations involving 'i' may occur for a variety of reasons. One example is a gas detector using a pellistor element. The pellistor element uses a catalytic effect to 'burn' the surrounding gas and usually needs to be contained in a 'd' enclosure in order to prevent the flame propagating to the surrounding atmosphere. At the same time the remainder of the circuit often can be certified 'i', so that the equipment is coded 'di' or 'id'. Other examples could be various forms of gas analysers or detectors, or maybe smoke or vapour cloud detectors, where for technical reasons sufficient power is required in the main circuits that they need to be 'd' or 'em', but for ease of maintenance the sensors themselves and maybe their calibration adjustments are 'i' so that they can be exchanged or adjusted at

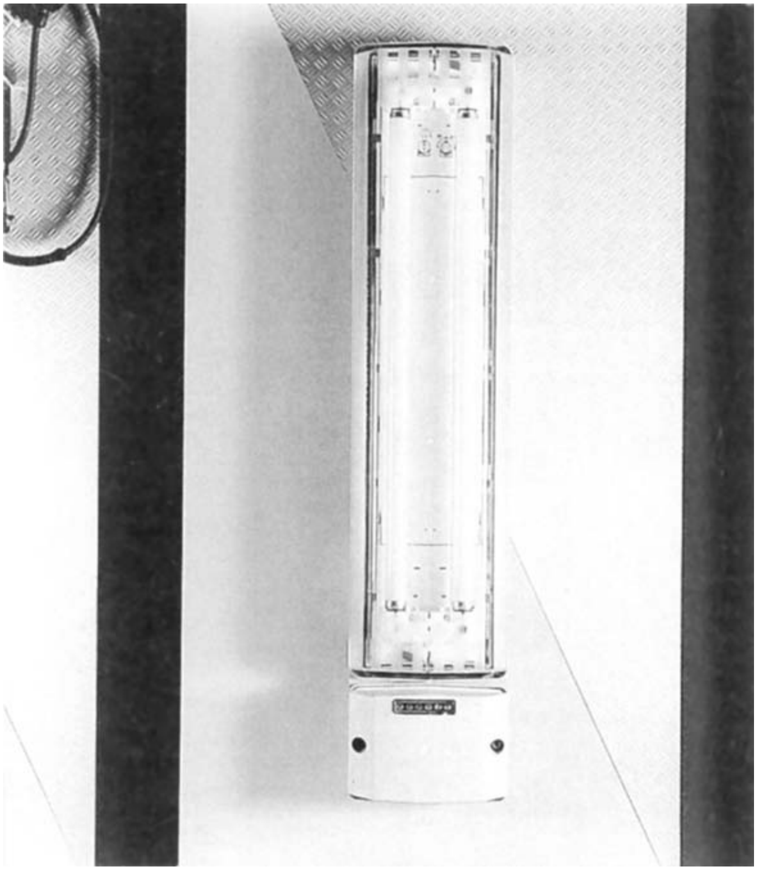


Figure 25.6 EEx 'e' Zone 1 emergency luminaire (courtesy of CEAG Crouse-Hinds)

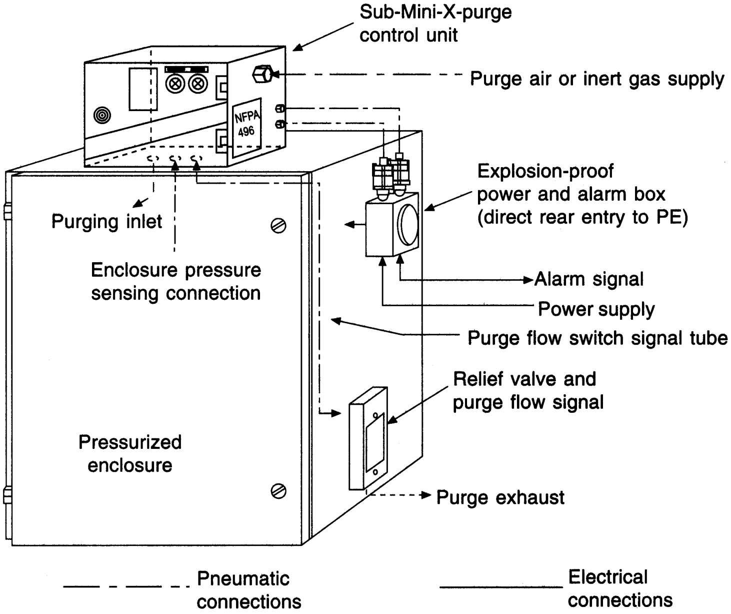

Sub-Mini-X-purge control unit
Purge air or inert gas supply
NFPA 496
Explosion-proof power and alarm box (direct rear entry to PE)
Purging inlet
Enclosure pressure sensing connection
Alarm signal
Power supply
Purge flow switch signal tube
Relief valve and purge flow signal
Pressurized enclosure
Purge exhaust
Pneumatic connections
Electrical connections

Figure 25.7 Main features of a pressurisation installation (courtesy of Expo Safety Systems Ltd)

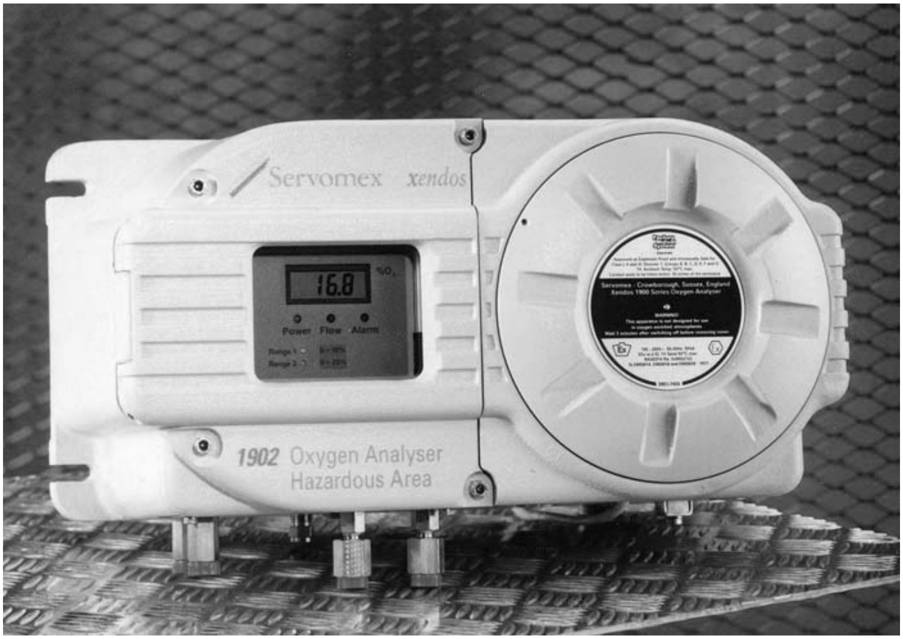


Figure 25.8 Example of a mixed concept apparatus which aids routine maintenance and adjustment. The circular threaded cover allows entry to a flameproof chamber and so must only be opened when de-energised, and under any other specific conditions which might be stated on the unit or in the associated documentation. The chamber behind the display contains only intrinsically safe circuits, thus permitting it to be opened at any time to allow access to, and live adjustment of, calibration or alarm setting controls, but subject of course to any extra operational conditions laid down by the manufacturer. Photograph courtesy of Servomex Group Ltd

any time without having to de-energise the equipment. It is worth noting here that is possible for equipment to be coded ‘dia’ for example, and there can be a slight conflict of understanding here as ‘d’ is a zone 1 concept, whilst ‘ia’ can be zone 0. The general rule to apply here is that in the absence of other information, the lower zone takes preference, and the overall equipment is only suitable for zone 1. There are situations where the ‘ia’ part may be placed in a zone 0, or be attached to pipework containing a zone 0 sample, and this will be explained fully in the certification and/or the manufacturer’s instructions for the particular equipment. (*Figure 25.8*)

Other combinations of codes are not precluded, but the author has not necessarily seen them in practice.

25.4.11 Dust protection

Historically some extra requirements for coal dust protection have been included in the standards mentioned above, to take account of mining situations. This meant that there was a fixed maximum permitted surface temperature of 150°C for any surface or component where coal dust could form a layer, or 450°C on any part where coal dust did not have access. Note that the temperatures shown here are for the very specific case of coal mining where coal dust and methane or firedamp are present. The prevention of access by coal dust was achieved by having an enclosure sealed to at least IP54.

For some time in North America, dust protection requirements have been included in the explosion protection standards for their class II and III, groups E, F and G, but with different (lower) temperatures and different enclosure sealing requirements, depending on the particular dust or fibre concerned.

In Europe, generalised dust protection was covered in various standards, but not specifically included in the explosion protection requirements until the advent of ATEX (see below). Now EN 50281-1-1, EN 50281-1-2, EN 50281-2-1 and others are being used to bring in dust protection here too. Equipment can be certified for use in dust areas only, or with dust and gas. Depending on the severity of conditions anticipated in service, enclosures may be IP5X (dust protected) or IP6X (dust sealed), and the temperature of the apparatus may be based on just a thin coating of dust, or on the possible increased temperature when covered with a thick blanket of dust. The marking used on the apparatus is dealt with under the ATEX heading.

25.5 ATEX certification

The European harmonised certification which came into operation during the 1980s had problems with respect to the introduction of amendments to—or new editions of—the relevant harmonised standards. The original EC directive was precise in specifying the exact issue level of each standard, so that each time a standard was changed, another modifying directive had to be issued to incorporate the new amendment level of the standard. Each country would put the standards or amendments in place ready for use, but they could not be used in the actual certificates until that modifying directive was issued. This caused confusion at times for all concerned, as manufacturers or users would obtain copies of the new standards and expect certificates to incorporate them immediately, whilst the certification bodies would have to refuse to include them until the appropriate EC directive appeared.

One of the main themes behind the ATEX directive is intended to overcome this problem. There is a phrase in the directive along the lines of 'taking account of technological advances', which means that certification bodies will use the latest published version of a standard at the time of the certification.

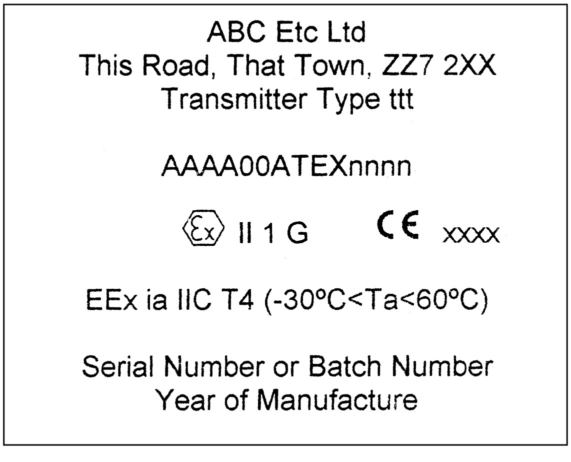
The ATEX directive uses a set of essential health and safety requirements, referred to at various times in other articles as ESRs or EHSRs. These are couched in general terms and cover many, if not all, aspects of hazard to persons, for example integrity of pressure systems and radiation of various types, instead of just possible ignition of explosive atmospheres. The EHSRs also effectively include mechanical as well as electrical sources of ignition, and cover protection against flammable gases or dusts.

At the present time ATEX certification can be used voluntarily as an alternative to the previous certificates of conformity, but from July 2003 any new installation will have to be in compliance with ATEX. Two situations are clear cut at that time: first, any brand new installation will have to meet ATEX; second, existing installations will be able to remain in service for the remainder of their useful lives. The area which is slightly unclear at present is that of repairs and maintenance on those installations. There is bound to be debate in some situations as to how much can be done in the name of a repair and where it becomes new equipment. Also the ATEX directive talks in terms of 'equipment placed on the market' after July 2003. At present there is a feeling that manufacturers or users could maybe 'stockpile' a reasonable number of units for repair purposes, but what is a reasonable number, and would the manufacturer be permitted to hold that stock, or would they have to be sold to the user before the due date and held by him? This is something to which there is no clear answer right now, and it remains to be seen how things develop once ATEX is fully in force.

25.5.1 Example of an ATEX label

AAAA is up to 4 letters identifying the notified body which issued the type examination certificate.

00 is the year of the certificate
ATEX shows it is an ATEX certificate
nnnn is the serial number of the certificate



ABC Etc Ltd
This Road, That Town, ZZ7 2XX
Transmitter Type ttt

AAAA00ATEXnnnn

Ex II 1 G CE xxxx

EEx ia IIC T4 (-30°C<Ta<60°C)

Serial Number or Batch Number
Year of Manufacture

xxxx is the number of the notified body carrying out the quality module of the ATEX requirements. This CE mark together with the four digit number may appear in addition to any other CE mark which the manufacturer may need to apply for other purposes.

The G in the II 1 G signifies protection for gas, and could be replaced by D for dust, or G D for both gas and dust. If the apparatus is protected for dust, there will be two further items on the label: the enclosure degree of protection, which must be at least IP54 for category 3 and IP65 for category 1 or 2, and the maximum surface temperature of the enclosure (shown as Txxx °C).

25.6 Global view

Both the earlier harmonised standards and ATEX arrangements were intended to ensure acceptability of certified apparatus in any European member state without repeating the certification every time.

From the certification body and manufacturer aspect, a strong trend has already developed towards one-stop assessment and testing for several other countries. Various agreements are in place between many organisations in Europe, Canada, America, Australia and Japan among others, whereby one certification body can provide assessment and testing of the design to several standards at the same time and issue a report to the other co-operating body. The equipment still has to be finally certified or approved by the appropriate organisation in the country where it is to be used, but the initial 'combined' assessment can help in achieving a 'universal' design first time and avoid having to keep redesigning the equipment for each authority in turn, with the resultant delays and inconvenience this was likely to cause.

Historically the various standards relating to intrinsic safety all used basically the same set of ignition curves originated by SMRE in the early 70s and supplemented by more research work by PTB in Germany in the 90s. Similarly, the test gas mixtures used in both intrinsic safety and flameproof testing tend to be the same throughout all the standards. However the exact application of factors of safety applied in testing or interpretation of the curves varies from standard to standard, which is one reason why it has been necessary to have equipment certified separately in every country where it is used.

Even when the equipment itself is certified to the harmonised standards within Europe, each member state tends to have its own installation requirements which need to be taken into account. The advent of EN 60079-14 (Code of practice for installation) lays the foundation for hopefully a more universal set of installation practices.

In the long term the IECEx scheme of certification (based on the IEC 60079 series of standards) is heading in the direction of providing certificates which would be acceptable in any participating country, so eventually full global certification is a possibility. At present the scheme is still only being used on the basis of an assessment and test report (ATR) produced in one country being accepted as the basis for certification in another country. At this time there are still too many national differences compared to the IEC Standards to enable direct acceptability of any country's certificate in any other country, but the long term aim is to reach that situation. around 20 countries have so far signed up to the IECEx scheme with around a further 5 showing interest.

It is also worth mentioning again here that the style of coding similar to EEx ia IIC T4 is becoming more and more the usual arrangement to look for, which should avoid the confusions which tended to occur in the past. The variations in this coding are likely to be only such things as the EEx changing to Ex for the IECEx scheme, or AEx for American approval, but the protection concept, gas group and temperature class will be recognisable throughout.

25.7 Useful websites

In this technological age, there is a plethora of information available on the internet. The following is a list of useful websites found by the author, with apologies to those not included.

http://www.baseefa.com
http://www.hse.gov.uk
http://www.dti.gov.uk
http://www.iee.org.uk
http://europa.eu.int
http://europa.eu.int/comm/enterprise/newapproach/
http://www.newapproach.org
http://www.hazloc.com
http://www.iecex.com
http://www.iec.ch
http://gost.ru
http://www.fmglobal.com
http://www.csa-international.org

Section F

Power Generation

26

Prime Movers

M A Laughton BASc, PhD, DSc(Eng), FIEE, FEng
Formerly of Queen Mary and Westfield College,
University of London

I G Crow BEng, PhD, CEng, FIMechE, FIMarE,
MemASME
Davy McKee (Stockton) Ltd
(Section 26.1)

H Watson BSc, CEng, FIMechE, FIMarE,
MemASME, FFB
Formerly Engineering Consultant
(Section 26.2)

W Rizk CBE, MA, PhD, FEng, FIMechE
W.R. Associates
(Section 26.3)

E Goldwag BSc(Eng), CEng, FIMechE
Engineering Consultant
(Section 26.4)

L L J Mahon CEng, FIEE, FIQA, FIBM
Consulting Engineer
(Section 26.5)

Contents

26.1 Steam generating plant 26/3
  26.1.1 Introduction 26/3
  26.1.2 Combustion 26/3
  26.1.3 Sources of chemical energy 26/3
  26.1.4 Thermodynamics and hydrodynamics of steam generating plant 26/5

26.2 Steam turbine plant 26/6
  26.2.1 Cycles and types 26/6
  26.2.2 Turbine technology 26/9
  26.2.3 Turbine construction 26/10
  26.2.4 Turbine support plant 26/12
  26.2.5 Turbine operation and control 26/14

26.3 Gas turbine plant 26/16
  26.3.1 Open cycle plant 26/16
  26.3.2 Closed-cycle plant 26/18
  26.3.3 Combined-cycle plant 26/18
  26.3.4 Cogeneration/CHP plant 26/19

26.4 Hydroelectric plant 26/20
  26.4.1 General 26/20
  26.4.2 Types of plant 26/21
  26.4.3 Power station 26/23
  26.4.4 Turbines 26/23
  26.4.5 Hydrogenerators 26/27
  26.4.6 Economics 26/27
  26.4.7 Pumped storage 26/28

26.5 Diesel-engine plant 26/29
  26.5.1 Theory and general principles 26/30
  26.5.2 Engine features 26/32
  26.5.3 Engine primary systems 26/32
  26.5.4 Engine ancillaries 26/34
  26.5.5 A.c. generators 26/35
  26.5.6 Switchgear and controls 26/36
  26.5.7 Operational aspects 26/37
  26.5.8 Plant layout 26/39
  26.5.9 Economic factors 26/40
  26.5.10 Cogeneration/CHP 26/41

26.1 Steam generating plant

26.1.1 Introduction

The production of steam at conditions suitable for supplying an engine or a turbine, or of providing steam for heating or process plant, is achieved by the transfer of heat from a primary energy source to water contained in a boiler, or steam generator, which in its most rudimentary form may be little more than a vessel heated from below. Currently, there are two major energy sources capable of supplying heat at a sufficiently high temperature and in a controlled and economically acceptable fashion. These are nuclear fission and the combustion of fossil fuels, such as coal, oil, natural gas and their derivatives. It is the latter, or 'conventional' energy source and its associated plant which are considered here.

26.1.2 Combustion

The combustion process for a fossil fuel is a rapid exothermic chemical reaction between the oxygen in the air and the combustible elements in the fuel. There are two such principal elements, carbon and hydrogen, which react with oxygen, thus:

$$C + O_2 = CO_2 + \text{heat}$$

$$2H_2 + O_2 = 2H_2O + \text{heat}$$

The heat release for carbon is approximately 7830 kcal/kg of carbon burned, and the equivalent figure for hydrogen is 33 940 kcal/kg. The requirements for the above reactions to proceed to completion are: first, a sufficiently high temperature to ignite the constituents; second, adequate mixing of the constituents; and third, sufficient time for the reaction to be completed. With regard to the ignition temperature, *Table 26.1* gives an indication of these values for a number of fuels.

The mixing of the constituents and the reaction times are both major elements in the design of the furnace of a steam generator, which must be large enough and provide sufficient turbulence to ensure that these two requirements are met.

The absorption of heat in the boiler is achieved in two ways. In the furnace, where all the combustion takes place, heat is transferred to the water by *conduction* through the walls of its containing vessel or tubes, which receive heat by radiation from the burning fuel. At the outlet of the furnace the products of combustion are still very hot (probaby in excess of 1000°C), so that they remain capable of transferring heat; this time by *convection*. This heat can be used in non-superheating boilers to further heat the water, while in superheating boilers, as the gas passes through the convection zone, it may progressively be used to superheat the steam first produced in the furnace, and eventually the cold incoming feedwater. Finally, the gaseous products of combustion are ducted to a chimney (sometimes via flue gas clean-up devices) and so any residual heat left in these gases is lost. Nonetheless, the efficiency of steam generating plant, as defined by the ratio between the heat energy of the steam leaving the boiler and the chemical energy of the fuel entering the furnace, can be high, with typical figures in excess of 85% for coal fired plant, rising to as high as 98% for compact oil fired boilers.

Table 26.1 Ignition temperatures for some common fuels in atmospheric air

Fuel	*Ignition temperature* (°C)
Charcoal	343
Bituminous coal	407
Anthracite	450–600
Ethane (C_2H_6)	470–630
Ethylene (C_2H_4)	480–550
Hydrogen (H_2)	575–590
Methane (CH_4)	630–750
Carbon monoxide (CO)	610–657

With regard to efficiency definitions, care should always be taken to establish the basis of the efficiency calculation, as the heat content of the fuel is generally assessed differently in Europe and the USA. In the USA the higher heat of combustion (Q_H) is used which incorporates the heat of vaporisation of water (that is, it is assumed that all water vapour formed by combustion is condensed). The lower heat of combustion (or lower calorific value, Q_L) is derived by assuming that the products of combustion remain in the gaseous state. It may be shown that

$$Q_L = Q_H - 578.24W \text{ kcal/kg}$$

where W is the weight of water formed per kilogram of fuel. European practice tends to favour the use of Q_L.

Calorific values are obtained experimentally by means of a bomb calorimeter in which combustion occurs at constant volume and the derived value Q_H. Q_L is obtained from the above relationship and, in fact, refers to combustion occurring at constant pressure.

The major heat loss from a boiler has already been identified as the heat loss up the stack, or chimney. Other losses may arise due to incomplete combustion of the fuel, leaving carbon in the ash or by producing CO rather than CO_2. In practical installations it is always necessary to use more than the theoretical air requirements to be sure of complete combustion. However, in order to keep the stack loss to a minimum this excess must be carefully controlled. For fuel oil fired installations, excess air is typically in the range 5–10% by weight, referred to the theoretical air requirements. Corresponding figures for a pulverised coal fired unit would be 15–20%, while for a stoker fired boiler the figures would be in the range 30–60%. The remaining inherent loss for a boiler arises due to the moisture content of the fuel and the production of water vapour in the combustion process, while the remaining avoidable loss arises from radiation from the boiler setting. This latter loss can, of course, be controlled by providing the unit with good insulation.

26.1.3 Sources of chemical energy

The fossil fuels available as suitable heat sources occur naturally in the earth and are the remains of organic materials once found living on the Earth's surface. It is not surprising, therefore, that the properties of individual fuels vary markedly from place to place and that the fuels contain many trace elements which can in some cases profoundly influence their combustion characteristics. It is necessary, therefore, that the boiler plant designer knows the type of fuel that will be burned and also fully understands the effects the properties of the fuel will have upon all his design parameters. In consequence, much effort has been directed towards understanding the combustion of conventional

Table 26.2 Examples of some fuel oils available

Grade of fuel	*Composition (% by weight)*					*Gross heating value* (kcal/kg)	*Description*
	C	H_2	$O_2 + N_2$	*S*	*Ash*		
Distillate (amber)	86.5	12.7	0.2	0.7	Trace	10 800	For general-purpose domestic heating
Very light residual (black)	86	12	0.5	1.5	0.02	10 500	For installations not equipped with preheating
Residual (black)	85.75	10.5	0.9	3	0.08	10 150	For installations equipped with preheating
Crude oil (Mexican)	83.3	10.9	2.2	3.6	—	9 665	
Coal-tar	89.5	6.5	3.5	0.6	—	9 020	

fuels which are commonly used, and considerable data are available on the fossil fuels (oils, natural gases and coals).

As well as fossil fuels, less important, but common, vegetable fuels such as peat, wood, wood bark, bagasse (sugar cane residue), grain hulls and residues from coffee grounds and tobacco stems are useful sources of heat. Sometimes gases, tars and chars produced as by-products in steel making or oil refining processes are used. In many instances when waste products are to be used they will be burned in conjunction with the major fuels (oil, coal and gas).

26.1.3.1 Liquid fossil fuels

Fuel oil, owing to its relative ease of handling and storage, the absence of large quantities of residual ash and its high calorific values, had by 1970 become a most attractive fossil fuel. Since then, however, its relative cost with regard to other fuels and the uncertainty of its long-term availability have both tended to reduce its competitiveness. Nonetheless, particularly in the case of smaller boiler installations, oil remains a viable fuel.

The type of oil burned for steam generation can vary from light distillates having low viscosity and low specific gravity to the heaviest residual black fuel oils with very high viscosities. Occasionally, crude (or unrefined) oil may be considered for use directly in the boiler furnace; however, the presence of highly volatile fractions (normally removed in a refining process) may prove troublesome. Typical characteristic of fuel oils are shown in *Table 26.2*. Despite the low percentages of ash indicated in *Table 26.2*, trace elements such as vanadium, sodium and sulphur can be responsible for a number of operating problems.

A further source of liquid fuel is oil derived from oil-shale, which is a fine-grained, compact, sedimentary rock containing an organic material known as kerogen. Shale oil is obtained by heating the rock to 470 °C. (The yield is normally about 115.1 per tonne of rock.) Substantial reserves are known to exist in North America, in some parts of the Middle East, in China and in Australia. The use of shale oil in steam generators may be restricted by cost, but consideration is currently being given to the direct combustion of oil-shale in boilers.

26.1.3.2 Gaseous fuels

In many ways gas can be considered the most easily used of all the chemical fuels. It is capable of easy transportation from the producing or gathering plant to the consumer, and the use of pipelines can eliminate storage problems, particularly for very small installations. Care is required, however, in large boilers to avoid the occurrence of explosions during unit ignition or at other times when ignition may be lost. As may be expected, for clean gases complete combustion with a low excess air requirement is possible and the combustion does not produce smoke; additionally, it is substantially free of ash. Typical analyses of gases burned in steam generators are shown in *Table 26.3*. Blast-furnace gas is included in *Table 26.3* for completeness, although it is usually available only as a by-product of a steelworks. It is normally heavily contaminated with dust, and great care must be taken to avoid plugging of fuel pipes or furnace fouling. Clean-up is usually achieved by a washing process.

26.1.3.3 Solid fossil fuels

Coal is a major fossil fuel throughout the world and has been in use for many decades by the industrialised nations. It is found in many forms, ranging from anthracitic through bituminous and sub-bituminous to lignitic. With so many sources and with such variations of type and quality, it has

Table 26.3 Examples of some gaseous fuels available

Gaseous fuel	*Composition (%)*							*Higher heating value* (kcal/kg)
	CO	H_2	CH_4	C_nH_m	CO_2	N_2	O_2	
Coal gas	4.7	2.5	38.5	35.9	10.3	7.3	0.8	8640
Water gas	71	6.4	0.3	—	14	8.3	—	3560
Producer gas	31.3	0.9	0.2	—	8.5	59.1	—	985
Natural gas	—	—	63–88	6–29	0.5–8	0.5–7.3	—	9000–11 050
Blast-furnace gas	30	0.2	1.3	—	14	55	—	5900

Table 26.4 Examples of some hard coals available

Country	*Typical analysis* (%)				*Higher heating value* (kcal/kg)
	Moisture	*Volatile matter*	*Fixed carbon*	*Ash*	
Australia	0.5–2.6	30–37	50–65	11–15	6500–7300
Canada	1.4–4.0	25–34	51–62	10–12.5	7000
China	2.5–5.4	25–30	40–45	20–31	5200–6000
Great Britain	0.7–7.5	11–36	57–85	2–4	7400–8000
Poland	4–17	21–32	40–59	6–22	4500–7500
South Africa (export)	2.0–5.5	16–27	50–70	9–16	5900–6800
USA (export)	1.5–3.5	20–40	—	5–9	7700

Table 26.5 Examples of some non-fossil solid fuels

Fuel	*Moisture* (%)	*Volatile matter* (%)	*Ash* (%)	*Heating value* (kcal/kg)
Peat (air dry)	37–45	38–40	1–2	2600–3000
Bagasse	40	45	1.315	2200–4860
Wood (air dry)	10–20	75	0.5–3	4200–5400
Wood waste with bark	15	70	2	3300

been found impossible to produce any generalised analysis for this fuel which will predict its behaviour in the boiler. As opposed to oil and gas, coal contains relatively high ash levels, which in turn will contain trace compounds (such as SiO_2, AlO_2, TiO_2, Fe_3O_4, CaO, MgO, Na_2O, K_2O, Mn_3O_4 and P_2O_5) whose presence and effects need to be recognised if efficient and reliable coal combustion is to be achieved.

With these reservations in mind, however, it is possible to divide coals into certain classifications. In Europe the International System of Classification is generally used, while in the USA the ASTM System is used. Both systems have attempted to group coal together by their heating values, moisture contents and volatile matter contents.

Anthracitic coals have a high heating value, are non-agglomerating, have low volatility and very high fixed carbon content (in excess of 85%). Bituminous coals have a heating value not less than 5700 kcal/kg, and are agglomerating, while sub-bituminous coals can have heating values as low as 4600 kcal/kg. Lignitic coals are characterised by high moisture contents (up to 60%) and heating values lower than 4600 kcal/kg.

To provide an indication of the variations of coals worldwide, some of the more significant producers are listed in *Table 26.4*, together with corresponding typical hard coal properties.

26.1.3.4 Other solid fuels

Other fuels, mainly vegetable or vegetable waste, have been found to be useful. Peat is, in fact, burned by certain public utilities. Wood and wood wastes as well as bagasse waste are also well-known fuels, while, more recently, trials have been conducted into the use of industrial and domestic refuse as a boiler fuel. Some indication of the properties of these fuels is given in *Table 26.5*.

26.1.4 Thermodynamics and hydrodynamics of steam generating plant

In a conventional steam generator the working fluid (water) receives heat from the chemical reactions occurring between its fuel and the air, and in an ideal situation the working fluid is at constant pressure. In a real boiler there is a pressure drop between inlet and outlet due to the effects of friction; however, the process can reasonably be represented by the line bcde in *Figures 26.1* and *26.2*. Initially, the incoming water or feedwater to the boiler has to be pumped into the boiler and this is represented by the line ab. At inlet to the boiler (b) the water enthalpy will be subcooled and will first need to be heated to saturation at c. At c its mass quality (x) is zero. Mass quality may be defined as,

$$x = W_g/(W_g + W_f) \qquad (26.1)\Leftarrow$$

where W_g and W_f are the flow rates of steam and water, respectively. A mass quality definition can also be stated as

$$x = (h - h_f)/h_{fg} \qquad (26.2)\Leftarrow$$

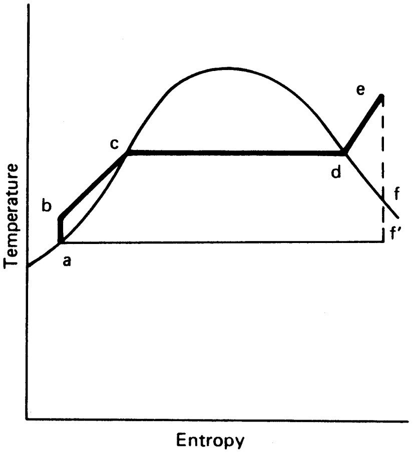


Figure 26.1 The temperature–entropy diagram

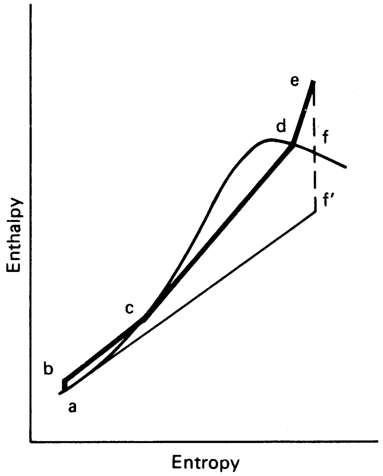


Figure 26.2 The Mollier or enthalpy–entropy diagram

where h is the enthalpy of the working fluid, h_f the enthalpy of saturated water at the same pressure and temperature, and h_{fg} the latent heat at the same conditions.

In equation (26.1) x can vary between zero and unity, while in equation (26.2) x can assume negative values if the fluid is subcooled and values greater than unity when the fluid is superheated. In the range $0 \leq x \leq 1$, the values of x derived from equations (26.1) and (26.2) are identical if thermodynamic equilibrium exists.

Between c and d the mass quality increases from zero to unity as steam is produced, until at d the working fluid is completely transformed into dry-saturated steam. If the boiler is equipped with superheaters, further heat can be added, to achieve a final condition at e. The processes between c and d occur at an essentially constant temperature (the saturation temperature equivalent to the pressure in the boiler), while in the superheat region the added heat causes the steam temperature to rise.

At point d or e the steam can be used to provide heat for process plant or be used to drive a work producing engine or turbine. If steam leaves the boiler at d, which is a dry-saturated condition, it is evident that the removal of energy from it by heat transfer or by work in an engine will cause it to start to condense, and x will fall. However, if the steam is initially superheated to e, it can provide an amount of work approximately equivalent to the drop in enthalpy ($h_e - h_f$) before any moisture is formed: this is the great advantage of superheat. Nonetheless, many applications, particularly in the process industries, require only heat and the added complications of providing large amounts of superheat are avoided, providing that process temperature demands are compatible with available boiler pressures.

By contrast, in cases of power generation, high pressure and temperatures are thermodynamically attractive, as in the expansion of the steam to vacuum in a condenser (point f′ in *Figures 26.1* and *26.2*), so making the term ($h_e - h_f$) as large as possible. The boiler may then be said to be working as part of a regenerative cycle with steam under vacuum at f′ being condensed along f′*a* to a pump at a, where it is pressurised and returned to the boiler at b.

Mixtures of these two systems are common, particularly in industry, where some electrical power as well as process steam is required. However, in all cases the steam generator must provide the steam at the desired conditions, and to achieve this the boiler designer must be fully aware of the processes involved within the working fluid as it moves from b to e.

For the simplest type of industrial boiler, the shell or fire tube boiler, where only modest steam pressures are available and little or no superheat is produced, the combustion occurs in furnace and fire tubes and the products of combustion pass down further tubes (called smoke tubes). The tubes are contained within a vessel (or shell) which contains essentially non-flowing water, heated by the tubes passing through it. The rate at which heat can be transferred to the water is controlled essentially by the properties of the water. If the rate of heat transfer is poor or a very high flux is provided by the furnace, overheating of the tubes may occur and the material may fail. Information on these heat transfer rates, either by experience or by experiment, are therefore required by the designer.

For higher pressures and capacities and for applications calling for appreciable amounts of superheat, the roles are reversed, in that water occupies the tubes, which now form the enclosure of the furnace. Such a boiler is known as a water tube boiler. In such designs, water flows in the tubes and it is induced to do so either by means of a pump or by the 'natural' effects induced by heating water in inclined or vertical tubes so that the steam so formed is free to rise.

26.2 Steam turbine plant

The steam turbine is a prime-mover well suited to the direct drive of two- or four-pole high-speed a.c. generators for power-supply networks. The associated steam-raising plant can be fired by a wide variety of fuels—fossil, nuclear and such unusual fuels as city refuse, sawdust and sugar waste. The turbine can accept a low exhaust temperature (making possible a reasonable thermal efficiency) and a moderate inlet temperature (permitting a long life, e.g. 200 000 h). The turbine is best suited to high power ratings, as parasitic losses tend to be independent of size; further, output ratings can be raised without a pro-rata increase in materials, so giving a higher power cost ratio.

Steam turbines are constructed in the range 5–1300 MW. At the highest ratings the only comparable prime-mover is the hydraulic turbine, while at the lower end of the range the steam turbine competes with the diesel engine and the gas turbine.

Besides generator drive, the steam turbine is applied to ship propulsion and to rotary compressors. This section deals exclusively with electric power generator drive applications.

26.2.1 Cycles and types

26.2.1.1 Reheat cycle

Most fossil-fuel-fired generating stations use the reheat cycle. Currently, outputs span the range 200–1300 MW. Common steam conditions are 16 MPa, 540 °C (2350 lbf/in.^{2}, 1000 °F). Occasionally supercritical conditions, e.g. 22 MPa and above, are met, and sometimes there are two stages of reheat; but the one-stage reheat subcritical cycle here described is more common. The British advanced gas-cooled nuclear-reactor plant also uses this cycle, with outputs currently of 660 MW.

In *Figure 26.3*, ABCD shows the conditions of steam throughout the cycle, and *Figure 26.4*(*a*) gives the plant arrangement. Steam is expanded in the high-pressure (h.p.) stage down to near-saturation and is then returned to the

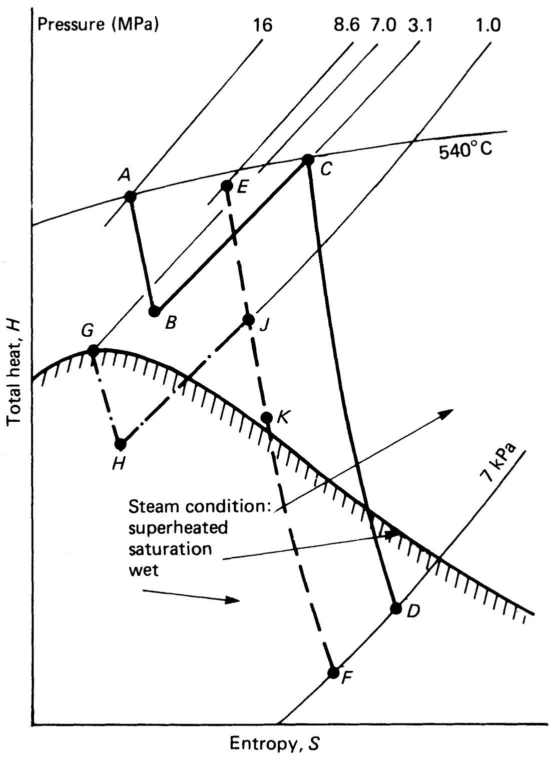


Figure 26.3 Cycles on a Mollier chart

boiler for reheat to the original temperature but at lower pressure. It then expands in the low-pressure (l.p.) stage until it becomes as wet (usually not more than 13% wet) as the final blades can tolerate.

The cycle is efficient and eases the design of last blading, which is usually the critical problem in turbine development. By using reheat, less steam per megawatt-hour is needed and this means shorter blades at exhaust. The cycle usually gives an exhaust somewhat drier than others, reducing the blade erosion hazard. However, all this is at the expense of some complication and extra boiler size and the smaller units—say under 100 or 200 MW—use the simpler non-reheat cycle.

26.2.1.2 Non-reheat cycle

This is shown by the line EF, in *Figure 26.3*, and the plant layout in *Figure 26.4*(*b*). Steam usually at 5–8.5 MPa (900–1250 lbf/in^2) is expanded through the turbine to exhaust from the condenser without return to the boiler. This simple and compact machine is usually constructed in a single cylinder. The plant has a lower first cost than that for reheat, but is 8% or more higher in fuel consumption.

26.2.1.3 Pressurised-water and boiling water reactor cycle

The steam conditions are given by line GHJF and the layout in *Figure 26.4*(*c*). Water-cooled reactors require a special cycle because steam is supplied at an unusually low temperature, about 280 °C (540 °F). The steam is saturated and when expanded, quickly becomes wet. Water in the h.p. stages can cause damage, so the steam is taken out of the turbine after the h.p. stage and put through a drier—usually a device that uses the motion of the steam to separate water from it. It is then reheated by steam bled from the main supply to a surface heat exchanger alongside the turbine.

The cycle is not efficient and turbine design must take care of the potentially damaging effects of water in the steam. However, it suits the pressurised-water boiling-water reactors and their economics. Machines of up to 1300 MW are in service using such cycles.

26.2.1.4 Extraction cycle

This cycle, line EKF in *Figure 26.3* with layout in *Figure 26.4*(*d*), is particularly economical where steam is used to supply a process as well as to generate: a typical example is the desalination plant fed with steam from a turbogenerator, an arrangement in demand in countries where industry and population develop in desert areas. Here, power units of 20–120 MW are in operation, and pass-out steam feeds multi-flash sea-water desalination plants.

Steam is expanded in the h.p. stage of a turbine and is then withdrawn. A control system acts to take off as much steam as the process requires. The steam that the process does not need is then returned to the turbine to expand and do work in the l.p. stages and ultimately to exhaust to the condenser under vacuum.

26.2.1.5 Back-pressure ('total energy') cycle

Figure 26.3 shows the cycle by line EK, and *Figure 26.4*(*e*) shows the plant diagram. While the turbine is efficient in itself, with stage efficiencies of around 80%, the basic cycle has a much lower efficiency (usually 30–38% overall for power stations), mainly because over half the energy in the steam is lost in the exhaust. With the exhaust pressure usually as low as 5 kPa (0.75 lbf/in^2) the steam is at only 33 °C (92 °F). It is difficult to find ways to use energy at such a relatively low temperature and so the very considerable heat in the exhaust is discarded to a condenser.

Some factories need steam for process work, e.g. at upwards of 150 °C and at a pressure useful enough to drive the steam to and through the process plant—say 350–1000 kPa. District heating schemes usually require water at 65 °C and obtain it from steam at around these conditions.

The turbine can be arranged to exhaust at these higher pressures and to pass all its exhaust steam to the process or to the district-heating scheme. Although the higher back pressure of, say, 700 kPa makes the machine much less efficient than with a more normal condensing turbine, the cycle as a whole is efficient because the turbine exhaust heat is not lost. Thus, a back-pressure plant generating electricity and exhausting to a factory process can have a thermal efficiency of 70% compared with 35% for a normal power station.

The basic problem with back-pressure units is that the heat required near to the turbine-generator rarely keeps pace with the electrical power required locally. Basically, it is easier to distribute electricity to more distant parts than to send heat: back-pressure units are therefore usually small. Confining their power and heat output to nearby use, they are rarely of more than 20 MW rating and often much less. There may well be a future for such cycles in the conservation of increasingly costly fuel, and the potential of such 'total energy' schemes must inevitably attract more attention.

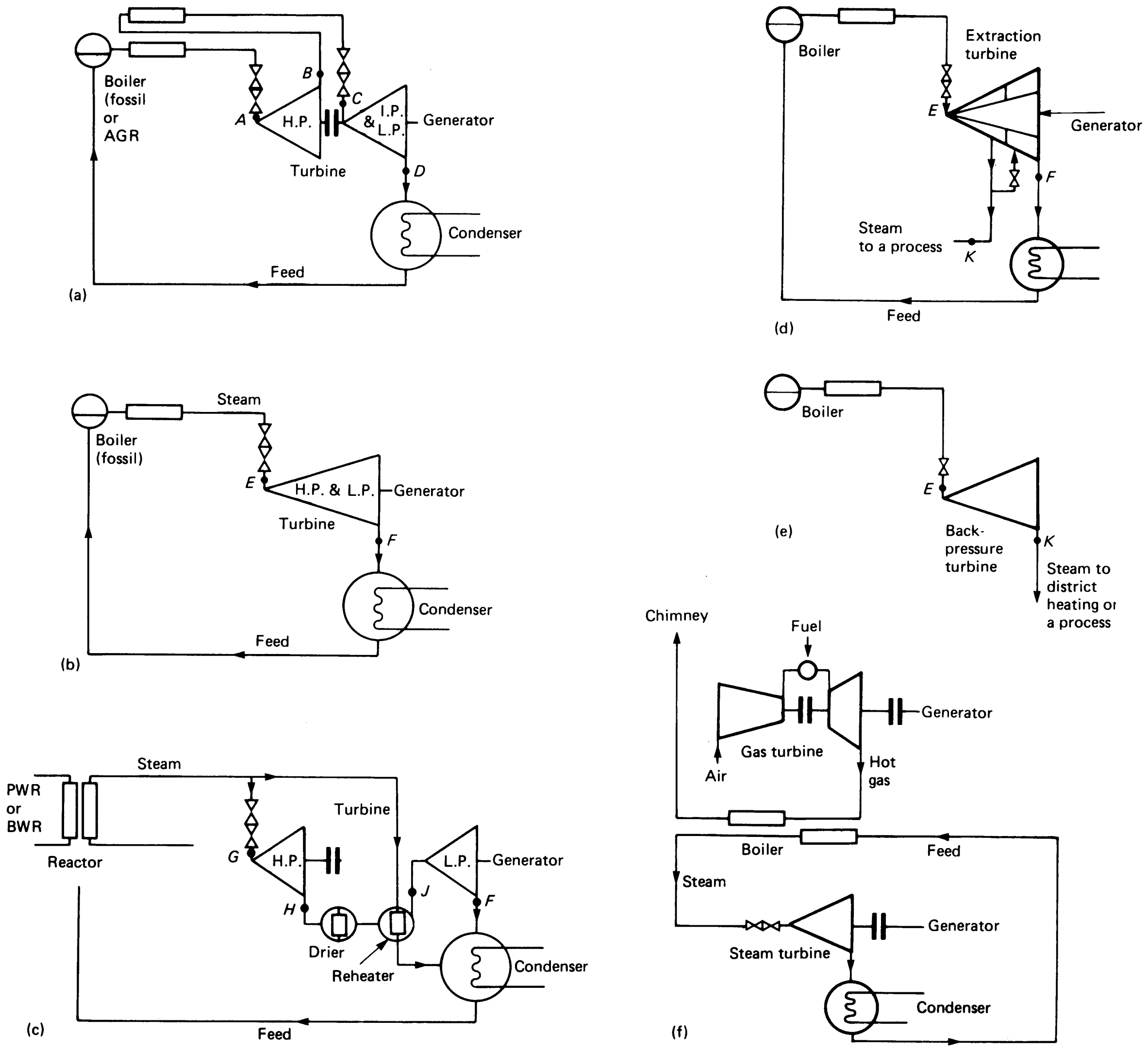

Boiler (fossil or AGR
B
A
C
H.P.
I.P. & L.P.
Generator
Turbine
D
Condenser
Feed
(a)
Boiler
Extraction turbine
E
Generator
F
Steam to a process
K
Feed
(d)
Steam
Boiler (fossil)
E
H.P. & L.P.
Generator
Turbine
F
Condenser
Feed
(b)
Boiler
E
(e)
Back-pressure turbine
K
Steam to district heating or a process
Chimney
Fuel
Generator
Gas turbine
Air
Hot gas
Boiler
Feed
Steam
Generator
Steam turbine
Condenser
(f)
Steam
PWR or BWR
Reactor
Turbine
G
H.P.
L.P.
Generator
J
F
H
Drier
Reheater
Condenser
(c)
Feed

Figure 26.4 Various turbine cycles. (a) Reheat cycle, commonly used in fossil fuel and AGR power stations; (b) Non-reheat cycle, used for lower power; (c) Cycle for pressurised-water and boiling-water nuclear reactors; (d) Extraction cycle; (e) Back-pressure or 'total energy' cycle; (f) Combined cycle, one of various steam and gas turbine arrangements

26.2.1.6 Combined cycle

The ideal engine has a high inlet thermodynamic (absolute) temperature T_1 and a low exhaust temperature T_2. The highest attainable Carnot efficiency is $(T_1 - T_2)/T_1$. The steam turbine has only a moderate T_1 but a very low T_2, so that its efficiency is normally higher than that of the gas turbine, for which T_1 is high but so is T_2.

The combined cycle aims to optimise conditions with a cycle that uses the high T_1 of the gas turbine and the cool exhaust of the steam turbine, *Figure 26.4(f)*. The cycle works as follows.

Fuel is burnt in the combustion chamber of a gas turbine, which generates electricity and exhausts to a heat exchanger. Here the hot gases boil water to raise steam which then drives a steam turbine which also generates electricity. The steam turbine exhaust is changed to boiler feed water in a condenser.

A number of such plants are in service and more are being built because of their high efficiency. They also have the quick-starting ability that the straight steam turbine lacks. Also they can offer power early in the building programme by installing the gas turbine first and the steam turbine later.

They have as yet a reliability somewhat lower than that of simple steam plant, and the gas turbine has a shorter life. The availability is more acceptable if natural gas is used rather than oil. A common arrangement is to take about two-thirds of the power output from the gas turbines and one-third from the steam turbine; say, two gas turbines and one steam turbine per unit.

While this cycle is as yet limited to gas and carefully prepared oil, there are prospects of its use with coal. The pollution problem is creating so strong an emphasis on clean burning that new combustion methods are being developed for coal. Pressurised, fluidised bed combustion is one; gasification is another. Such arrangements promise to produce a gas clean enough for gas turbines to use. So there is a prospect of using the combined cycle to give thermal efficiencies of around 45% using coal, which is the most profuse fossil fuel.

Combined cycles are not necessarily arranged as in *Figure 26.4(f)*. For instance, fuel can also be burnt in the gas-turbine exhaust to raise more steam (see also Section 26.3.3).

26.2.1.7 Alternative fluids

From time to time research is done into fluids alternative to steam. The physical properties of freon and ammonia have certain advantages, but so far the disadvantages have over-ruled them. Energy costs, however, may yet impose a change from the steam cycle.

26.2.2 Turbine technology

26.2.2.1 Blade types

There are two basic types of blade—impulse and reaction. In the impulse type (*Figure 26.5(a)*) all the expansion of steam is done in fixed nozzles and the high-velocity jets so created drive the rotor blade. In most machines the expansion is in successive stages, of which there are usually as many as 20 from inlet to exhaust, each comprising a row of nozzles and a row of moving blades.

Some turbines use a 'Curtiss' stage (*Figure 26.5(b)*) to drop the temperature quickly at inlet or to shorten the machine. This stage takes a more than usual pressure drop. An exceptionally high jet velocity results and this is converted to work by driving two or more rows of blades from the one jet. The sketch shows how this is done using guide blades to re-direct the steam leaving the first row.

In reaction blading (*Figure 26.5(c)*) the moving blades are driven not only by the steam jets from preceding nozzles: they act as nozzles themselves and the jet they create helps to propel them. Therefore, pressure is lost both in the fixed and moving blades and power is produced both by impulse and jet reaction.

In practice, few machines are purely impulse and none are purely reaction. In the modern so-called impulse machine all blades have a degree of reaction and generally this increases towards the exhaust. In some designs the amount of reaction is more marked, depending on the background experience of the manufacturer. Properly designed, there is little difference in efficiency between the impulse and reaction types.

In all but a few small machines, the steam flows along the shaft, parallel to the axis. There are two unusual designs in which this is not so. One is the Ljungström turbine, in which steam flows radially outwards through concentric rings of counter-rotating blades. The other is the Terry turbine where the steam jets travel at an angle inwards to drive against hollows in a rotating disc. The Ljungström turbine is confined to medium outputs, while the Terry turbine is used only for very low powers, e.g. for driving small auxiliaries.

Both these types are now rare. Most modern machines have axial flow with either reaction blading or a combination using h.p. stages of mainly impulse and l.p. stages of mainly reaction design.

26.2.2.2 Size effects

The steam turbine is notable in that higher outputs can be achieved without a pro-rata increase in mass and cost. A power increase of 20% is achieved by using blades 20% longer and this increases the mass of the machine by much less than 20%. So the mass/output ratio is a curve such as that shown in *Figure 26.6*. Also, parasitic losses such as gland loss, etc., do not increase at the same rate as power does when greater outputs are used. Thus, increased size produces some gain in efficiency (*Figure 26.7*).

Taking these two factors together there is a financial gain to be got from increased size (*Figure 26.8*). Two important points to be noted, however, are:

(1) If size is achieved at the expense of reliability, the financial gain may well become a loss; and

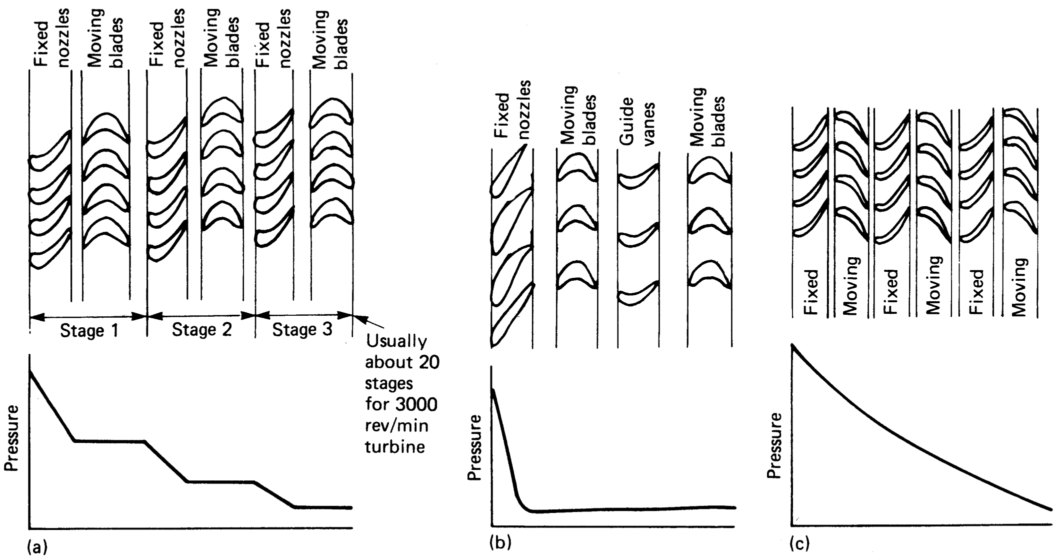


Figure 26.5 Blade types: (a) impulse, pressure compounded; (b) Curtiss stage; (c) reaction

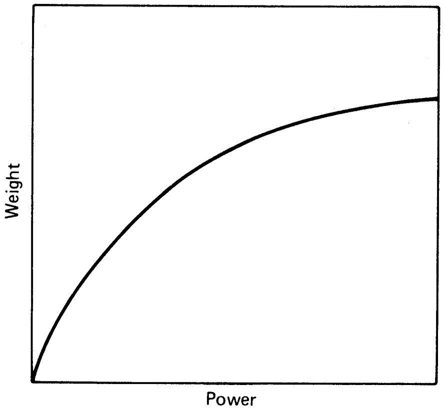


Figure 26.6 Weight of turbines of higher output

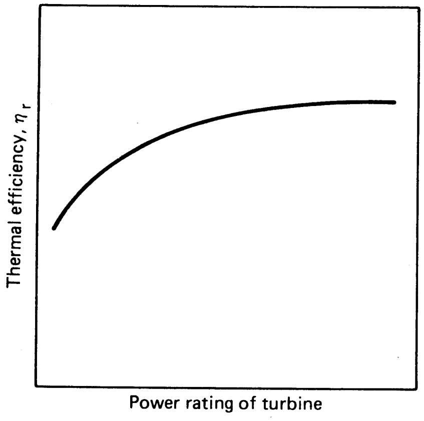


Figure 26.7 Trend of efficiency as outputs grow

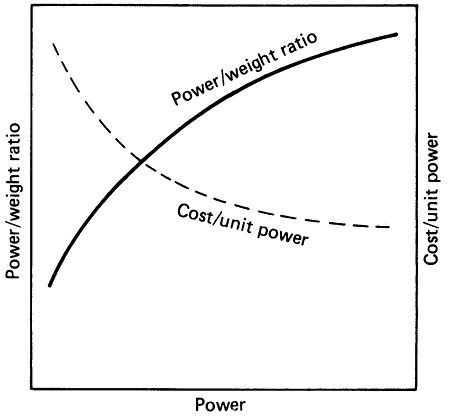


Figure 26.8 Power/weight and cost/unit power trends

(2) The gain due to higher outputs is best achieved if blades are lengthened and the change confined to this and to directly related issues. It will be only partially realised if, say, increased exhaust area is achieved by using more exhausts and hence more cylinders, rather than if the change is confined to longer blades.

26.2.2.3 Part load

The steam turbine performs best at full load, where the whole system of nozzles and blades operates with proper steam velocities, with minimal throttle losses in valves. It will be less efficient on part load: energy is wasted by throttling down the pressure in valves. The steam jets are slow and do not strike the blades at the right angle, and the exhaust is too hot. If a turbine has to be used frequently at part load, it can pay to adopt nozzle rather than throttle control. The arrangement shown in *Figure 26.9* could show advantage over that shown in *Figure 26.10*.

In nozzle control, part load is achieved by using fewer nozzles and unthrottled steam. The technique is more common overseas than in the UK, though mechanical difficulty tends to make it less suitable for large machines.

26.2.3 Turbine construction

26.2.3.1 Form and arrangement

It would be ideal to have all turbines as simple as the single-cylinder machine, this is shown diagrammatically in *Figure 26.11(a)*. In fact, these are built for outputs up to 80 or 90 MW. Above these, the longer blades make two or more exhausts necessary and these need a separate shaft. The next step is to *Figure 26.11(b)*, which in the non-reheat form provides outputs of 250 MW or more.

When reheat is adopted there are two hot inlet flows, one at high pressure and a second at an intermediate pressure.

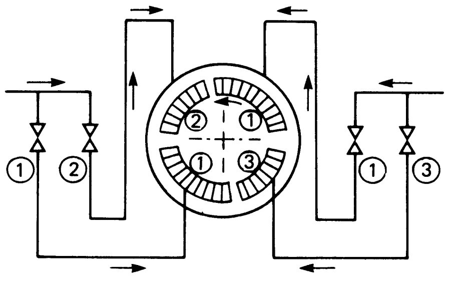


Figure 26.9 Steam inlet for nozzle control

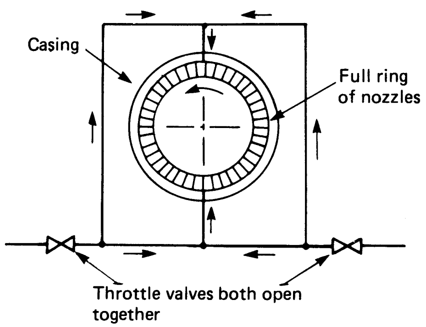


Figure 26.10 Steam inlet arrangement for throttle control

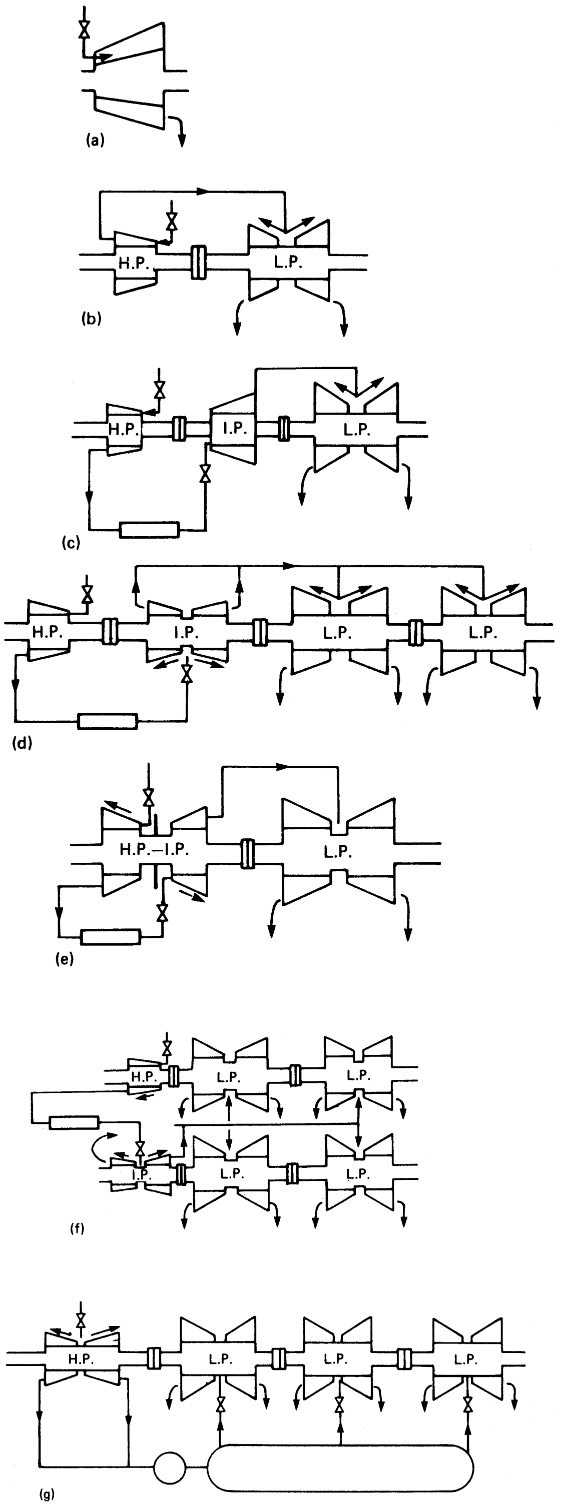


Figure 26.11 Forms of turbine: (a) simple cylinder, non-reheat; (b) two cylinder, non-reheat; (c) three cylinder, reheat; (d) four cylinder, reheat; (e) reheat with combined high power and interim power; (g) nuclear tandem for pressurised-water reactor

Some makers mount each on a separate shaft, giving the h.p., intermediate-pressure (i.p.) and l.p. arrangement of *Figure 26.11*(*c*). Others take both inlets to a common h.p.–i.p. rotor, *Figure 26.11*(*e*).The latter is common American practice.

The arrangements so far described are for rated speeds of 3000 or 3600 rev/min. An important design factor concerns the last blade size: increase in blade height permits the number of exhausts to be reduced. Current UK designs for 3000 rev/min turbines include blading heights up to 0.95 m on a mean diameter of 2.65 m. For 300 MW these use two exhausts as in *Figure 26.11*(*c*) and four for 600 MW, as in (*d*), or alternatively six where the higher efficiency obtained is economically justified.

Developments may lead to even longer blades, capable of 600 MW with two exhausts as in *Figure 26.11*(*c*) provided that low coolant temperatures are feasible. In some areas of the world these cannot be obtained, and only low area exhausts are justified.

Where the power rating is such that the exhaust area is beyond the capacity of the largest blades, the necessary area can be provided by a two-shaft or cross-compound arrangement, e.g. *Figure 26.11*(*f*), or by halving the speed as in the 1800 rev/min 1300 MW machines such as that in *Figure 26.11*(*g*). These arrangements accept, as most do, that eight exhausts on a single high-speed shaft line would not be a preferred solution.

There is a compromise between the two methods shown by using a cross-compound with a high-speed, h.p.–i.p., line and a low-speed, low-pressure line, either 3000/1500 or 3600/1800 rev/min depending on grid frequency. Inherent in these schemes is the fact that the lower speeds make it possible to use much longer blades and so enlarge the exhaust area and machine output. However, they do so at the expense of more metal for the low-speed rotors and casings, which are massive in comparison with those for a high-speed set.

Engineering solutions to the problem of increased output are always evolving and much larger turbine sizes than the current 1300 MW units could, no doubt, be devised. The limit tends to be imposed not by technical aspects of turbines, but by economics and by factors other than the turbine itself. An influence that the turbine contributes to this limiting of growth, however, is that the return tends to level off as size increases and can easily be nullified or even reversed if availabilities of the large machines fall even slightly beneath those of the smaller ones.

26.2.3.2 Foundations

The large turbine generator is supported some 10–12 m above the basement, a space that accommodates pipes, condensers and drainage. Supports were formerly mostly of reinforced concrete, but steel supports are increasingly preferred. They can be accommodated more readily into a building programme and are more predictable in their dynamic characteristics.

The block usually takes the form of an entablature supported on columns. It must be designed so that the shaft runs smoothly even though it may be slightly out of balance. The foundations must sometimes accommodate earth tremors. Both these requirements are better served if the structure is in steel than in concrete, because with steel the stresses, loadings and masses commercially used give a lower and more predictable and natural frequency. Vertical frequencies of about 10 Hz and sway frequencies somewhat lower are typical of such designs. These frequencies are substantially less than the shaft frequency of 50 Hz or half-shaft frequency of 25 Hz, both of which must be avoided if resonance is not to occur.

Concrete foundations tend to give calculated frequencies nearer 50 Hz, either above or below it. If the softness of the earth beneath the foundations is allowed for, these fundamental frequencies fall usually to well below 50 Hz, but not so low as with steel. While concrete foundations have given good service, their response is more difficult to predict and they are less likely to accommodate shaft and earth dynamics so acceptably as do steel foundations. Also, steel is usually more amenable to a quick station-building programme. However, in many countries concrete may be more economic and the choice has to be made on the circumstances.

26.2.3.3 Lubrication

The security of the turbine depends on a continuity of oil supply to its bearings and on the safe functioning of its hydraulic control gear. Further, the fire record in power stations shows that the oil system may play a part in starting or aggravating a fire. It pays to invest thought and money to make the oil system completely reliable and safe. Oil systems involve quite large flows with large turbines, partly because the bearings (especially the thrust block) have turbulent conditions in them unlike the less power-absorbent laminar conditions in former 30 MW sets. Special measures aim to reduce these power losses but they can still reach 3.5 MW for a 660 MW set.

Most machines have shaft-driven main oil pumps, usually centrifugal, primed by jet pumps or pumps driven by a hydraulic motor. The main pumps are backed by a.c. full-duty pumps and by a battery-operated d.c. emergency pump to provide oil until the shaft stops should there be an a.c. failure. Sometimes a small a.c. pump is fitted for use on barring. Shafts are generally jacked by h.p. oil to help starting.

In many arrangements the main oil pump also supplies the control gear with fluid; in others, separate pumps are used and feed control gear at much higher pressures than are used in the lubrication system. A fire-resistant fluid is sometimes used for control but is rarely considered for lubrication. The lubricant fluid is still oil, which can burn: it is therefore essential to design the oil system and fire-fighting plant to reduce fire risk. Attention is needed especially to pipe flange and coupling design. Any item that may fracture or leak oil should be eliminated or designed to reduce risk.

26.2.3.4 Governing

The governing and protection system operates control valves both to keep the speed steady within narrow limits and to prevent overspeed on sudden loss. The large turbine has stop and throttle valves at its inlet, and stop and interceptor valves at the i.p. inlet from the reheater. Usually, speed control is by adjusting the h.p. throttles. Overspeed on sudden load loss is prevented, first by closing the throttle valves and interceptors and then, should speed still rise, closing the h.p. and i.p. stop valves. Overspeed prevention requires fast valve action, usually well under 0.5 s closure and in some cases down to 0.15 s. The modern large machine has a high power/mass ratio and a sudden loss of full load can cause accelerations of 10–15% of full speed per second. This allows only a short time to cut off the steam supply, a situation aggravated by the stored energy of steam already in the pipes and spaces downstream of the valves. It is an important aspect of machine arrangement to minimise such spaces, e.g. using short steam loop pipes.

Other protective circuits besides overspeed operate the valves. The list usually includes low bearing-oil pressure, high vibration and poor vacuum.

Mechanical governors of large machines operate the steam control valves via hydraulic relays and power amplifiers and, eventually, by a servo-motor whose position is determined by speed level and held by a pilot valve and feedback. There are variations on this theme, some having an acceleration element.

Electric governing is becoming more general. Speed accelerations are sensed electrically and translated to an error signal in some adjustable way within the electric governor. The final movement and positioning of the steam valve is still hydraulic.

Some small machines use mechanical relays between speed sensor and valve, but most use hydraulics or electrical signalling and hydraulic power for valve movement. Lubricating oil is normally used, with pressures taken from the main oil pump at 0.35–2 MPa (59–300 lbf/in^2) but occasionally separate pumps are used for control and with pressures as high as 17 MPa. Fire-resistant fluids are used in the control systems of some large machines. In US practice, double piping is often called for if lubricating oil is used. The reason is that oil leaks are retained within the pipes rather than spill to atmosphere and cause a fire hazard.

26.2.4 Turbine support plant

26.2.4.1 Condensing plant

The condenser itself is large and costly consisting essentially of tubes through which cooling water is pumped and about which steam and air move. The steam is condensed and the air extracted. The condensate is drawn out by extraction pumps and in an efficient condenser it is nearly at the same temperature as the steam. It is returned to the boiler to use again. For every 1 °C by which the exhaust temperature is lowered in a typical 100 MW set, up to 0.5% extra fuel is burnt. Thus, a larger surface area of tubing allows the exhaust to be brought down more nearly to the coolant temperature; but, of course, this means a higher plant cost.

Having decided the surface area, however, it is important to use it properly. A particular offender is air, which can increase the exhaust temperature if it enters the condenser in large quantities, blanketing the tubes from steam. Excluding air is important: the condenser must extract such air as does enter without its forming pockets and blankets.

In the past, tubes have usually been made of brass. Where there is a severe corrosion risk, cupro–nickel is sometimes used. Dosing is usually employed, e.g. by ferrous sulphate injections, especially where there is a risk of corrosion or of erosion. Sometimes a cleaning gear is used in which shoals of floating balls pass through the tubes, are recollected and put back into the tubes.

Thin tubes of titanium are now becoming more usual. The evidence so far is that they offer much greater resistance to corrosion and erosion and often promise economic advantage.

With the high-steam conditions now used, it is vital to keep salts from the boiler and tube fixtures usually involve expansion or welding. Packings are rare. Sometimes anti-leak measures extend to the use of double tubeplates, with pressurised distilled water filling the space between the two plates. This ensures that, if there is a leak into the steam space, it will be of clean water and can be monitored. Large condensers usually have arrangements for on-load cleaning.

Ejectors are used to remove air from the condenser. Currently, steam-jet ejectors appear to be in favour but there are various alternative electrically driven ejectors.

The cooling water system can be quite elaborate and usually involves a considerable amount of both mechanical and civil engineering. In a typical British coastal 660 MW unit, for every 1 kg of steam that enters the condenser, about 30 kg of cooling water must be pumped through it. This means a flow of some 1000 m^3/min. A 1300 MW nuclear unit on the Californian coast uses as much as 3000 m^3/min of cooling water.

There is no standard cooling system using sea-water: the arrangement and equipment are made to suit the circumstances. Most have one or all of the elements shown in *Figure 26.12*. The system shown is elaborate because it is designed for the extreme conditions met, for instance, in the Mediterranean, where pollution causes growth of sea-grass. Periodic disturbance by storm can bring large quantities of this into the cooling system of a power station. The choking which results can take the turbines off load and is always a worry as regards chemical attack of the tubes. The weed has to be trapped and extracted before the condenser is reached if trouble is to be avoided.

Sand can also be a hazard, especially if it is hard and sharp. Settling tanks are often used to take sand out of the system.

In such systems the water inlet is taken well out to sea. This usually avoids the areas of higher weed density. The inlet is kept well above the seabed to avoid intake of sand, yet well beneath the surface to avoid oil slick. Water inlet positions must be planned most carefully with all these factors in mind and with others, such as avoiding warm water from the discharge re-entering the system. The work involves a careful on-the-spot survey with seasonal observations. It usually entails a hydraulic model.

Power stations are often sited on rivers. In some estuaries the water may be seriously short of oxygen or have aggressive pollutants, which promote tube damage. In others, the water is warm through use by other power stations or may be in seasonal short supply. Cooling towers then become necessary. Their essential principle is shown in *Figure 26.13*. Warm cooling water is sprayed into the atmosphere within the tower and passes its heat to the air, which in turn becomes warm and rises. The tower is thus filled with air which is warmer than that outside and the thermal syphon effect creates a through-draught. Wind across the top of the tower can stimulate the process.

In a sea-water system the turbine exhaust temperature is set by the temperature of the sea plus the temperature difference required by the condenser to transfer heat. In cooling-tower plants a further temperature rise occurs because the air is normally warmer than the sea; thus the station efficiency is lower and the cooling plant more costly.

Some water is lost by a wet cooling tower. About 1.4% of the total throughput is carried away by the air—about 27 000 t/day for a 660 MW plant. At some sites a reasonable source of water is too remote for economic use and a dry cooling system must be adopted.

Figure 26.14 shows the principle of the dry cooling tower. There is no direct contact between the water coolant and air, so no water is lost. Cooling takes place in large heat exchangers at the base of the tower across which air is drawn either by natural draught or by fans. The cooled water, which must be pure enough to use in the boiler, is sprayed into the steam entering a jet condenser. Such a system was installed at Rugeley in England in 1961 as an experiment at 120 MW and others have been built overseas since. An alternative arrangement is to use a surface condenser instead of a jet condenser. Raw water is then used as a coolant instead of feed water.

The dry-tower system is less efficient than the wet because the terminal temperature differences are higher so poorer vacua are obtained. Also the tower itself is much larger and so the system costs more. It can, however, enable large amounts of power to be produced in a dry terrain.

Where lower-power units are required, an air-cooled condenser (*Figure 26.15*) may be more economical than a dry tower. In the air-cooled condenser, the turbine exhausts to a heat exchanger mounted some distance above it. Steam goes through tubes and is condensed by air drawn or blown across them. The choice between air condensers and dry towers depends on many factors, but currently the former is used below about 100 MW and dry towers above about 200 MW.

26.2.4.2 Feed plant

The feed water produced in the condenser is taken back to the boiler through a train of devices. First the extraction pump raises it to well above atmospheric pressure and

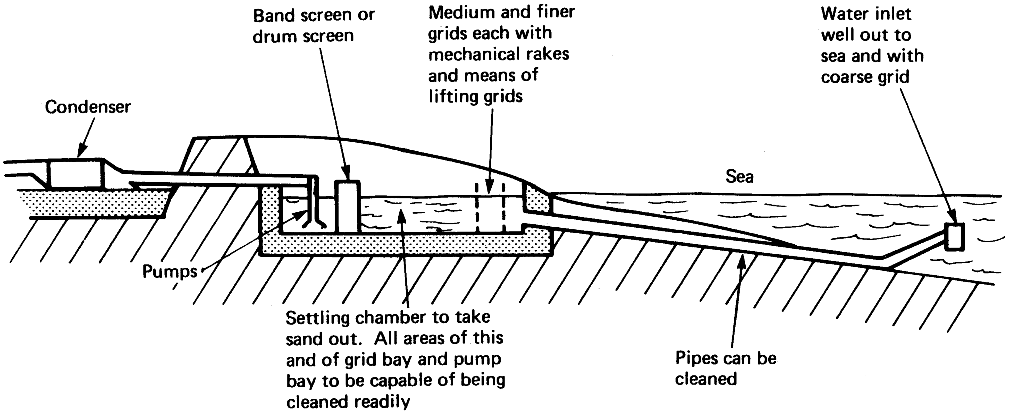

Band screen or drum screen
Medium and finer grids each with mechanical rakes and means of lifting grids
Water inlet well out to sea and with coarse grid
Condenser
Sea
Pumps
Settling chamber to take sand out. All areas of this and of grid bay and pump bay to be capable of being cleaned readily
Pipes can be cleaned

Figure 26.12 Sea-water cooling system where weed and sand are a problem

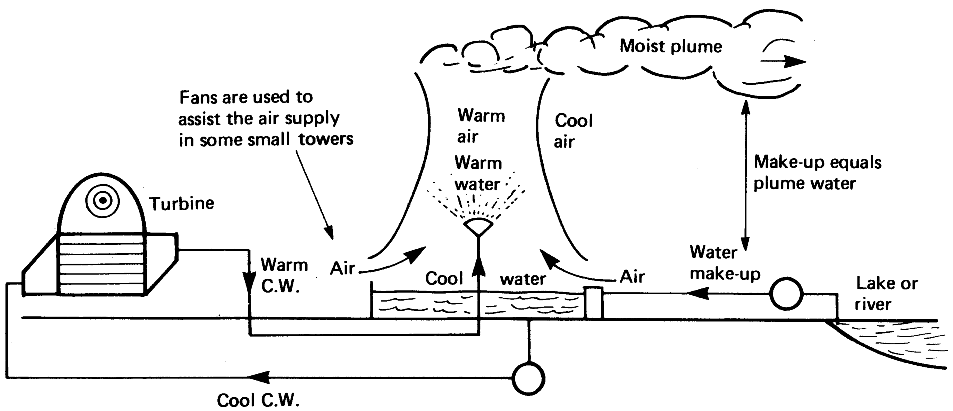

Moist plume
Fans are used to assist the air supply in some small towers
Warm air
Cool air
Warm water
Make-up equals plume water
Turbine
Warm C.W.
Air
Cool water
Air
Water make-up
Lake or river
Cool C.W.

Figure 26.13 Principle of wet cooling tower system

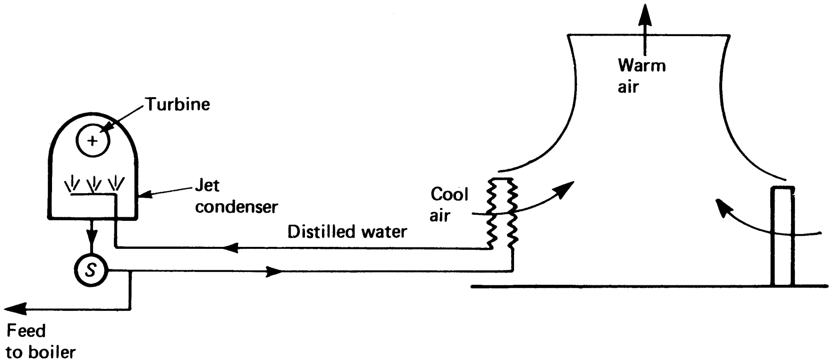

Warm air
Turbine
Jet condenser
Cool air
Distilled water
S
Feed to boiler

Figure 26.14 Dry cooling tower with jet condenser

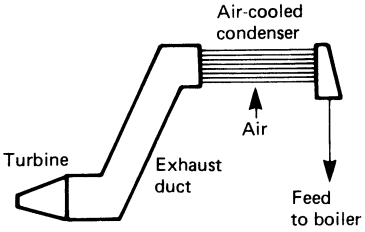

Air-cooled condenser
Air
Turbine
Exhaust duct
Feed to boiler

Figure 26.15 Air-cooled condenser

drives the water through a series of l.p. heaters, which warm it with steam drawn from the turbine. The heaters are usually of tubular type, the direct contact type is out of favour after accidents in which water spilled back into the turbine.

After leaving the l.p. heaters, the feed water is lifted to the deaerator where it is mixed with more steam bled from the turbine, and releasing dissolved air that would otherwise appear in the boiler. Sometimes the deaerator is incorporated in the condenser itself.

Water is then passed through a network of further tubular heaters and the feed pump, before passing to the boiler via flow-regulating valves. A typical arrangement is shown in *Figure 26.16*.

The feed-heating system is complicated, expensive and often takes up almost as much station space as the turbine itself. In spite of the fuel saved by it (and even for a small station and medium steam conditions this can amount to over 10%) there is a trend to simplify by reducing the number of feed heating stages, more particularly on stations for developing countries where simple robustness is desirable and where fuel is not expensive.

The feed pumps of a large power station absorb considerable power—up to about 15 MW for a 600 MW generator. There is much to be gained by driving them in an economical way.

26.2.5 Turbine operation and control

There are two basic ways of arranging and controlling the steam supply. Where boilers have substantial storage volume in their drums, the steam is arranged to pass direct from the boiler to the turbine (*Figure 26.17*(*a*)). This is common practice in Britain. A system more frequently met overseas is *Figure 26.17*(*b*), which uses bypasses. This is essential for supercritical boilers, for no drum is used and the boiler has but little steam storage capacity. But the bypass system is also advantageous where only small drums are used. It allows the boiler to operate on part-load and pass steam via the bypass system to the condenser without its going through the turbine. Thus, should the turbine

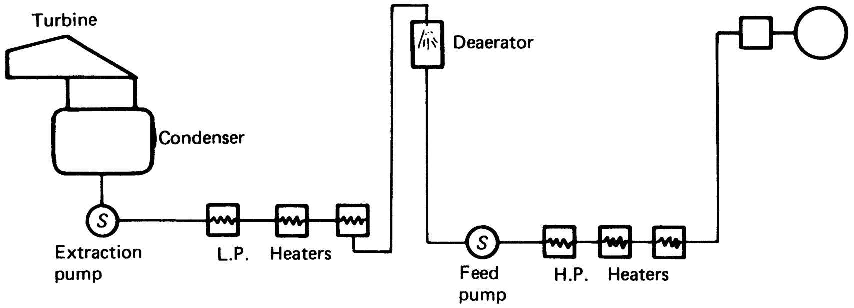

Turbine
Condenser
S
Extraction pump
L.P. Heaters
Deaerator
S
Feed pump
H.P. Heaters

Figure 26.16 Typical feed system

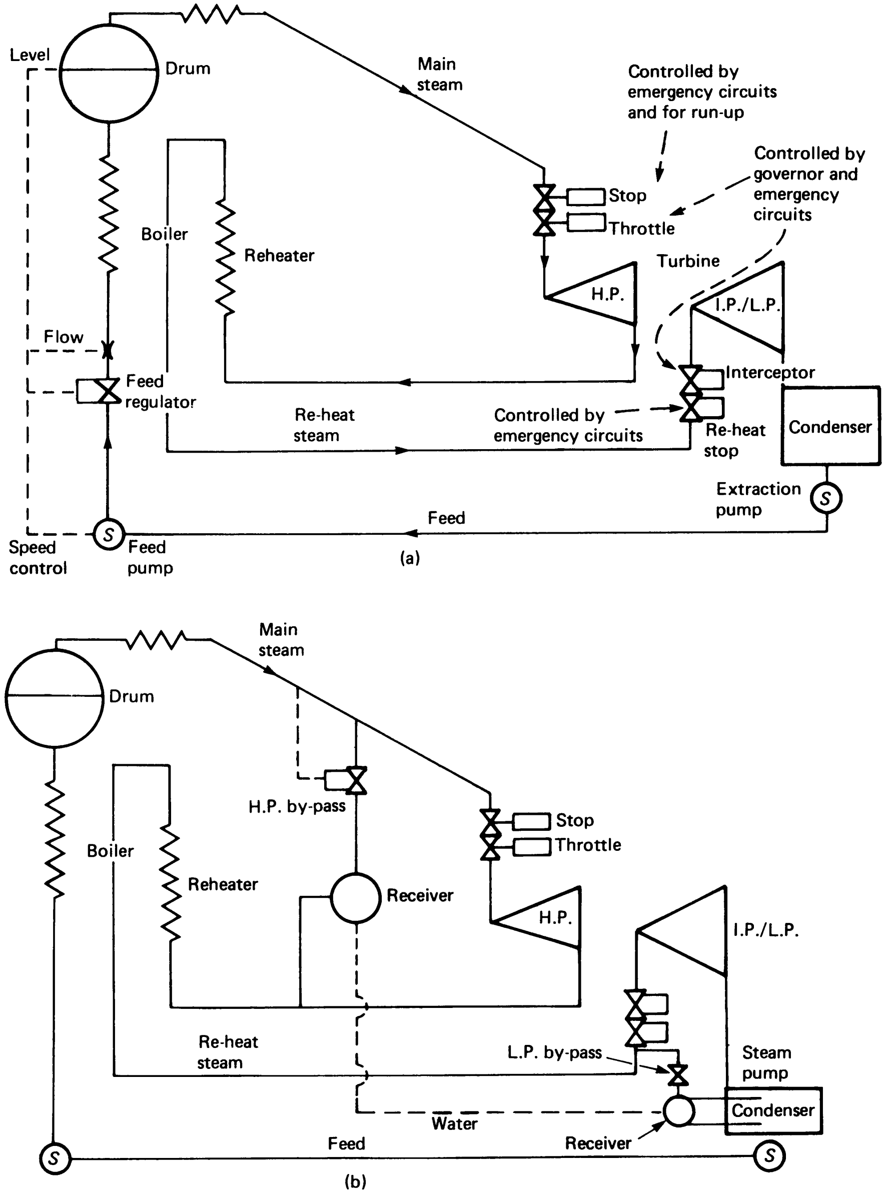

Level
Drum
Main steam
Controlled by emergency circuits and for run-up
Controlled by governor and emergency circuits
Stop
Throttle
Boiler
Reheater
Turbine
H.P.
I.P./L.P.
Flow
Feed regulator
Interceptor
Re-heat steam
Controlled by emergency circuits
Re-heat stop
Condenser
Extraction pump
S
Feed
S
Speed control
Feed pump
(a)
Main steam
Drum
H.P. by-pass
Stop
Throttle
Boiler
Reheater
Receiver
H.P.
I.P./L.P.
Re-heat steam
L.P. by-pass
Steam pump
Condenser
Water
Feed
Receiver
S
S
(b)

Figure 26.17 (a) Basic steam supply and control system; (b) system with steam bypass lose full load, rather than blow the safety valves the bypass will open and take steam from the boiler until the furnace heat has been adjusted to the load condition.

The bypass system has other advantages. For instance, the i.p. stage can be warmed through in parallel with the h.p. rather than in series with it, so applying hot steam sooner to the large i.p. cylinder and quickening the start of the unit as a whole.

When on-load in either system, the control of the turbine is by the governing system operating the throttle valves. Emergency signals of overspeed, oil loss, vibration, vacuum loss, etc., can close or reduce both stop and throttle valves by over-riding or acting through the governor. Instruments record the more important parameters of steam pressure, temperature, eccentricity, sliding, differential expansion, etc., and alarms operate from selected parameters.

Starting the machine is mostly automatic, or can be made so. There is a set sequence for taking each action, starting the oil system, barring the shaft, raising vaccum, etc. The most difficult operation and the one that requires most judgement is the admission of steam. The walls of the turbine valve chests and cylinders are thick and steam must not be admitted so quickly that severe thermal stresses develop. Many machines are equipped with instruments to measure the temperature difference across critical sections of chest or casing and so guide the operator on whether the warming through and starting is proceeding safely.

The operator must watch other factors too during a start. Most machines require special care to ensure that axial clearances are not taken up and these and the ease of sliding of the critical guidance and expansion surface are shown by appropriate instrumentation.

26.3 Gas turbine plant

During recent years many electric supply authorities have installed gas turbines, in unit sizes up to 150 MW, primarily for peak load or stand-by, but occasionally (where fuel is cheap) for base-load operation. Many industrial plants have also used gas turbines in connection with total energy schemes.

The major advantages of a gas turbine over steam plant are low capital cost, quick starting, small erection time and the facility for using a wide range of fuels from heavy oil to natural gas; also remote control or fully automatic operation is readily achieved. The disadvantages are the low thermal efficiencies (typically 25–30%) unless considerable complication is involved; the use of suitable materials to withstand very high temperatures combined with attack from undesirable elements in the combustion products is still a critical issue.

In the *open-cycle* plant, air is taken from the astmosphere, burnt with the fuel in a combustion chamber, passed through the turbine and finally exhausted to the atmosphere. In the *closed-cycle* plant air (or gas) is circulated around a closed circuit, heat being supplied to the air through a heat exchanger from an external combustion chamber. The former is the more common; its lower capital cost and adaptability to the relative importance of capital cost and efficiency are cogent factors.

26.3.1 Open-cycle plant

The simplest form of gas turbine consists of a compressor, a combustion chamber and a turbine as shown in *Figure 26.18*. This is the 'single shaft' gas turbine which is generally used for electricity generation applications. Driven by the turbine, the compressor delivers compressed air to the combustion chamber where there is continuous combustion of the injected fuel at constant pressure. In order that the temperature after combustion may not be too high, the ratio of fuel to air, by weight, is of the order of 0.017. This ratio is too low for combustion, so the air from the compressor is divided at entry to the combustion chamber. Part of it is used in the combustion zone and part acts as a coolant. The two streams are thoroughly mixed before entering the nozzles of the turbine at the permissible temperature. *Figure 26.19* shows the outline of a gas turbine combustion chamber. The combustion of the fuel must be as nearly 100% efficient as possible to ensure a high plant efficiency and to reduce carryover of carbon matter to the turbine blades. The quantity of fuel injected controls the output of the plant but it is essential that flame stability shall be maintained over a wide range of fuel injection rates.

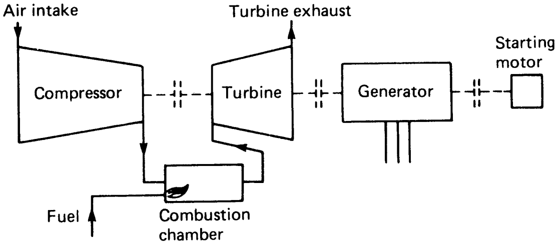


Figure 26.18 Simple open-cycle single-shaft gas turbine. Note that the turbine may also be split into high-pressure and low-pressure elements to form a two-shaft gas turbine

The heated air from the combustion chamber, including the products of combustion, is expanded in the turbine and is then exhausted to the atmosphere. Approximately two-thirds of the power developed by the turbine is used up in driving the compressor, the balance is used for producing the net output power for the generation of electricity or other purposes.

26.3.1.1 Simple power relations

Figure 26.20 shows the ideal pressure–volume diagram representing the ideal cycle work derived from the turbine and the ideal work necessary to drive the compressor. In practice the work done in the turbine is less than that for the 'ideal', i.e. the isentropic efficiency η_t of the turbine is less than unity. Similarly the work done in the compressor is greater than for the ideal case, i.e. the isentropic efficiency η_c is also less than unity. The useful work per unit mass of air is the difference between the work done by the turbine and the work required by the compressor, namely

$$W = c_p[T_2(1 - 1/k)\eta_t - T_1(k - 1)/\eta_c]$$

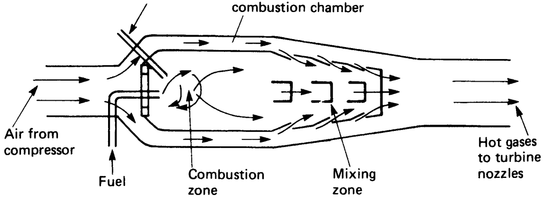

Igniter
Air stream around combustion chamber
Air from compressor
Fuel
Combustion zone
Mixing zone
Hot gases to turbine nozzles

Figure 26.19 Scheme to show the layout of a combustion chamber

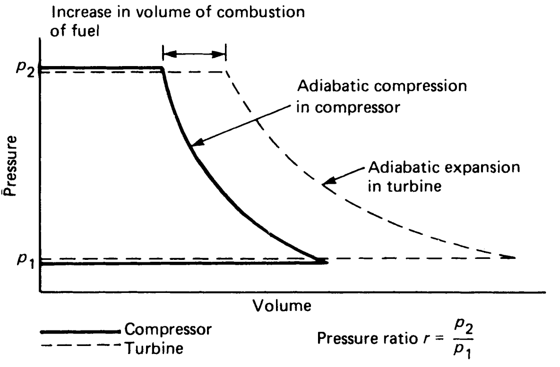


Figure 26.20 Pressure/volume diagram for an ideal gas turbine cycle

where $k = r^{(\gamma - 1)/\gamma\psi}$, r is the pressure ratio of compression and expansion, $\gamma\psi$ is the ratio of specific heats, c_p is the specific heat at constant pressure, T_1 is the absolute temperature at compressor inlet, and T_2 is the absolute temperature of the air after combustion and before entry to the turbine.

It is evident that if η_t, η_c and T_2 are too low this expression could have a negative value. The turbine output would not be sufficient to drive the compressor and there would be no useful work. Hence the success of the gas turbine depends on the attainment of high isentropic efficiencies in the compressor and turbine and the ability to operate at a high turbine inlet temperature.

The gas turbine must operate at very high temperatures to develop efficiencies comparable with those attained in steam turbines and diesel engines. The turbine blades must withstand both high stresses due to rotation and the high temperature of the combustion gases passing over them. The success of the gas turbine has been made possible by the development of steels which will permit stressing at high temperatures without excessive creep. To allow the use of higher gas temperatures the turbine blades are often cooled by fluid circulating through passages in them. The designed maximum temperature of operation is determined by the required 'life' of the turbine. A lower temperature of operation will reduce the creep rate of the stressed parts and increase the number of hours of running life of the turbine. Temperature gradients in rotors and blades also induce stresses and alternating stresses due to blade vibration could cause failure by fatigue. With the materials and design experience now available, efficient and reliable gas turbine plants are being built and many are in successful operation.

The gas turbine is not a self-starting machine; it must be brought up to speed by a starting motor until the compressor is running fast enough to attain an adequate pressure ratio and component efficiencies for a self-sustaining or viable cycle. After ignition of the fuel the machine will increase in speed until the turbine drives the compressor and provides the useful output. If the gas turbine drives a generator with separate exciter, the exciter may be used as a starting motor.

As in the operation of a steam turbine, starting up and shutting down must take place with due precaution to avoid excessive temperature gradients, particularly in the rotors.

Contemporary gas turbines have compressor pressure ratios (delivery pressure/inlet pressure) of as much as 30:1 resulting in combustion chamber pressures up to 3000 kN/m^2 (435 lbf/in^2 gauge).

26.3.1.2 More complex plant

The efficiency of the plant described above is somewhat low (25–30%). It has advantages of simplicity, low cost and independence of water. Higher efficiencies (30–40%) may be attained at the expense of greater complication and cost.

The addition of an efficient heat exchanger has a most pronounced beneficial effect on efficiency. *Figure 26.21* shows a gas turbine plant in which a heat exchanger is included. Heat in the turbine exhaust is transferred to the air delivered by the compressor so that less heat is required from the combustion of fuel. The output of the plant is not increased but the reduction in the amount of fuel burned improves the efficiency. To justify its space, weight and cost the heat exchanger must have a high thermal ratio. It usually consists of a shell containing nests of small-bore tubes over which the turbine exhaust passes giving up heat to the compressor delivery air passing through the tubes on its way to the combustion chamber; however, heat exchangers have not been adopted widely for reasons of cost, bulk and reliability.

For electric power generation the gas turbine is essentially a constant-speed machine. The simple, single-shaft open-cycle gas turbine is not generally flexible as it needs to operate at full load at constant speed to give its highest thermal efficiency. Some flexibility is possible by adopting a 'split shaft' (sometimes known as a 'two-shaft') machine, but much better flexibility and performance can be obtained at the cost and inconvenience of complexity. There are many possible combinations of compressors, intercoolers, turbines, combustion chambers and heat exchangers. *Figure 26.21* shows a plant in which the compressor and turbine are each in two stages. Interstage cooling of the compressed air increases the efficiency of compression. Another combustion chamber may be interposed between the h.p. turbine and l.p. turbine to increase the output of the plant. The separation of the turbine producing useful output from that driving the compressors makes the plant more adaptable to load variations with less reduction of efficiency at part load. The cost of more complex plant is, of course, justified only when the load factor and fuel costs are high.

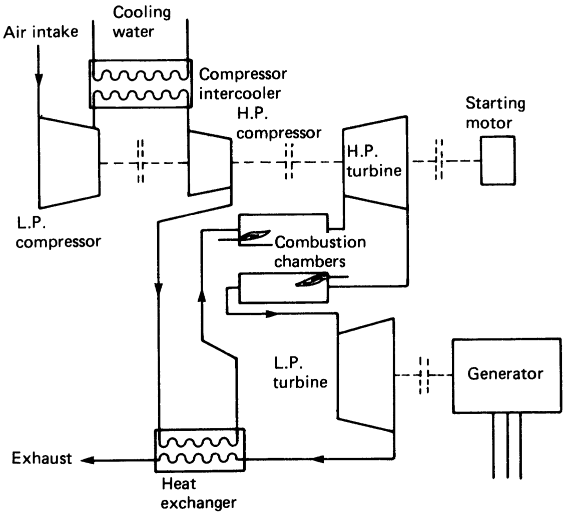


Figure 26.21 Compound-cycle gas turbine

26.3.1.3 *Aircraft-type gas turbines*

Since the 1960s there has been widespread application of aerojet gas turbines in which the propulsion jet is coupled to a heavy duty power turbine. The development of this type of turbine makes use of the extensive research carried out by the aero-engine industry and employs one or more aircraft jet engines to discharge into a gas turbine. Such units are reasonably inexpensive, compact and have exceptionally quick starting properties, e.g. full load from a cold start in 2 min.

The jet engine on aircraft has frequent skilled maintenance and normally operates at a high altitude with low atmospheric pressure. By de-rating the engine and incorporating certain small modifications it can be operated successfully at sea level at a power between its normal flight and take-off ratings and can give 20 000–25 000 h operation between overhauls. Its compactness also makes it suitable for mobile power plants.

26.3.1.4 *Free-piston gas generator*

Instead of the compressor and combustion chamber of the conventional gas turbine, one or more free-piston gas generators may be used to produce the hot gas for discharge into the turbine. This comprises a cylinder containing two free pistons; fuel is admitted between the pistons and, by compression ignition and appropriate operation of valves, hot gas is emitted from the exhaust ports. The chief advantage is a higher overall efficiency (30–35%) than the conventional gas turbine, but cost and maintenance requirements are generally greater so that applications are limited. A few successful plants of between 1 and 10 MW are, however, in operation.

26.3.2 Closed-cycle plant

In the open-cycle plant a continuous supply of air from the atmosphere is drawn into the compressor. The intake pressure and density of the air are, therefore, fixed at atmospheric conditions. By using the closed system shown in *Figure 26.22*, in which the same air is circulated continuously, the pressure and density of the air may be increased. More power is obtained from a given size of plant and output may be varied by changing the pressure and mass flow of air, the pressure ratio and speed being retained at the optimum values.

Combustion must be external to the air stream and the heat transferred to the air in an air heater. Waste heat is extracted from the system in a cooler placed before the compressor intake.

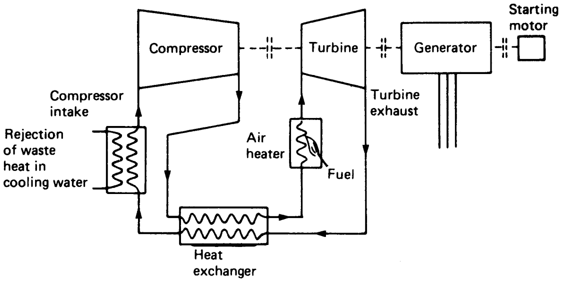


Figure 26.22 Closed-cycle gas turbine

A further advantage of the closed-cycle plant is that the air is not contaminated by products of combustion and dust drawn in from the atmosphere. Other gases than air might be used if their physical properties were superior for the purpose, e.g. helium has been used for nuclear reactor cycles.

The main disadvantage of the closed cycle is the air heater. It is very difficult to attain high rates of heat transfer to a gaseous fluid like air, and excessively high tube temperatures are easily developed with consequent failure. Increasing the air velocity through the tubes improves the heat transfer but increases the pressure drop of the air in passing through the tubes. A similar problem occurs in the design of heat exchangers where improvements in heat transfer by higher velocities and greater turbulence result in excessive pressure drop. The solution is always a compromise between these conflicting factors.

26.3.3 Combined-cycle plant

The combined cycle, in its conventional power generation form, recovers much of the gas turbine's unutilised exhaust heat in a heat recovery steam generator (h.r.s.g.). The h.p. steam generated by the h.r.s.g. is used in a conventional Rankine cycle where the steam is expanded in a condensing steam turbine/generator set to provide an additional conversion of fuel energy to electrical energy. Unfortunately, because not all the exhaust energy can be recovered by the h.r.s.g. and the efficiency of the Rankine cycle is unlikely to exceed 38%, considerable energy is still eventually lost to atmosphere as low temperature heat, principally from the condenser. (The Rankine cycle thus acts as a bottoming cycle for the gas turbine's higher temperature cycle.) Two typical arrangements are shown in *Figure 26.23*.

Figure 26.24(*b*) illustrates the use of a high efficiency simple gas turbine in combined cycle mode. The situation has considerably improved and overall power generation efficiencies in the range 44–55% are currently available. Application is, however, still restricted due to the need for

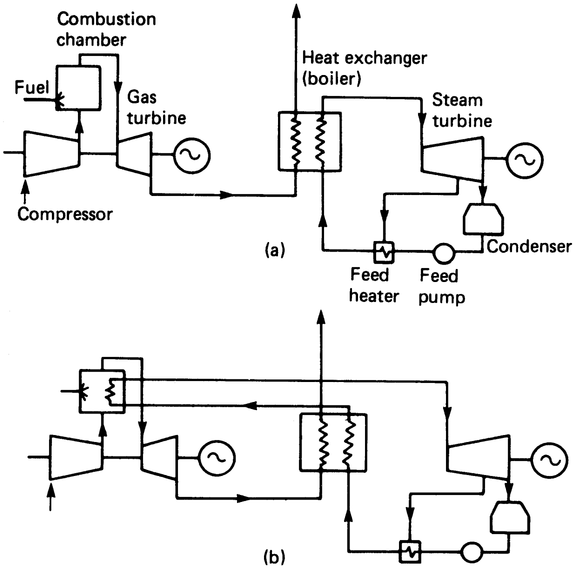


Figure 26.23 Gas/steam cycles: (a) simple cycle; (b) additional evaporator in the combustion chamber

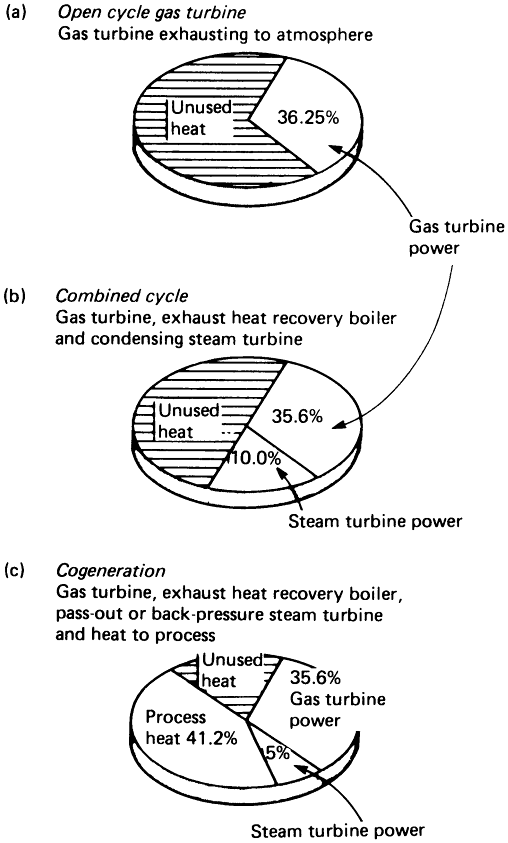


Figure 26.24 Energy usage in various gas turbine systems

clean but expensive fuels such as distillates and natural gas. However, the above-mentioned application of fluidised bed combustion systems capable of burning all coal types, peat, refinery bottoms and cokes, is anticipated to change this. The increasing influence of emission controls also favour fluidised bed combustors which can absorb sulphur oxides (SO_x) and produce inherently low-nitrogen oxides (NO_x) without performance penalties. The addition of flue-gas scrubbers to conventional steam plant to absorb SO_x and NO_x causes a reduction in plant efficiency of 2–3 percentage points and 25–30% increases in capital cost. This will give a considerable incentive to utilities to give serious consideration to the new generation of fluidised bed combined cycles that are now available.

26.3.4 Cogeneration/CHP plant (see also Section 26.5.10)

If instead of using recovered exhaust heat solely for additional power generation in a bottoming cycle, it is used directly as heat for a process requirement, further substantial improvements in energy utilisation can be achieved. This is known as 'cogeneration' (of both heat and power), or sometimes as 'combined' heat and power (c.p.h.).

Figure 26.25 shows diagrammatically some of the heat utilisation forms that are possible with gas turbine cogeneration plant. In each case exhaust heat is used to eliminate, or significantly reduce the requirements for additional combustion of fuel in separate heat generation plant (boilers, kilns, combustors, fired heaters, etc.). In addition, since a number of the applications involve eventual heat rejection at temperatures close to ambient, most (if not all) of the exhaust heat can be credited to the overall thermal efficiency of the cogeneration plant. Accordingly, cogeneration installations can achieve very high overall efficiencies, up to 90%, depending on the nature of the heat utilisation.

From the point of view of economically effective plant installations, the two most interesting forms of heat use shown in *Figure 26.25* are the direct use of the gas turbine exhaust gases and the use of exhaust heat for steam generation. The first is of interest because no expensive intermediate heat exchange equipment is required to recover the exhaust heat; this possiblity arises because the gas turbine exhaust gases are relatively clean, particularly when the gas turbine is operating on high-quality natural gas, and they contain a large percentage of oxygen which can be used in further combustion processes downsteam of the gas turbine.

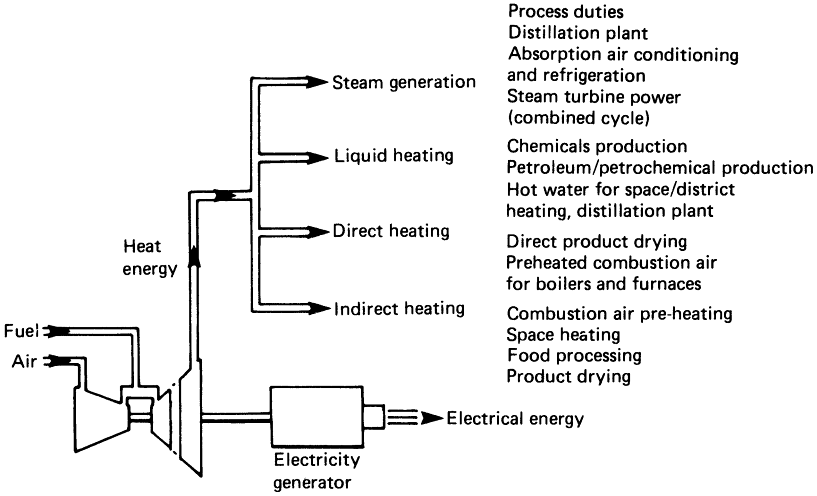

Steam generation
Process duties
Distillation plant
Absorption air conditioning and refrigeration
Steam turbine power (combined cycle)
Liquid heating
Chemicals production
Petroleum/petrochemical production
Hot water for space/district heating, distillation plant
Direct heating
Direct product drying
Preheated combustion air for boilers and furnaces
Indirect heating
Combustion air pre-heating
Space heating
Food processing
Product drying
Heat energy
Fuel
Air
Electrical energy
Electricity generator

Figure 26.25 Gas turbine cogeneration utilisation

Accordingly, the exhaust heat can be used directly for duties such as product drying or preheating of combustion air for product heating in fired heaters (petrochemical industry), kilns (brick and ceramics industry), coke ovens (steel industry), etc.

The steam generation function is of interest because it is often found that the steam is required at relatively low process pressures, say at an absolute pressure of 274–620 kPa. Since it is economic to generate steam in an h.r.s.g. at much higher pressures, say 1825–6310 kPa, the opportunity arises to expand the steam from its generation pressure to the required process pressures using a back-pressure steam turbine, with steam extraction for intermediate pressure requirements. Thus, valuable extra electrical megawatts can be generated, thereby improving the economics of an installation, while serving the process steam requirement. Such plant would be properly referred to as combined cycle cogeneration equipment, although frequently only the cogeneration term is applied.

Figure 26.24(*c*) shows the energy utilisation of a typical combined cycle cogeneration plant. The figure shows that overall energy utilisation efficiencies well in excess of 80% can be achieved with this plant configuration. The performances of typical systems are shown in *Table 26.6*.

26.4 Hydroelectric plant

26.4.1 General

Hydroelectric plants convert potential energy of water into an electrical output. Rivers, upland lakes, coupled with their catchment areas, estuarial tidal cycles and upper reservoirs of pumped storage plants can provide the source of energy to be converted. The process involves flow of water from the source to the turbine tailrace, which acts as a sink. In the process of conversion, use is made of water turbines, of associated civil structures and of rotating electrical machinery.

The power supplied to the turbine, P_d (kilowatts) is given by the product of the rate of mass flow ζQ (tonne per second) and of the net head across the turbine H_n (meters) corresponding to this flow:

$$P_d = 9.81\,\zeta Q H_n \tag{26.3}$$

where ζ is specific mass (tonne per cubic metre) and Q is the volumetric discharge (cubic metres per second).

The net head (H_n) represents only a fraction, however large, of the total or gross head (H_g) which for all types of hydroelectric plant is defined as the difference in elevation between the water levels at the upstream (intake) and downstream (tailwater) limits of the installation, when there is no flow. The difference between these two heads represents the losses within the plant, but outside the confines of the water turbine. These losses are either due to flow related phenomena or arise because of the need to set certain types of impulse turbines well clear of the tailwater level. The ratio H_n/H_g is designated as the hydraulic plant efficiency (η_p) which, as will be shown later, can represent a significant parameter when evaluating the worth of alternative designs of the civil works.

While hydroelectric projects are normally considered in terms of the gross heads they create, water turbines are invariably designed bearing in mind the need to maximise the weighted efficiency at a number of net heads. Exploitable heads vary from a few metres to 2000 m.

It will be apparent that, even at the highest heads, the available energy levels per unit of mass flow, are sub-

Table 26.6 TYPICAL SYSTEMS (simple-cycle, combined cycle and cogeneration)

	PG5371(PA)	PG654(B)	PG9171E	PG9301F
Base load (kw)	26 300	38 340	123 400	212 200
Overall efficiency (%)	28.45	31.42	33.78	34.13
COMBINED-CYCLE SYSTEMS				
Gas turbine base load (kw)	25 950	37 980	122 750	210 600
Steam turbine (kw)	12 250	18 800	66 000	124 900
Total base load (kw)	38 200	56 780	188 750	334 500
Overall efficiency (%)	41.9	46.6	51.5	54
COGENERATION SYSTEMS				
Base load (kw)	25 950	37 980	122 750	210 600
Heat in exhaust at 12°C (kw)	48 770	62 520	182 330	306 670
Overall efficiency (%)	81.9	82.5	83.2	83.5
Steam produced at 100°C (kg/s)	21.2	27.2	79.2	133.2

Basis for nominal performance ratings:

15°C, sea level, no intake or exhaust losses for open cycle, 245 mm H_2O back-pressure loss in exhaust heat recovery systems.
Performance at power turbine coupling.
Steam production figures assume 98% boiler efficiency.

Cogeneration systems are tailored for specific projects and the above examples are typical.

stantially lower than those associated with thermal plants. Typically a conventional 660 MW thermal unit would require a water mass flow of 2000 t/h to achieve its full output. A similar rate of flow in a hydroelectric unit, operating at 2000 m head would produce an output of under 10 MW. At this flow, and at a head of 20 m, the output would be below 100 kW, the capability of a small mini hydroelectric unit. However, notwithstanding the very low levels of specific energy, high output per unit can be achieved, even at moderate heads. This clearly involves very high rates of mass flow.

The greatest outputs, on modern units, have been achieved at net heads of around 120 m where flow rates of 700 t/s yield outputs of 715 MW. Similar rates of flow have recently been considered on some very low head, tidal installations, currently under study. These would result in outputs of 40 MW or so. In the former case, the diameter of the water turbine runner is 8.5 m; in the latter case a 9.0 m runner is envisaged. The largest runner diameter already in service is believed to be 9.3 m. It will be appreciated that machines of these sizes would operate at very low rotational speeds. The 715 MW units operate at 90.9 and 93.3 rev/min on the 50 and 60 Hz systems of Paraguay and Brazil, respectively.

The proposed synchronous speed of tidal units amount to only 50 rev/min and even that speed represents a compromise between the optimum speed of the turbine for energy capture (47 rev/min) and the need to ensure that the generator design parameters, such as the number of poles and the rating, could be achieved without recourse to the use of an uneconomical, oversized rotor. It is the quest for higher rotational speeds and for higher turbine efficiencies which provides the motivation for the continuing development of turbine designs.

26.4.2 Types of plant

The very wide range of possible operating conditions has led to the development of a number of diverse designs of rotodynamic machinery. However, as far as the conceptual design of plant is concerned, there are in the main only two basic types of arrangement of the powerhouse within the scheme; there are, of course, some variations on the theme.

One of these arrangements is known as a 'run of the river' plant (see *Figure 26.26*). Here the powerhouse is either incorporated in the dam, or is located alongside it, i.e. is local to the dam. In such an installation, the gross head is determined by the difference in levels across the dam. The majority of such installations operate at heads of less than 100 m. However, even substantially higher heads, amounting to 200 m or so can be accommodated with this arrangement, as evidenced by the example of the Boulder Dam in the USA. This arrangement is also employed on tidal schemes, where the gross heads involved are well under 10 m.

The alternative arrangement is known as a 'diversion' plant (see *Figure 26.27*). Here the supply is taken from a dammed river or lake, from which water flows through a head race canal to a head pond or forebay in the vicinity of the remote powerhouse. From the forebay, the water flows to the turbine through a system of pressurised pipes, known as penstocks. The purposes of the forebay is to ensure that sudden changes in rates of flow, dictated by turbine controls, can be accommodated without producing unacceptable changes of the levels in the canal. In a variation on this design, the canal is replaced by a low pressure tunnel. In such a case, when control considerations justify it, a surge chamber or surge tank is provided.

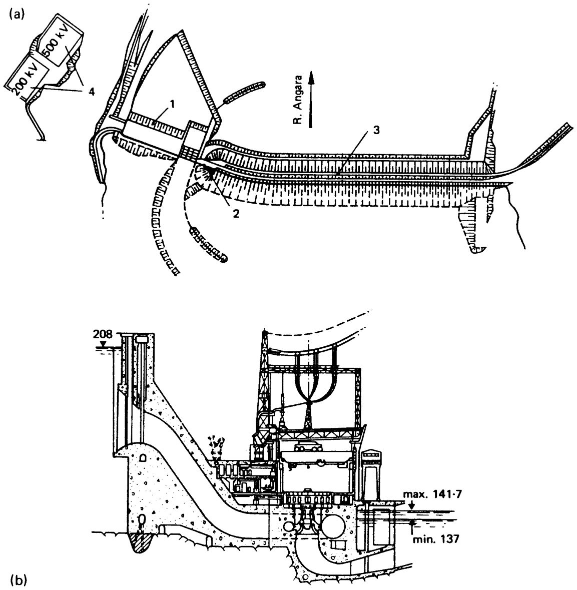


Figure 26.26 A 'run of the river' plant (Boguchauy hydroproject). (a) Layout: 1, powerhouse; 2, spillway; 3, rockfill dam; 4, switchyard. (b) Section through the dam and centre line of a unit

The purpose of either of these devices is to protect the low-pressure tunnels from water hammer caused by turbine controls and to permit quick starts of the turbine, without producing significant loss of head, caused by the need to accelerate quickly the water column within the tunnels.

The powerhouse in a diversion-type plant may occasionally be located underground. Here, the tailrace may take the form of a tunnel. In cases where the tunnel runs full and is of sufficient length for the effect of the inertia of the water column to become significant a surge chamber may have to be provided downstream of the turbine draft tube. Thus it is possible to have plants of this type with no surge protection or with one surge chamber upstream or downstream of the plant or, alternatively, with two surge chambers or tanks.

Finally, in micro units which, because local demand for power is limited, may take only a fraction of the total discharge of a river, it is often possible to divert sufficient flow without the use of a dam, simply by providing a parallel channel.

Thus, it will be noted that in diversion plants the gross head is either only partly or not at all dependent on the level above the toe of a dam. In any case, a large proportion of the gross head, even in a system containing a dam may be due to the geodetic head difference between the toe of the dam and the tailwater level. Thus a diversion type of plant makes it possible to create high heads without the need to build tall dams.

Another important aspect in the consideration of the design of hydroelectric plant is concerned with the question of how best to develop the potential of a river or a basin. Given the statistical information concerning the 'run off' over a period of many years, the designer must consider the various ways in which the potential can best be exploited. The local load demand and its probable rate of growth must also be taken into consideration.

Where the hydraulic potential is large but the local demand is small or even non-existent, the designer must

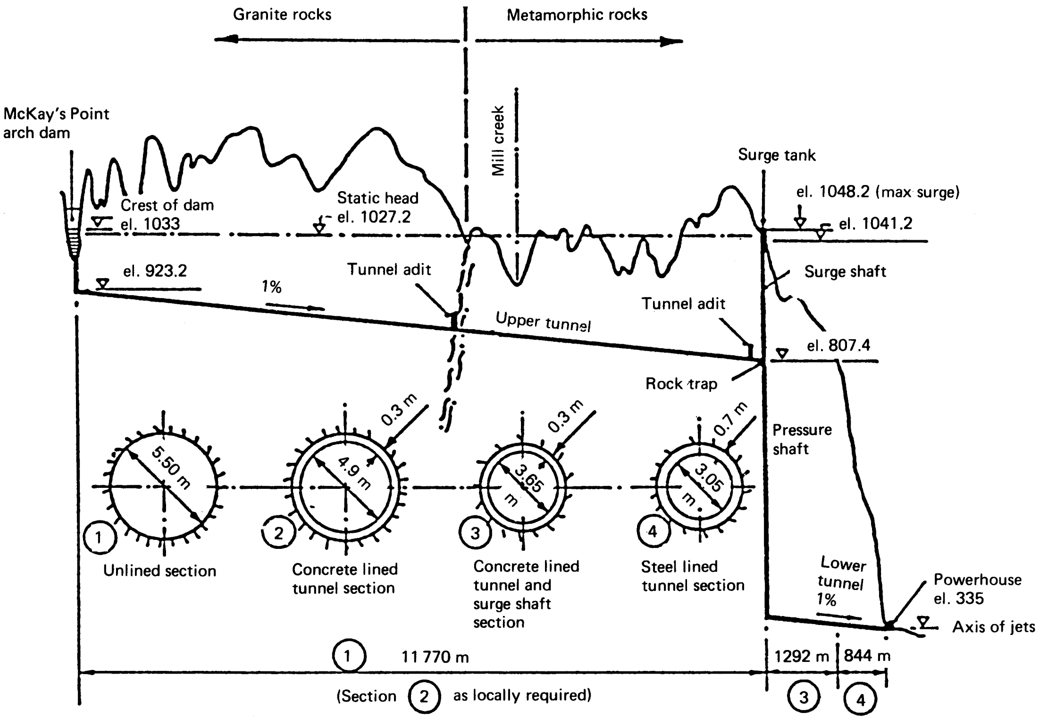

Granite rocks
Metamorphic rocks
McKay's Point arch dam
Mill creek
Surge tank
Crest of dam
el. 1033
Static head
el. 1027.2
el. 1048.2 (max surge)
el. 1041.2
el. 923.2
1%
Tunnel adit
Surge shaft
Upper tunnel
Tunnel adit
el. 807.4
Rock trap
0.3 m
0.3 m
0.7 m
5.50 m
4.9 m
3.65 m
3.05 m
Pressure shaft
1
2
3
4
Unlined section
Concrete lined tunnel section
Concrete lined tunnel and surge shaft section
Steel lined tunnel section
Lower tunnel 1%
Powerhouse el. 335
Axis of jets
1
11 770 m
1292 m
844 m
(Section 2 as locally required)
3
4

Figure 26.27 A diversion plant (Collierville)

consider the viability of transmitting large amounts of power to distant load centres, as was done in the cases of Churchill Falls in Labrador (the power output of which is exported to the USA) and Cabora Bassa in Mozambique (which was constructed on the assumption that it would export power to South Africa). Yet another possibility is presented when a substantial load such as an aluminium smelter can be brought to the vicinity of the plant, as was the case on the Volta River in Ghana.

Where a scheme is to meet a demand which is expected to grow substantially with time, economic considerations dictate that, if at all possible, the scheme should be built in stages. Where damming of the source of power is involved, this of course must be done before any power at all is produced. It is possible, however, to install only the amount of machinery commensurate with the initial demand. Alternatively, a scheme may be evolved for using only a portion of the total potential head by dividing the overall scheme into a number of stations to be developed sequentially.

Another aspect of conceptual design concerns the case of a potential site where the available gross head is substantial, the site topology favourable, but the 'run off' insufficient to support the required levels of output. In such a case, diversion of 'run off' from neighbouring catchment areas through low-pressure tunnels is often employed. Only when the flow-duration data including normal seasonal variations and effects of unusual dry and wet years has been established and the storage capacity determined can the most economic mode of operation be considered. In addition, the existing load demand and its anticipated rate of growth must be taken into consideration.

Where there is no storage capacity or where the storage capacity is only small, the output, up to the rated capability of the plant, will tend to be in step with the flow. At times of flood, excess of flow over that used for generation must be spilt over the dam. Even in such a plant, it may often be economical to increase the capability of the rotating plant well in excess of the rating based on time averaged rates of flow. This is so because the cost of the civil works represents a major component of costs and the marginal increase in costs arising as a result of the use of oversized rotating plant could be more than compensated for by the value of the additional energy recovered during the period of floods.

Where the inflow to a scheme is seasonal, as is typically the case where melting of snow or of glaciers is the source of water, provision of storage capacity is essential. Where storage capacity is provided, it is not only possible, but often very desirable, to increase the capability of the plant well beyond that corresponding to a plant that would operate on a base-load principle. This is due to the fact that hydroelectric plants are inherently suitable for use as peak-lopping plants. Where pondage is small, this type of operation can be achieved by providing a pumping facility for replenishing the stored energy. On any but the very high-head plants using impulse machines, the ability to operate both as a pump or as a turbine can be achieved in a single reversible pump-turbine unit. During pump operation the generators would operate in the motor mode. From the foregoing it will be clear that in the majority of hydroelectric plants the rated capability of the plant does not relate directly to the firm volume of output. While steps can be taken to reduce seasonal variations by provision of storage, the main uncertainty stems from variations in precipitation, be it in the form of rain or snow.

It is therefore desirable to install either a suitable mix of schemes or to ensure a certain level of stand-by capacity in thermal plant. Only very occasionally, on small schemes, taking their supply from large natural lakes, can the rated capacity, with suitable allowances for outages, be taken as firm power.

26.4.3 Power station

Depending on the layout of a scheme, the turbines, their auxiliaries and the electrical plant may be housed either at or below ground level. Run of the river situations are invariably housed at ground level and are located either inside or alongside the dam. On very low head schemes they may be housed within the structure of a submerged weir, with provisions for spilling excess flow, over the roof and side of the station. In diversion schemes, the power station is housed either in a purpose-built structure at ground level or in an underground cavern.

By comparision with thermal stations, the number of auxiliary systems that must be housed either within, or in the vicinity of, a hydroelectric power station is fairly limited. Such systems that are essential include the following:

(1) Gates and/or valves used for isolating the turbine, together with the associated pumping sets and pressure receivers. Note that closure must be effected either under the action of gravity or by utilising stored energy.
(2) Governors, actuators and servomotors and a power oil system containing sufficient stored energy for operation of a closing–opening–closing cycle.
(3) Duplicated de-watering system with a proportion of the de-watering pumps supplied from a secure battery based supply.
(4) A compressed-air system capable of charging and topping up the pressure receivers under (1) and (2). Where synchronous condenser operation is involved or where pump turbines are employed, the compressed air system should have sufficient stored capacity to permit the blow down of one unit at a time and have the additional capacity to make up air losses following the blow down.
(5) Cooling systems for generators, transformers, pumping sets and for thrust and journal bearings of the turbo-generator.
(6) A heating and ventilating plant capable of maintaining the required degree of comfort.
(7) Voltage regulators and controllers.
(8) Instrumentation and controls for monitoring and operating the units and their auxiliaries, including automatic synchronising equipment.

In addition to the generators, the electrical plant housed within or in the neighbourhood of the plant consists of low-and high-voltage switchgear, transformers and generator bus-bars. In connection with the latter it should be noted that, as the generator voltages employed in hydroelectric sets tend to lie in the range 8–15 kV, considerations of the cost of bus-bars dictate that the main transformers must be placed in close proximity to the generators. Thus in underground stations the main transformers are normally housed in galleries running alongside the turbine hall.

On installations on which large pump turbines are employed, pump-starting equipment has to be housed within the stations. This equipment operates on the variable-frequency principle and comprises transformers, converters and inverters, together with the associated switchgear, controls and bus-bar connections. The use of a de-watered pump turbine ensures a soft start prior to synchronising with no significant surge currents.

26.4.4 Turbines

26.4.4.1 General principles

Consideration of hydraulic conditions at the turbine runner leads to the division of water turbines into two main groups: the impulse turbines represented in modern turbine practice mainly by Pelton wheels; and reaction turbines, a group covering both mixed flow and axial flow machines.

Impulse turbines are driven by jets of water issuing from one or more nozzles distributed tangentially around the periphery of the wheel. The power output is controlled by adjusting the opening of the nozzles. As such adjustment does not affect the direction of the jet, the part-load performance of such machines is on a par with that at the design point.

With a few exceptions, reaction turbines are normally equipped with movable guide vanes. These are disposed symmetrically around the runner and control both the velocity of flow and its direction at the entry to the runner. The majority of reaction turbines have runners whose geometry is fixed and invariable. Modern, mixed flow machines, equipped with such runners are known as ‘Francis’ turbines. The flow at inlet is invariably inward and the flow at exit is usually axial. The axial-flow machine with a fixed geometry runner is known as a ‘propeller’. Turbines in which the guides are movable but the runner blades are fixed are said to have ‘single regulation’. Part-load performance of such machines (see *Figure 26.28*) tends to be poor. In order to improve performance, machines with both movable guide and movable runner blades have been introduced. Such machines are said to have ‘double regulation’.

Axial-flow turbines with double regulation have been employed since the 1920s. They are known as ‘Kaplan turbines’, after their inventor. Double regulated mixed-flow turbines have been used since the 1950s. This type of turbine is now designated as either a ‘diagonal’ or a ‘Deriaz’ turbine, after its inventor. It may be of interest to note that the Deriaz turbine was first developed in Rugby, England. Perusal of *Figure 26.28* will show the degree of improvement, of part-load performance, resulting from the use of double regulation, on both axial- and mixed-flow machines. It will be appreciated that double regulation increases both the size of a machine and its costs. However, the resulting improvement in performance under a range of operating conditions can make their use economically justified, either where single or very few machines are installed, or where substantial head variations are encountered.

The range of heads for which the various reaction machines operating as turbines and as pump turbines are suitable is shown in *Figure 26.29* This figure also indicates the outlines and the proportions of the runners of such machines. The shape of runners corresponding to progressively higher heads are shown from left to right.

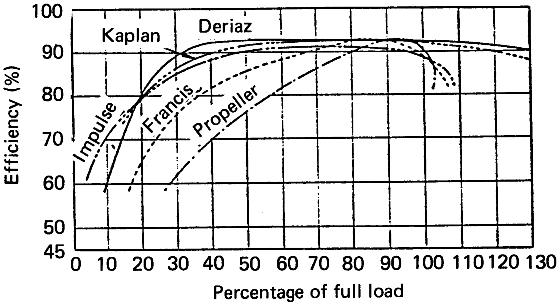


Figure 26.28 Comparative turbine efficiency curves

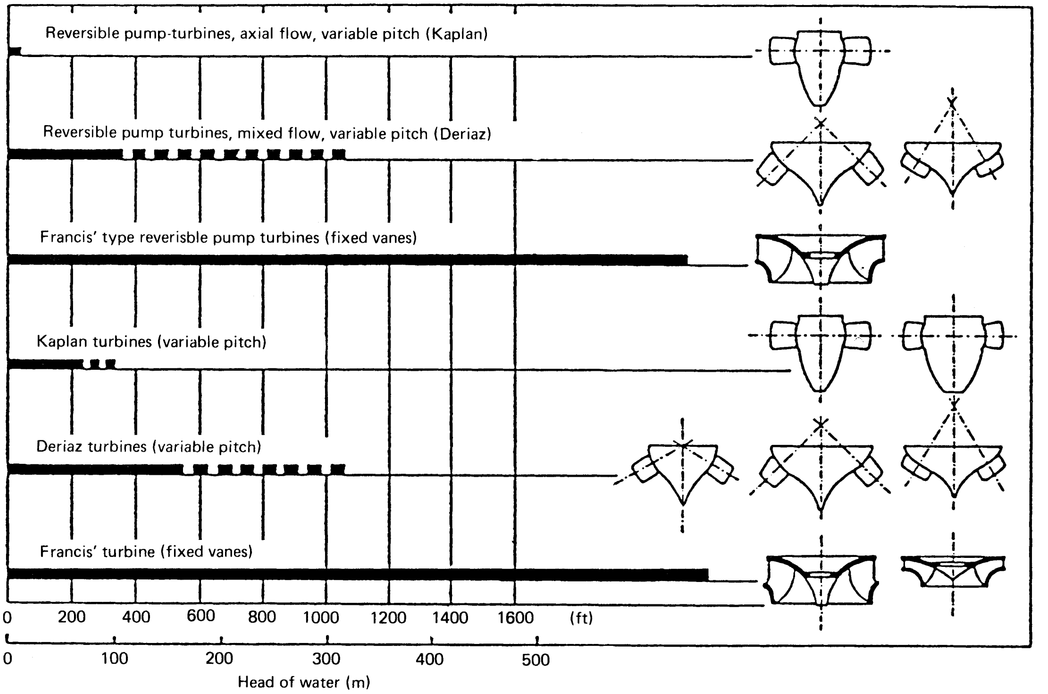


Figure 26.29 Heads under which single-stage pumps and turbines can operate

With the advent of physically large units operating at very low and varying heads, as encountered on some run of river schemes and on tidal applications, yet another variation on the design of an axial flow turbine has been introduced. This type of turbine, known as a 'Kapeller', has fixed guide vanes but the runner is equipped with movable runner blades. Such a combination produces a plateau of relatively high turbine efficiencies, along the line of maximum output for a substantial range of heads.

In an impulse turbine all the available energy is converted into velocity, before the water enters the runner, while in a reaction turbine the process of conversion takes place partly before and partly after the water has entered the runner. The division of water turbines into these two groups is based on general usage and does not imply any difference in the method of energy transfer between the water and the runner. At its simplest the principle of a water turbine is that of a rotating duct, through which flows a stream of water. The stream and the duct interact; the stream is deflected and, as a result, a force is exerted on the duct. The moment of this force, about the axis of rotation of the duct is equal and opposite to the change in the moment of momentum of the stream. If the mass of water, flowing in a unit of time is ζQ, the driving torque M is given by

$$M = \zeta Q(r_1 C_{u_1} - r_2 C_{u_2}) \qquad (26.4)\Leftarrow$$

where C_{u_1} and C_{u_2} are the whirl velocities at the inlet and exit, and r_1 and r_2 are the corresponding radii. Note that only the whirl components of the absolute velocities of flow contribute to the driving torque.

The driving torque can also be determined from considerations of power output from the turbine and of angular velocity ω. Hence it can be shown that

$$\begin{aligned} H_n\eta_t &= (r_1 C_{u_1} - r_2 C_{u_2})\omega/g \\ &= (u_1 C_{u_1} - u_2 C_{u_2})/g \end{aligned} \qquad (26.5)\Leftarrow$$

where u_1 and u_2 are the peripheral velocities of the runners at the inlet and exit, and η_t is the turbine efficiency.

If the effect of the small mechanical and volumetric losses on turbine efficiency is neglected and the concept of effective head $H_e = H_n\eta_t$ is introduced, we get the Euler turbine equation in its canonical form

$$H_e = u_1 C_{u_1}/g - u_2 C_{u_2}/g \qquad (26.6)\Leftarrow$$

This equation demonstrates the interdependence between head, velocities and the geometry of the turbine, at entry and exit from the runner.

It should be noted that, although equations (26.4) to (26.6) are constantly used in the literature to describe the performance of water turbines, they are only strictly applicable to machines in which the flow can be fully described by a single representative stream line.

26.4.4.2 Fundamental similarity

Turbine characteristics are normally established on the basis of tests carried out in hydraulic laboratories on homologous models of the machine and of certain associated water passages, be it distributor pipework and nozzles in the case of a Pelton wheel, or a portion of the inlet pipe, the spiral casing and the draft tube in the case of a reaction machine. For the model to be homologous it must be:

(1) *geometrically similar*—the requirements are that if any representative dimensions defining a water passage are in a given ratio, then all the dimensions of the units must be in the same ratio, i.e. $D_a/D_b = D_{ma}/D_{mb} =$ constant; and

(2) *kinematically similar*—i.e. it is necessary for all the fluid velocities to change in a given ratio when going from one system to a corresponding point of another. This requirement can be expressed mathematically as

$$\text{Flow coefficient} = \phi\psi\!=\!\Leftarrow\frac{\text{velocity of flow}}{\text{rotational velocity}}$$

$$= \frac{V_a}{u_a} =\!\Leftarrow\frac{V_m}{u_m} = \text{constant} \qquad (26.7)\Leftarrow$$

The above relationships can be expressed in the form

$$\phi\psi\!= (Q/D^2)/nD = Q/nD^3 = \text{constant} \qquad (26.8)\Leftarrow$$

where n is the rotational velocity and D is the diameter of the runner. In addition all the fluid velocities must have the same direction at similarly located points. These two conditions can be summarised by stating that all the velocity triangles at all corresponding points must be summarised by stating that all the velocity triangles at all corresponding points must be similar.

(3) *dynamically similar*—i.e. any forces, be they inertia, viscous in origin or those arising out of cavitation phenomena, acting at similar points of fluid should have corresponding magnitudes and should be so arranged that the ensuing flow can satisfy the conditions of kinematic similarity and follow similar paths.

In order to meet these conditions, in machines operating at speeds which are inversely proportional to size it is necessary to ensure that Reynolds numbers for both the model and the prototype are sufficiently high to assure similarity of viscous forces. In addition, the consistency of head coefficients ( ) at corresponding points must be assured. This last requirement can be expressed mathematically as

$$= u^2/2gH_n = \text{constant} \qquad (26.9)\Leftarrow$$

Using the relationships established above, and if small differences in the respective prototype and model efficiencies are disregarded, it is possible to derive a number of parameters in terms of which the performance of a whole family of homologous machines, of various sizes and operating at differing heads can be reduced to just a single diagram, applicable to the whole range of similar turbines. This is achieved by reducing all the data to those appertaining to a turbine with a runner of unit diameter, operating under unit head. These parameters are:

Unit speed $N_{11} = nD/\sqrt{H_n}$ (26.10)⇐

Unit quantity $Q_{11} = Q/D^2\sqrt{H_n}$ (26.11)⇐

Unit torque $M_{11} = M/D^3H_n$ (26.12)⇐

Unit power $P_{11} = P/D^2H_n^{3/2}$ (26.13)⇐

In the case of turbines, the performance is nowadays most commonly expressed in terms of unit speed against unit quantity, with efficiency curves (mussel curves) superimposed on the diagram. Occasionally, cavitation data are also superimposed. In the case of pump turbines the above diagrams would normally be extended to cover both turbine and pump operation. In addition, a four-quadrant plot of unit speed against unit torque is often given.

Typical turbine characteristics for a high-head Francis and a propeller turbine are shown in *Figures 26.30* and *26.31*. *Figure 26.32* shows a four-quadrant characteristic for a reversible pump turbine.

26.4.4.3 Specific speed

In order to extend codification of hydraulic performance from that covering just one family of homologous turbines at a time to all turbines at large, it was found necessary to introduce yet another parameter. Theoretically, this parameter, known as 'specific speed' (N_s), should be obtained by combining unit speed and unit power at the point of best efficiency (b.e.p.) in such a manner as to eliminate the size of the unit from the resulting expression

$$N_s = N_{10}\sqrt{P_{10}} =\!\Leftarrow\frac{n\sqrt{P}}{H_n^{5/4}} \qquad (26.14)$$

However, the published data are more often than not based on the value of this parameter for the rated output rather than the b.e.p.

It should be noted that specific speed is a dimensional quantity and, therefore, the system of units must be specified. Thus in SI units, speed (n) is in revolutions per minute, power (P) is in kilowatts and head (H_n) is in metres. Unfortunately most of the available data are classified in terms of specific speed expressed either in metric or British units. In both systems, the horse power is taken on the reference unit and heads are in metres or feet, as appropriate.

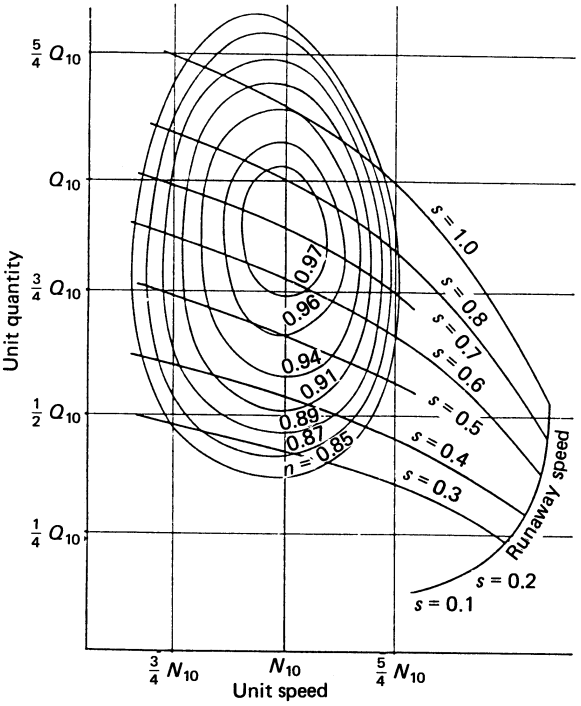


Figure 26.30 Typical high-head Francis characteristic (mussel curves)

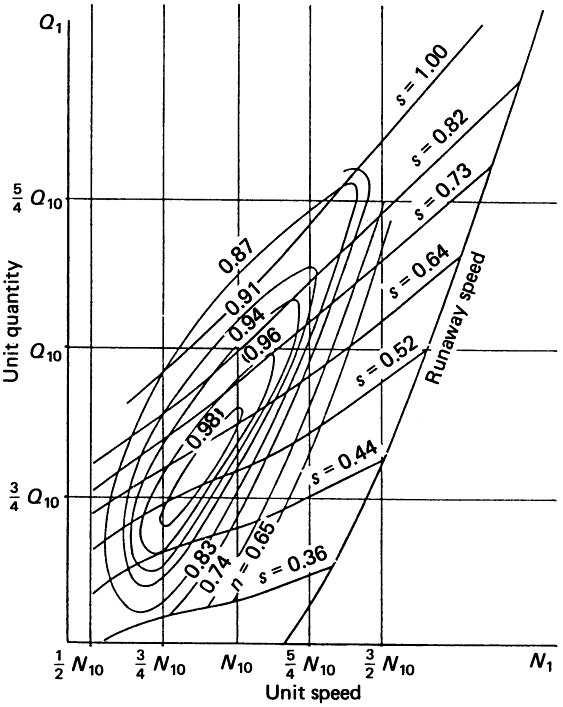


Figure 26.31 Typical propeller characteristic (mussel curves)

Thus, when referring to such data, it should be borne in mind, that

$$1.165(N_s)_{SI} = (N_s)_{metric} = 4.65(N_s)_{British}$$

Whichever system of units is used, the numerical value of the specific speed represents the rotational speed at which a homologous machine would operate if it were designed to give an output of unit power, under unit head, when operating in the same fluid and at the same efficiency.

At this point, it should be mentioned that, while machines having different specific speeds cannot be geometrically similar, it is possible for machines of the same specific speed to be geometrically different, because the same hydraulic conditions can be met by different designs. Thus at certain ranges of specific speed there are overlaps between machine types. The two such ranges of interest concern the overlaps between Francis and multi-jet Pelton wheels at the lowest specific speeds and between Francis and propeller turbines at the highest specific speeds corresponding to the Francis range.

In selecting an appropriate specific speed for large reaction turbines of a particular installation, use may be made of a number of empirical expressions which summarise the current achievements world-wide. These expressions are normally given in the form

$$N_s = A/\sqrt{H_n} \qquad (26.15)\Leftarrow$$

where A is a constant whose magnitude has been steadily increasing with the passage of time, following the adoption of higher specific speeds.

In the case of major reaction machines this constant is now taken as 2400 or so. Some 25 years ago, on Francis units

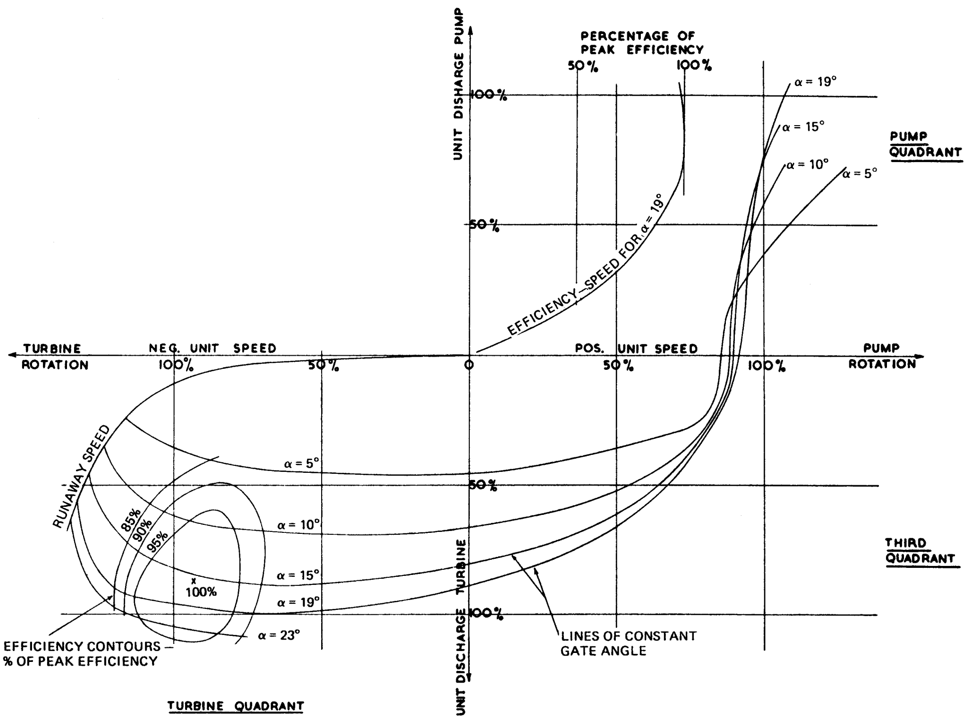


Figure 26.32 Characteristic curve of a Francis reversible pump turbine

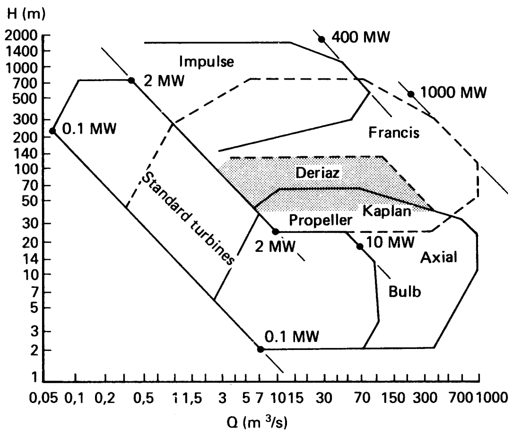


Figure 26.33 Application ranges of water turbines (After Eseher Wyss)

US practice was based on half of this value. This improvement was made possible by better understanding reached through research and testing of the conditions at the turbine runner, where the danger of damage by cavitation represented the major risk. The steady increase in the magnitude of this constant represents a substantial economic benefit because the higher specific speed leads to smaller runner diameters and higher running speeds.

In the case of impulse machines the specific speed is based on a per-jet principle. Thus if a wheel is equipped with four jets its specific speed would be twice that of a wheel with just one jet. The range of specific speeds covered by Pelton wheels equipped with a single jet extends approximately from 10 to 80 rev/min, the lower figure corresponding to the highest available heads.

Because of the overlap between Pelton wheels and Francis turbines, the choice of machine type for a given combination of head and flow becomes too complicated to be expressed by a simple formula. Instead, the reader is referred to *Figure 26.33* which demonstrates the design trends.

For many years specific speed has been used as a general guide to the shape, relative size and speed of machines. It also provides access to information regarding typical efficiencies, flow and head coefficients as well as to values of runaway speeds appropriate to a given class of machinery. The relevant information is summarised in *Figure 26.34*.

26.4.5 Hydrogenerators

Because of the system frequency, the maximum design speeds encountered in practice are not greater than 1000 or 1200 rev/min and salient-pole generators are almost universally employed. Small induction generators of up to 5 MW are used in isolated cases. The two requirements specific to hydroelectric installations are the need to take into consideration the very high runaway speeds and the need to provide sufficient rotary inertia to assure both the quality of speed control of the turbine and the stability of the electrical system. In the case of vertical-shaft units the specification often calls for the bore of the generator stator to be of sufficient size that the turbine runner can be withdrawn without the need to remove the stator. Thrust bearings are normally provided on the generator shaft. It is thus necessary for the generator designer to be acquainted with the magnitudes of hydraulic thrusts appertaining to a range of operational conditions.

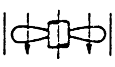
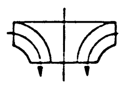
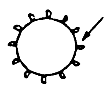
Head (m)	Low 4-50	Medium 20-700	High 200-2000
Storage	Pondage or run-of-river	Storage or pondage	Ample storage
Turbine type	Propeller or Kaplan	Francis	Impulse (Pelton)
Speed (rev/min)	50-250	150-600	200-1200
Runner			
Specific speed (kW-m)	300-1000	60-450	10-80
Peripheral speed of runner / Water speed	2-3	0.75-0.9	0.5
Runaway speed / Normal speed	2-2.2	1.8-2.1	1.8

Figure 26.34 Data for basic types of water turbine

As a result of the restrictions on size and weight due to consideration of transport and access to remote sites, generator rotors and stators are often assembled on site using preformed, transportable components. In the case of major schemes it is often economic to construct a dedicated factory at the site of the scheme, rather than to transport large components and subassemblies to the site.

26.4.6 Economics

The capital costs per unit output of hydroelectric schemes are normally very much higher than comparable costs for thermal stations. This is due to the cost of the extensive civil-engineering works involved and to the very long periods of construction, during which costs are incurred and interest has to be paid, without receipt of any compensating income. However, the civil works are permanent and the life of much of the plant may well be 60 years between major overheads, partly because of its durability and partly because the problem presented by obsolescence hardly arises. Operating costs are very low, comprising only taxes, wages and salaries of the staff and low levels of expenditure arising from the need for maintenance of the plant. Because a very large portion of the lifetime costs is incurred before a scheme is operational the cost of borrowing is one of the major parameters to be considered when assessing the viability of any scheme. As a result, the construction of many power-producing schemes can only be justified by incorporating them within larger schemes producing additional benefits such as irrigation, flood control or navigation.

The cost of the environmental effects of a scheme cannot be easily assessed. Hydroelectric schemes often provide excellent recreational facilities, first-class roads and river crossings. However, they lead to drowning of valleys, interfere with the migration of fish and can lead to the deposition of substantial amounts of silt upstream of dams. Tidal schemes can badly affect the ecology of an estuary, especially during the construction period, and on any future schemes it will be necessary to ensure that at no time will a tidal reach of a river be turned into a temporary sweet water or brackish lake.

Notwithstanding the difficulties arising from the need to assess the ecological consequences of major power schemes, their effects must be seen to have been taken into account in any assessment of the viability of any future scheme. Otherwise, major costs will be incurred by the population at large rather than the polluter.

26.4.7 Pumped storage

26.4.7.1 Historical development

In many countries the available water-power resources are becoming fully utilised and, as environmental considerations lead to resistance to further encroachment by developments leading to the inundation of large tracts of land, modern hydroelectric practice is tending towards construction of pumped storage schemes. Currently some 20% of the world's largest turbines are used in this mode.

The location of such installations is not critically dependent on the quality of the catchment area, but there must be sufficient water to fill and make good any losses due to seepage and evaporation. However, any natural inflow to the upper reservoir produces a bonus in output.

As the cost of machinery per unit output depends on the available head, there is an incentive to construct such schemes at sites providing potential heads in the range 300–600 m. Should such sites not be available, lower heads could still be exploited, but at a greater cost. The pump turbines employed on schemes with heads of up to 600 m would nowadays invariably be of the single stage, reversible type. Where higher heads are encountered, the pump turbines would employ several stages of runners/impellers. Alternatively, the scheme itself can be based on the use of two or more pump turbines operating in series. However, such a solution would involve substantial additional costs.

Wherever possible, the existing storage capacity of conventional hydroelectric plant is used as the base for the pumped-storage plant. In the absence of such facilities, large existing underground excavations may be used as the lower reservoir with the pump turbines installed in a deep shaft, well below the free level of the underground reservoir.

Historically, on pumped-storage schemes separate pumps and turbines were employed with the generator doubling up as a motor in the pumping mode. The four 90 MW units operating at a maximum head of 320 m installated at Ffestiniog, North Wales, are typical of this type of installation. All the machinery is arranged in line and the length of the main shaft is around 30 m. When pumping, the turbine is run in the dry and the thrust load is supported on two separate thrust bearings, one of which is located within the generator and the other under the pump. A coupling is provided between the turbine and the pump. Separate water conduits lead to the turbine and the pump and these are guarded by separate sets of valves (see *Figure 26.35(a)*). In 1960, when the scheme was commissioned, Ffestiniog was the largest pumped-storage project in the world. Even at the time, the design concept was already obsolescent. The next generation of pumped-storage plants depended on the use of reversible pump turbines which were developed at that time. *Figure 26.35(b)* demonstrates the simplification of layout arising out of the use of reversible units.

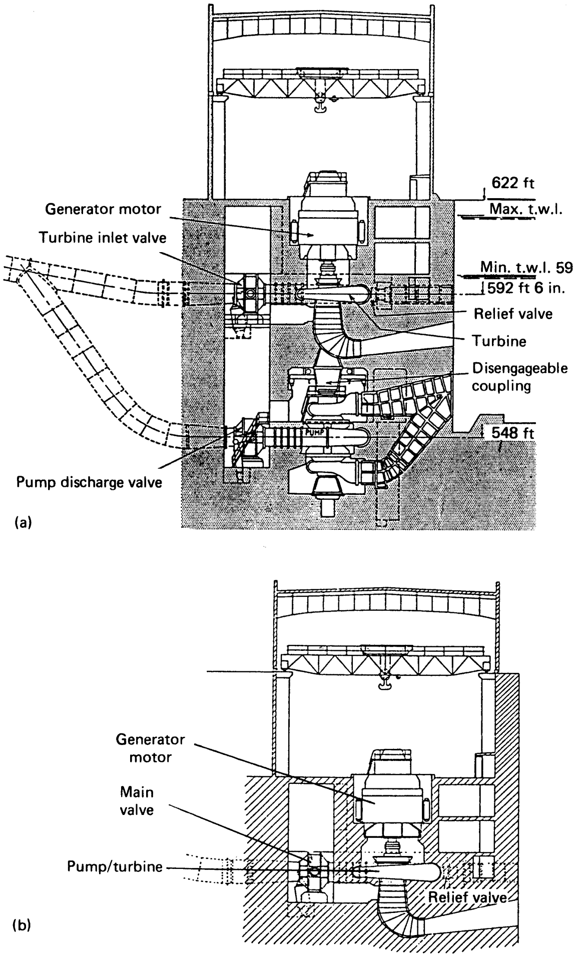


Figure 26.35 (a) Section through the pumped-storage station of the Ffestiniog site, North Wales. (b) Possible arrangement of a reversible installation based on the Ffestiniog site

In Great Britain, because of constraints imposed by the terrain, conventional hydroelectric installations provide a very small fraction of the electrical output. Such schemes as are in existence are invariably small and are mainly located in the north of Scotland. However, there is a substantial investment in pumped storage. In addition to the 360 MW installation at Ffestiniog, there is a four 100 MW turbine installation at Loch Awe, operating at a maximum head of 365 m; a two 150 MW turbine installation at Foyers, Loch Ness, operating at a head of 180 m and the six 300 MW turbine station at Dinorwig, North Wales, operating at a maximum head of 545 m.

At the time of writing, the reversible pump turbines giving the largest output are installed at Bath County, USA, where six 457 MW units operate at a maximum head of 329 m. Next in terms of output come the units at Helms, USA, operating at heads of up to 540 m with an output of 400 MW per unit. Comparable outputs are obtained at Racoon Mountain, USA, where the operating heads are 310 m.

At the heads beyond the range of Francis turbines, where Pelton wheels are employed, application of pumped storage involves the use of multistage centrifugal pumps. Because of the need to ensure that the Pelton wheels are set above tailwater level, while high-head pumps require substantial submergence, the generator can no longer be used to drive the pump.

The Roncovalgrande plant in Hago Delio, Italy, operating at a maximum head of 747 m is equipped with eight Pelton wheels and a single pump. Similar arrangements have also been used at San Fiorano, Italy, where four Pelton wheels and one pump are installed. These operate at heads in excess of 1400 m.

Yet another solution to the problem presented by pumped storage at high head is presented by the use of multistage reversible pump turbines. An example of such an application is provided by the Edolo plant in Italy where eight such machines each rated at 132 MW operate at a head of 1287 m. In connection with the use of such machines it should be mentioned that, while their performance at or about the design point can be excellent, their output cannot be varied other than by natural changes in head. However, with eight units installed, the output of the scheme can of course be varied in stepwise fashion between 12.5 and 100%.

26.4.7.2 Mode of operation

Pumped storage plants provide the means for system regulation, peak lopping capacity and spinning reserve. Where a correct mix of thermal stations and pumped-storage plant has been installed, the system operator can maintain continuous output from base-load plant even at times of low-load demand. At times of high-load demand, the pump turbines provide additional output, and when operated in the spinning reserve mode they are capable of very quick response to system demands. Their use, unlike that of conventional hydroelectric units, does not in the main introduce the utilisation of renewable energy. Their economic and operational benefits arise mainly from a reduction of the high costs incurred in both fuel and in the consequences of physical damage incurred by thermal stations when operated in a cycling mode, such as 'two shifting', so common in the UK and elsewhere. They also ensure that the uneconomic use of gas turbines in their peak-lopping mode of operation is no longer necessary, thus removing the need for the provision of such facilities.

The rapid response capacity of pumped storage plant can be seen from consideration of the operating times, under automatic control, achieved at the Dinorwig power station:

(1)	standstill to no-load generation	90 s;
(2)	no-load generation to 1320 MW station output	10 s;
(3)	full load to no load but still synchronised	10 s;
(4)	pump synchronised to full pump discharge	
	for a single pump	80 s;
	for six pumps	220 s.

Modern reversible pump turbines when starting in the pump mode are nowadays run up to speed and synchronised while the generator that acts as a motor is under the control of static, variable-frequency convertors. These convert the fixed voltage, mains frequency supply into a variable voltage, variable frequency output. Fixed-blade machines are normally started with the runner operating in the dry, in compressed air. Movable-blade runners are normally started with the unit watered, but with the blades fully feathered, to minimise resistance.

Starting a set for generation simply involves, after starting the essential auxiliaries, admitting water to the turbine. Where separate pump and turbine units are installed, like at Ffestiniog, the start of pumping involves: running the set up to speed by the turbine and then synchronising; closing the turbine guide and inlet valves so that the set is motoring; dewatering the turbine casing using compressed air; and opening the pump discharge valve. At Dinorwig changeover times from pumping to turbining under normal procedures take just under 10 min. In an emergency this time can be reduced substantially. It is claimed that at Dinorwig under an emergency procedure the changeover can be achieved in 90 s. The changeover from turbining to pumping is much slower, mainly because the rates of acceleration of the unit in the pump start mode are controlled by the capability of the static, variable-frequency converter. However, as already pointed out, the initial reduction in load, from full to no load, is only a matter of 10 s or so.

26.4.7.3 Economics of pumped storage

The overall energy efficiency of a modern pumped-storage scheme is usually about 75%; the losses arise chiefly from the need for double energy conversion and from loss of water from the upper reservoir due to seepage and evaporation.

The running costs of the station comprise the cost of the energy used in pumping (including the above losses), wages and maintenance, and the costs of transmission losses arising from the remoteness of the station. Pumping energy can be purchased at off-peak times at between one-half to two-thirds of the price at which it can be sold at times of high demand.

Capital costs depend critically on the topography of the site and on the availability of efficient turbo-machinery of sufficiently high specific speed. The latter determines the size of the power station and the cost of generators.

Availability of pumped storage increases the efficiency of operation of the whole interconnected system and, because of its rapid response capability, removes the need for investment in gas-turbine based peak-lopping units and of operation of thermal units in the spinning reserve mode. Given the correct mix of pumped storage and thermal stations, the need for two shifting is largely eliminated.

In Europe, the USA and Japan, many pumped-storage plants have been built. Many more are projected or are in the course of construction. Pumped plants comprise the single most important area of growth of hydroelectric power generation.

26.5 Diesel-engine plant

Generators powered by diesel engines are employed in three main roles:

(1) on primary or base-load duty in locations where there is no utility supply or as an independent power source to ensure security of supply where a public supply system is available;
(2) for peak-lopping (or peak-shaving) duty to supplement and/or reduce the cost of supply from a utility source; and
(3) as stand-by to a power supply from a utility.

The speed of crankshaft rotation basically determines the weight, size and cost of an engine in relation to its output power. Engines are generally accepted as being divided into three classes:

High speed	>1000 rev/min,
Medium speed	400–1000 rev/min,
Low speed	<400 rev/min.

The maximum size of diesel plant for primary power generation is for all practical purposes between 150 and 200 MW per station. Output ratings available are in unit sizes from 1 kW to 30 MW. The most significant range for generating plant in all three utilisation categories (continuous, peaking and stand-by) lies between 250 kW and 3.5 MW unit sizes, in the medium- and high-speed classes.

The choice between low, medium or high speed engines must be related to evaluation of power supply security against operational economy. Security of supply is essentially a function of the availability of engines and the number of units and spare capacity installed in relation to the average load demand.

26.5.1 Theory and general principles

26.5.1.1 Working cycles

The compression ignition engine operating on liquid fuels, or in the dual-fuel mode, works on the principle of fuel being injected into a charge of compressed air and spontaneously ignited by the high temperature of the induced air by the heat of compression. The process converts the heat energy of the fuel into mechanical work.

The two basic working cycles are four-stroke and two-stroke. These are represented diagrammatically in *Figures 26.36* and *26.37* together with the appropriate indicator diagrams, which portray the events within the engine cylinder during each cycle.

In the two-stroke engine the working stroke occurs in each revolution of the crankshaft, whereas in the four-stroke engine it occurs once in every two revolutions. It does not follow that, because the two-stroke engine has twice as many power strokes as the four-cycle engine it will produce twice the power. The down stroke of the two-cycle engine (*Figure 26.37*) combines both power and exhaust strokes. As the intake and exhaust ports are cleared by the piston some mixing of fresh air charge and burned gases takes place (scavenging). Not all the burned gases are exhausted, which prevents a larger fresh charge of air being induced into the cylinder. The resulting power stroke has, therefore, less thrust.

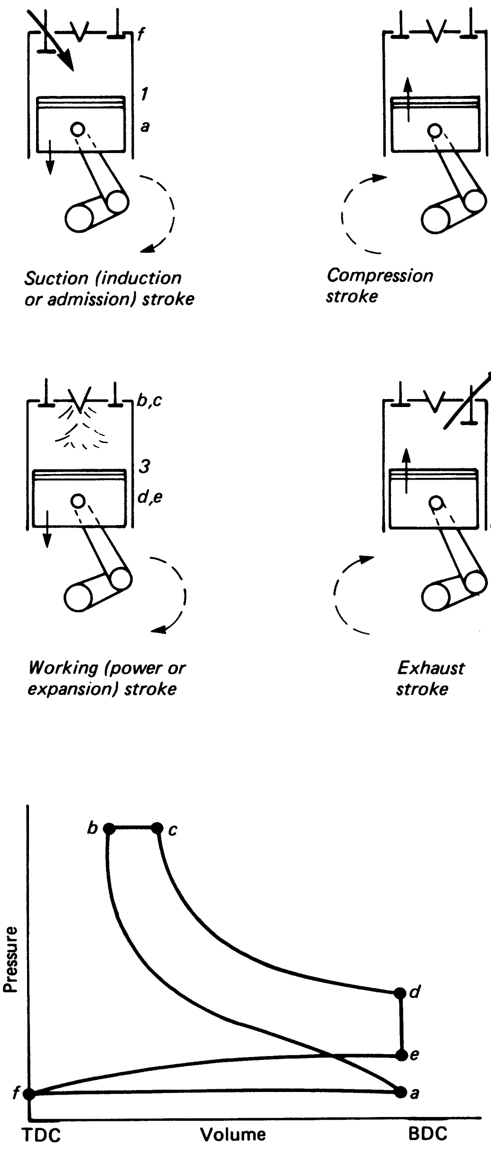
In the four-stroke engine, however, nearly all the burned gases are forced out of the combustion chamber by the up-stroking piston. This allows almost a full air/fuel mixture to enter the cylinder since a complete piston stroke is devoted to induction of the mixture. The power stroke therefore produces relatively more power than its two-cycle counterpart.

26.5.1.2 Combustion

The major advantage of the reciprocating internal combustion engine is that its design is not limited by the properties of the materials of its construction, since none of its parts is required to work continuously at maximum-cycle temperature. This allows high maximum-cycle temperatures to be used, which results in a high thermal efficiency—this is a measure of the efficiency with which the fuel is burned during the combustion process to produce engine power.

Whilst compression-ignition engines are generally about 5% more efficient than their prime-mover competitors, there is appreciable variation amongst them in thermal efficiency. Much depends on the size of engine and the type of combustion chamber.

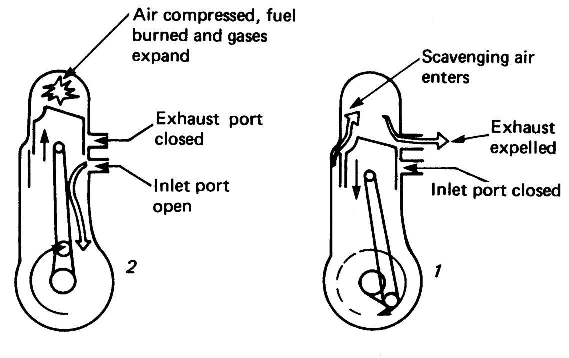
Combustion chambers are basically of two types: those designed for indirect injection and those for direct injection. The former employ pre-combustion chambers in the cylinder head into which a relatively coarse fuel spray is injected at low pressure. They are popular with European and American engine manufacturers and have the advantage of being able successfully to handle a wide range of fuels. When required to operate over a wide band of environmental conditions, however, they compare unfavourably with direct-injection chambers on fuel-consumption performance. Also, since heat loss from the pre-combustion chamber is high, cold starting can be difficult without prolonged cranking or recourse to external heating (such as glow plugs).


f
1
a
b
2
a
Suction (induction or admission) stroke
Compression stroke
b,c
3
d,e
f
4
e
Working (power or expansion) stroke
Exhaust stroke
b
c
Pressure
d
e
f
a
TDC
Volume
BDC

f–a	Induction of fresh charge	Stroke *1*
a–b	Compression	Stroke *2*
b–c	Combustion	Stroke *3*
c–d	Expansion	
d–e	Blowdown	
e–f	Exhaust	Stroke *4*

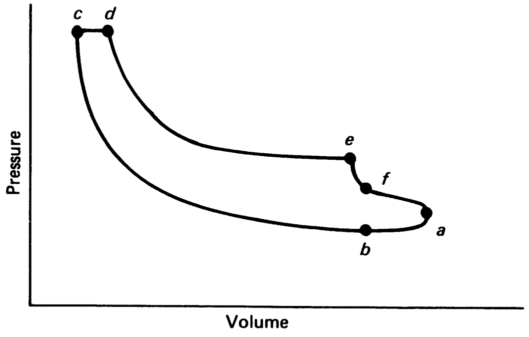
Figure 26.36 Four-stroke cycle

In direct-injection systems the underside of the cylinder head is usually flat and clearance volume on compression is mainly contained within the piston crown. Crown depressions are so shaped as to effectively induce swirled air



Charging and combustion stroke

Scavenging and induction stroke



a–b	Scavening and induction	Stroke *1*
b–c	Compression	
c–d	Combustion	Stroke *2*
d–e	Expansion	
e–f	Exhaust blowdown	
f–a	Exhaust and scavenging	

Figure 26.37 Two-stroke cycle (valveless form)

turbulence, as the piston rises on its compression stroke. Fuel is then injected in the same direction as this flow of swirling air. The direct-injection principle is almost universally employed in modern medium-speed and on many high-speed engines.

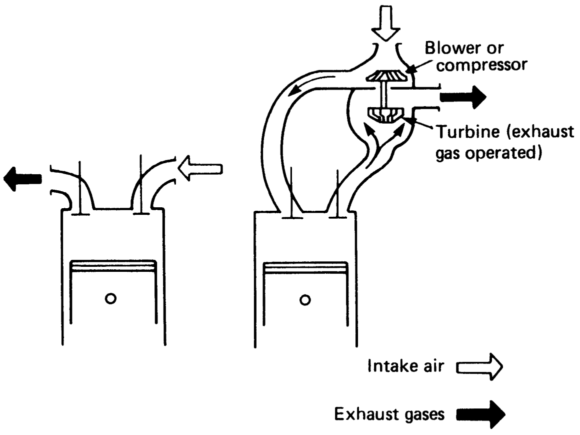
Small-bore engines tend to have lower thermal efficiencies because their high surface-area/cylinder volume ratios give larger heat losses. Again, the greater heat losses from the larger exposed surfaces of indirect-injection systems means that they give lower thermal efficiencies than direct-injection ones. For these reasons, small indirect-injection engines may have thermal efficiencies as low as 28% whilst larger engines, particularly those using direct-injection techniques, may have efficiencies as high as 40%.

26.5.1.3 Pressure charging and inter-cooling

In the naturally aspirated four-stroke engine the working cylinder is almost (but not fully) charged with fresh air at atmospheric temperature and pressure at the end of the suction stroke. The density of this aspirated air regulates the weight of fuel which can be burned during the working stroke and this in turn determines the maximum power that can be developed. If a compressor were to be employed to supply the engine with intake air at a pressure higher than atmospheric, the mean effective pressure (and, therefore, the power output) of the engine would be increased without altering crankshaft speed or cylinder volume. This is effectively what pressure charging does. It can increase output by as much as 50% over an equivalent naturally aspirated engine of similar speed and dimensions. Furthermore, appreciable reductions are achieved at all loads in specific fuel consumption rates and the less arduous working conditions at the cylinders give increased engine reliability and reduced maintenance. On the debit side, however, it is impractical to expect a pressure charged engine to accept more than about 85% of its full load capability in one step in less than 10 s from crank initiation.

Several types of compressor (driven by chain or gearing from the crankshaft) are available, but since they deprive the engine of a portion of its shaft output, they are not as economical as the turbocharger which utilises the otherwise wasted energy of the engine's exhaust gases. The turbocharger very simply consists of a gas turbine, driven by the exhaust gas flow, mounted on a common spindle with a blower or compressor placed in the air intake path. *Figure 26.38* illustrates in schematic form the application of the turbocharger to a four-stroke engine. Further engine upratings are now being achieved in some medium-speed, four-stroke engines by employing two-stage turbocharging to give higher air intake densities.

The full potential of this increase in air inlet density by pressure charging is, however, marginally offset by an increase of air temperature due to adiabatic compression in the turboblower. This loss is recoverable by the use of charge air coolers (inter-coolers) placed downstream of the turboblower, which have the effect of increasing the fuel/air ratio, allowing more fuel to be injected into the cylinder and so raising the engine's power output. The lower air intake temperature has the further effect of reducing not only the maximum cylinder pressure but also the exhaust temperature, and with it the engine's thermal loading. Increase in engine power over a straight turbocharged model is usually of the order of 20–25% and thermal efficiencies of over 40% are obtainable.



Naturally aspirated

Turbocharged

Figure 26.38 Four-stroke engine with exhaust-gas-operated turbocharger

Air-to-air charge cooling is usually carried out in a section of the engine's radiator. An alternative water-cooled arrangement uses either a separate radiator circuit for charge-air cooling water or, in marine installations, sea-water-cooled heat exchangers with finned tubes carrying the sea-water and over which the charge air passes.

26.5.2 Engine features

26.5.2.1 Basic classification

In generator drives, the compression-ignition engine is usually a multi-cylinder unit classified by the synchronous speed required, the type of fuel to be used, and the mechanical arrangement (i.e. the geometric arrangement of the cylinders).

26.5.2.2 Synchronous speed

Consistent with the output power required, the operating speed N (revolutions per minute) is determined by the frequency f (hertz) and the number p of generator pole-pairs in accordance with the relation $N = 60f/p$. Thus a two-pole generator must be driven at 3000 rev/min for 50 Hz or 3600 rev/min for 60 Hz; and for an eight-pole generator 750 or 900 rev/min.

26.5.2.3 Fuels and operating modes

(1) In its compression-ignition form the internal combustion (i.c.) engine may be run on liquid fuels using either distillate or light or heavy residual oils.

(2) As a high compression unit in its compression-ignition form, the i.c. engine may operate in a dual-fuel mode using a mixture of gas and air ignited in the cylinders by the injection of a small pilot charge of liquid distillate fuel. The pilot fuel consumption is between 5 and 10% of the normal full load quantity required for straight diesel operation and it remains fairly constant throughout the load range unless there is a gas shortage, in which case any input energy shortfall is made up by an increase in liquid fuel injected. Should the gas supply at any time fail or become inadequate for the load demanded, the engine automatically reverts to the straight diesel mode. Because of the need to modulate both oil and gas flows, the control and protection system needed is more complex than for straight diesel or gas engines.

The engine may be switched at any load from dual-fuel to diesel operation, and vice versa. The same output rating must therefore be selected for both diesel and gas operation.

(3) As a spark ignition unit the i.c. engine may operate on gaseous fuels such as natural gas, propane or sewage gas.

(4) Finally, in an alternative-fuel form, an i.c. engine may be designed to incorporate both fuel-injection and sparkignition systems to give operation on either liquid or gaseous fuels. More often than not change to either type of fuel requires engine stoppage, but link mechanisms can be fitted to disconnect the fuel pump drive and close the air intake whilst simultaneously energising the electrical ignition and turning on the gas input. Changeover may then be effected with the engine running, much as in the dual-fuel mode described above. It must be appreciated that output rating will vary with the type of fuel used; it will be lower for gas operation owing to the change in compression ratio.

26.5.2.4 Mechanical arrangements

Perhaps the most widely applied engine format is the vertical design in which the cylinder axes are perpendicular and in-line. In the medium-speed range vee-form engines offering high power in a small bulk are attractive, particularly for trailer-mounted and transportable generator plants. The included angle of the vee may range from about 35° to 90°. Other arrangements occasionally used are horizontal designs, opposed piston engines of various forms and vertical twin crankshaft or double-bank engines.

26.5.3 Engine primary systems

26.5.3.1 Fuel injection

Most modern compression ignition engines have a mechanical or airless fuel injection system embodying jerk-pumps. Medium-speed engines tend to use individual camshaft-actuated pumps for each cylinder. Higher-speed engines employ block pumps within which all the jerk-pump elements are incorporated and driven from a self-contained camshaft which, in turn, is coupled to an auxiliary drive from the engine. Certain two-stroke engines use the common rail system, wherein fuel is maintained at a constant pressure by a pump and a hydraulic accumulator. A fuel valve at each cylinder, driven and timed from the main camshaft, delivers the fuel to the engine.

The injector, the device that introduces the fuel by spray into the combustion space, is in essence a spring-loaded needle valve, whose tip covers the injector nozzle hole(s). The number of holes, their angles and the angles of spray are largely dependent upon the shape of the combustion chamber. Since fuel systems must have small passages and nozzle holes, it is of paramount importance that the fuel within these systems is well filtered.

26.5.3.2 Lubrication

The lubricant in an engine performs many tasks. In addition to reducing friction (and the potentially considerable power loss due to it) and minimising wear, its purpose is to provide:

(1) cooling (either under-crown and/or in piston ring areas);
(2) cleaning and flushing of impurities; and
(3) absorption of shocks and impacts between bearings and other engine parts.

It also affords a seal between piston rings and cylinder walls to reduce the seepage of gas that passes between the piston and cylinder walls from the combustion chamber into the crankcase.

Most engines use a pressurised or force-fed system to circulate the lubricant from an external drain tank or from a sump in the base of the crankcase. The main components of any system are: the circulating pump, which may be either of the gear or the multi-lobed rotor type; a pressure relief valve; oil filter(s) and an oil-to-engine coolant heat exchanger fitted between the feed pump and the filters. Delivery pressure is normally in the range of 50–200 kPa (0.5–2 kg/cm^2) but it may even be as much as 400 kPa in high-speed engines.

In the so-called dry-sump system two pumps are employed. Oil that has circulated through the engine oilways returns by gravity to the crankcase pan. The task of the first pump is to transfer this oil from the pan to a reservoir tank external to the engine. The second pump draws oil from this tank and delivers it via the heat exchanger and filters to the bearings, etc.

The arrangement, mainly used on high-speed engines, in which the crankcase oil pan is in itself the reservoir is known as a wet-sump system.

Probably the most severe cylinder wear conditions occur just after an engine has been started, when piston lubrication is at its poorest. For this reason pre-priming systems are incorporated into both the manual start procedures and automatic start controls of low-speed and higher-rated medium-speed engines. It is also advisable to use periodic priming of lubricant on the larger high-speed engines, when these are operating in the automatic stand-by mode. Oil fed from a separate electric-motor-driven pump is circulated at periodic intervals through the oilways to flush the liners and generally wet the engine moving parts in readiness for an automatic start. Periodic priming also has the advantage of reducing the severe wear that can occur on cylinders due to condensed combustion products when an engine is at standstill.

26.5.3.3 Cooling

Engines may be either air cooled, using a mixture of air and oil as cooling medium, or water cooled with water and oil as the cooling media. Air cooling is simpler and is satisfactorily applied to high-speed engines up to approximately 400 kW output. Air is drawn into an impeller (usually secured to the engine flywheel or vee-belt—driven off the crankshaft) and discharged through shrouding across the finned external surfaces of the cylinder and cylinder heads. Whilst very low output engines may not require separate oil-to-air lubricating oil heat exchangers, they are standard features on larger engines.

Good installation of air-cooled engines is critical especially in high ambient conditions and in confined spaces. Care must be taken with the design and application of air intake and hot air outlet trunking to avoid, in particular, the possibility of hot discharged air being recirculated within the generator housing.

Water-cooling by circulation of water through cylinder jackets is the cooling method most frequently applied to engines in generating plants. Detailed arrangements vary considerably but most installations employ some form of closed-circuit jacket cooling system to transfer the engine heat to a heat exchanger. This may be a fan-cooled radiator or a shell-and-tube type heat exchanger. Thermostatic elements are incorporated within the systems to bypass the heat exchanger when starting from cold so as to allow the engine to attain its operating temperature more quickly.

Radiators may either be mounted on the same baseframe as their engine-generator assemblies, or remote mounted. Set-mounted radiators usually have engine-driven cooling fans, whilst remote units incorporate single or multiple fans driven by electric motors. At ratings below 1 MW, single or double sectioned radiators are employed. On larger engines multi-sectioned units are used to provide separate circuits for jacket water, lubricating oil and charge air cooling.

Since most modern medium- and high-speed engines are designed for high-temperature coolant conditions they employ pressurised closed circuit cooling systems. System pressures vary between 30 and 70 kPa (0.3–0.7 kg/cm^2) in practice. They equate to system boiling points at sea level of 107°C and 115°C, respectively, so that engine makers may design for corresponding maximum operating temperatures at engine outlet of 80°C and 90°C.

Where a plentiful and cheap raw water supply of good quality is available at less than 30°C, it may be used as a secondary fluid to circulate through an engine's tubular heat exchangers before it is run to waste. Pre-knowledge of the character and quality of the water is necessary to ensure that correct selection of materials for the tubes and shells of the heat exchangers is made.

Where a raw water supply is of dubious quality, clean (or treated) water may be used in a secondary closed circuit, which is in its turn cooled by the raw water in, for example, a cooling tower. *Figure 26.39* illustrates diagrammatically an arrangement for a system of this kind. Whatever the system, it is advisable to use a high rate of jacket circulation with a small temperature difference between engine inlet and outlet, rather than a slow circulation and a large temperature rise.

26.5.3.4 Induction

Every engine should have air-intake filtration/silencing equipment. The induction system must be designed to supply clean dry air to the engine at as near ambient temperature as is possible and with the minimum of restriction. Engine makers stipulate the maximum permissible restriction at induction manifold or turbocharge inlet. The quality and quantity of the air supply has a direct bearing upon the engine output, fuel consumption and life.

Choice of filter depends upon the plant environment and the service life required. The following suggest the duty category in which the various filter types may be applied. The engine maker's recommendations should be sought and complied with.

(1) For plant installed in sheltered and low dust concentration conditions: oil bath or paper element (dry) type filters, both types without pre-cleaner stages.
(2) For installations in temperate, relatively dry and moderately dusty conditions: oil bath or dry element filters with centrifugal pre-cleaner stages and with greater dust holding capacity than those in (1).
(3) For severe dust concentration applications or where regular maintenance is not always possible: heavy-duty paper element filters with highly efficient (90–95%)

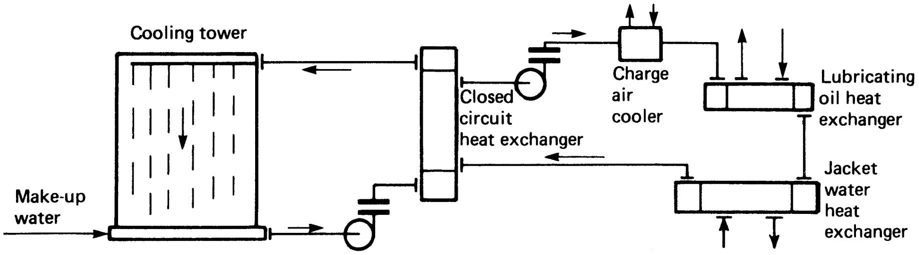

Cooling tower
Make-up water
Closed circuit heat exchanger
Charge air cooler
Lubricating oil heat exchanger
Jacket water heat exchanger

Figure 26.39 A typical closed-circuit secondary cooling circuit

centrifugal pre-cleaner stages and preferably with self-emptying or dust unloading arrangements. Filters of this type usually offer safety elements as an optional feature. Their purpose is to protect the engine in the event of main element perforation or act as temporary substitutes when the main elements are being serviced.

It is good practice to fit air restriction indicators to dry type filters to warn when the elements have reached a pre-set limit of fouling.

26.5.4 Engine ancillaries

26.5.4.1 Starting equipment

The two main energy sources for engine starting are batteries for electric-start systems and air receivers for air-start systems. On small high-speed engines hydraulic energy or spring type inertia systems are occasionally employed. The simplest method of starting is of course manual cranking, but it is practicable only with the smallest engines. On the very largest low-speed engines small pony i.c. engines are sometimes employed to give crankshaft rotation at starting through a clutched pinion engaging with the engine flywheel gear ring.

Most electric-starting systems use motors fitted with Bendix-type pinions on their armature shafts to mesh with a toothed rim on the engine flywheel. Lower output engines in the high-speed range may require only one such motor whereas larger engines in that range, and medium-speed engines, need two motors with a common synchronising control.

Starter motors are either of the axial or co-axial type. In the former the complete armature assembly and pinion move forward axially to engage with the flywheel teeth. On the latter type only the pinion dome moves forward to engage under reduced power. This minimises engagement shock and reduces wear on the gear teeth. Starter windings may be of the 'hold-on' or 'non-hold-on' type. The non-hold variety is usually applied to motors for remotely or automatically started engines. The most frequently used voltages are 12 and 24 V, but 6 V for small engines and 32, 48 or 64 V for the top end of the electrically started range of engines are not unusual.

A cost-effective alternative to starter motors may be applied on generating sets below 15 kW rating, making use of a special d.c. starting winding within the generator to motor the engine. The same winding is used for charging the starter battery when the engine is running.

Starter batteries are either of the heavy-duty lead–acid type or alkaline type. The latter, in its nickel–cadmium form, has been favoured for stand-by generators because it retains its charge better over longer periods without use. Disadvantages are: that it is bulkier than the lead–acid battery of similar capacity; it is more expensive; and it tends to have a larger terminal voltage drop with heavy starting current drain. Maintenance-free, sealed storage batteries, using gas recombination technology, have been increasingly applied on stand-by plants in the last 5 years.

As a typical illustration the capacity of a lead–acid battery required to give six consecutive 20 s cranking periods with 5 s rest periods between each at ambient temperatures down to $-7°$C on an eight-cylinder vee-form engine rated at 600 kW is 236 A-h at the 5 h rate for a 25% discharged condition. At this temperature the steady cranking current demand at 80 rev/min is of the order of 1 kA corresponding to an engine cranking torque of 1 kN-m. Breakaway current and torque figures are as much as 150% above these values.

Starting difficulties, particularly in low ambient temperatures, are more usual on the smaller high-speed engines because their large surface area-to-cylinder volume ratios tend to dissipate the heat of compression. Moreover, restrictions on the size and weight of starting equipment limit the amount of starting torque available. Various proprietary starting aids are available: devices such as electric glow plugs or those using ether–air mixtures pumped into the air inlet manifolds to give access to the engine combustion chambers to promote combustion. Decompression mechanisms may also be used to hold off either the inlet or the exhaust valve on each cylinder, during the initial starting period. Once the engine is up to steady cranking speed, full compression is restored. The engine should then fire and run up to self-sustaining speed.

Compressed air for engine starting may be expanded either in an air motor engaging with the flywheel or directly within the engine cylinders, to move the pistons downwards on their working strokes until firing occurs. As in the electric starter the drive of the air motor is through a sliding pinion. Air motors may be applied to the range of engines that otherwise uses electric starter motors.

Direct air starting applies to the larger medium-speed engines and to low-speed units. Air is directed to each of the engine cylinders, in their proper firing sequence, through non-return valves either from a camshaft driven distributor or through mechanically operated valves.

Compressed air for either form of starting is stored in one or more receivers at pressures between 100 and 300 kPa (1–3 kg/cm^2). Air charge is maintained by a small single-stage auxiliary compressor driven by an electric motor, an i.c. engine or by the main engine itself. On large or critical installations it is usual to provide back-up auxiliaries; the primary compressor being perhaps electric-motor-driven with an i.c. engine-driven stand-by unit.

26.5.4.2 Governors

Diesel generators use variable speed governors set to operate at the predetermined synchronous speed. Choice of governor is dictated by:

(1) The engine type and its application. For example: independent operation feeding an isolated load or parallel operation with similar generators or with a utility supply.
(2) The standard of governing required, i.e. defining the limits for speed regulation (speed droop), steady-state stability and dynamic behaviour. (Classes of governing accuracy and their parameters are defined in BSS 5514: Part 4 and its corresponding International Standard ISO 3046: Part 4.)
(3) The available inertia or flywheel effect of the combined engine and generator. It may be possible to employ a relatively simple mechanical governor in conjunction with a higher inertia flywheel to meet a tight governing specification in an economical manner without recourse to a more sophisticated and expensive governing system. Engine makers will calculate the minimum generating set inertia required for each eligible type of governor to fulfil the requirements of any governing system.

Governors vary from the simple all-speed mechanical type through various forms of mechanical-hydraulic and electro-hydraulic types to all-electric or electronic types.

The mechanical types use rotating flyweights to measure engine speed. The weights move radially and assume a position related to the speed of the engine. In the straight mechanical governor this position is directly translated to the fuel pump rack. Speed setting is usually fixed (or adjustable through only a very narrow speed range) and speed droop is non-adjustable. Pre-selection of speeder springs provides a choice of droop settings—between 4 and 12%. Most high-speed diesel generators up to 1 MW capacity use block type fuel pumps with an all-speed mechanical governor fitted to one end of the injector pump housing. Mechanical–hydraulic governors amplify flyweight movement through a hydraulic servo system to the fuel rack(s). Droop is readily adjustable from 0 to about 8%.

In the electronic governor, engine speed is measured through a magnetic pick-up usually mounted in the flywheel housing to detect the gear teeth on the flywheel rim. The signal so derived is compared with a potentiometer-set speed reference. Any detected difference is amplified within a control unit to adjust the signal to an hydraulic or electrical actuator fitted to the engine fuel rack to correct the fuel and return the engine to its preset speed. Response to speed changes is much faster than with the mechanical–hydraulic types and fully isochronous load sharing is possible on paralleled generators with these types of governor. Moreover, they may be used with a wide selection of control modules to provide fully automated and integrated multi-generator installations.

Where a governor is independently mounted from fuel injector pump(s) it is critical that the engine be fitted with some form of overspeed shutdown device, either acting directly on the fuel pump rack(s) or completely cutting off the intake air flow to the engine. This is necessary to prevent over-fuelling of the engine should any casual jamming of racks and fuel control levers take place or should any of the elements within the governing system fail.

26.5.4.3 Engine monitoring

It is essential that the strategic temperature, pressure, speed and flow parameters of an engine system are regularly monitored to interpret its behaviour and performance and relay this information to planned maintenance operations.

Maker's instruction books give a good indication of the degree of instrumentation required. Much also depends upon the build specification for the particular engine but parameters such as those listed below may be monitored by instruments, wherever applicable.

(1) Jacket and raw water temperatures, pressures and flows.
(2) Lubricating oil temperature and pressure.
(3) Differential pressures across fuel and lubricating oil filters.
(4) Jacket water temperature at outlets from individual cylinders.
(5) Exhaust temperatures at individual cylinders and before or after turbochargers.
(6) Charge air temperatures and pressures.
(7) Starting air pressure.
(8) Engine, pumps, compressors, and other relevant speeds.
(9) Fuel temperature and pressure.
(10) Fuel and lubricating oil tank levels.
(11) Mechanically driven or transducer operated cylinder pressure indicators.

Abnormal operating conditions may be detected by sensors whose output signals are fed into alarm/shutdown logic controls. Various combinations of indicative and protective action are possible:

(1) two-stage alarm and shutdown;
(2) simultaneous alarm and shutdown; and
(3) alarm only (visible and audible).

The conditions to be covered may include:

(1) high jacket and raw water temperatures;
(2) low lubricating oil pressure;
(3) high exhaust temperature before turbocharger;
(4) high differential pressure across both fuel and lubricating oil filters;
(5) high charge air temperature;
(6) excessive vibration; and
(7) engine overspeed.

This list is by no means exhaustive and the engine maker's advice should be sought on the extent of protective insurance to be taken.

26.5.5 A.c. generators

Generator technology is dealt with in Chapter 20. Engine-driven generators for outputs of 20 kVA and above have direct-compound or brushless excitation.

Compounded generators use the load current to provide part of the excitation. They have good overload capacity and rapid voltage recovery, and may be preferred for marine application where large induction motors have to be started.

Brushless excitation is common for diesel-generator plant. The exciter is a three-phase machine connected to a shaft-mounted rectifier diode assembly the output from which is fed to the main generator field. The brushless machine may be self-excited from its output terminals through a solid-state automatic voltage regulator (a.v.r.). The d.c. output from the a.v.r. feeds the stator field of the a.c. exciter to control the output voltage of the main generator. Alternatively, the exciter may obtain its field supply from a shaft-mounted permanent-magnet pilot exciter in combination with an a.v.r.

For medium- and low-speed generators the rotor is usually of laminated form with fully interconnected damper windings. In some high-speed sets the poles have solid bolted-on shoes, which provide eddy-current damping.

26.5.5.1 Construction

Types of construction and mounting arrangements are designated in BS 4999: Part 107 (IEC 34-7).

Almost all diesel-generator sets employ direct coupling of the prime-mover and generator. The latter may be treated either as a separate machine flexibly or solidly coupled to the engine, or flange-mounted to the flywheel housing and close coupled.

Most two-, four- and six-pole machines incorporate end-shield-mounted, grease-lubricated ball and roller bearings. Above about 3 MW, plain bearings are sometimes employed, self- or oil-scavenge lubricated. The large slow running machines tend to use single or dual pedestal outer bearings, mounted on a common baseframe with the generator stator casing.

26.5.5.2 Protection

Where generators are housed in buildings or canopies and enclosures, protection to IP23 of BS 4999: Part 105 is sufficient. (The technically equivalent standard is IEC 34-5.) Ventilation is provided by an internal shaft-mounted fan drawing air from the non-driving end and discharging it at the driving end. Dust filters may be fitted to this type of enclosure when the inlet air contains fine dust, sand, moisture or oil vapour. An output power reduction factor of about 0.95 should be applied to compensate for the resulting restricted cooling air flow. It is advisable to use thermistor temperature-sensing probes, either embedded in the stator windings or placed in the cooling air flow, to guard against damage by overheating, if the filters are not cleaned at regular intervals. For operation in extremely dirty conditions and on outdoor sites such as quarries, a totally enclosed construction, using closed-air-circuit ventilation, should be specified. The closed-circuit air is directly cooled by air-to-air (TECACA) or air-to-water (TECACW) heat exchangers, which may be attached to the generator casing. These enclosures are defined as IP45.

26.5.5.3 Voltage

In the UK the preferred output voltages for three-phase 50 Hz supply, and the appropriate ratings of engine-driven generators, are:

Voltage (kV)	0.415	3.3	6.6	11
Rating (MVA)	≤1.5	0.5–6	0.8–10	1–20

Voltages of 2.4, 4.2, 6.9, 13.2, 13.8 and 15 kV are encountered in American and European practice.

BS 4999 requires the output voltage to be held to within ±5%, but most a.v.r.s will control to $\pm 2\frac{1}{2}\%$ over the load range. Regulation down to ±1% is possible with solid-state devices in closed-loop control systems.

Standard classification for insulating materials is given in Chapter 7. Most machines up to 3 MVA employ class E or F insulation, or combinations thereof, for stator and rotor windings.

26.5.5.4 Parallel operation

Active power load-sharing is a function of the engine and its governing system. The sharing of reactive power is determined by excitation and synchronous impedance: proper sharing is obtained by applying quadrature current compensation (q.c.c.) to give automatic controlled droop of the output-voltage/reactive-current relation. The amount of droop should be the same for each generator paralleled into a power system: it is typically about 5%.

26.5.5.5 Short-circuit performance

Fast-response brushless and compounded machines can have subtransient short-circuit current levels of 6 to 10 times full-load current. But where excitation is derived from the output voltage, a short-circuit removes the excitation supply and the generator voltage collapses.

In distribution networks it is essential that an adequate 'permanent' short-circuit current is maintained by the energy source in order to allow discriminative operation between protective devices such as circuit-breakers, fuses and overcurrent releases. This is necessary if faults on final circuits are to be cleared as quickly as possible. Maintenance of short-circuit current at a level 2 to 3 times full-load current may be achieved by excitation power derived from current transformers in series with the generator output leads. This power is fed to the exciter field through a relay which closes when the output voltage collapses.

26.5.5.6 Generator selection

In selecting the right size of generator for a specific application the following factors, all influencing the rating, should be considered.

(1) *Application*: the mode of operation, e.g. stand-by or continuous; single running or paralleled with a utility supply or with similar generators; load power factors.
(2) *Location*: altitude, ambient temperature, humidity and other environmental conditions.
(3) *Dynamic loading*: any limitations imposed on voltage transient performance in the starting of large induction motors or on waveform characteristics where static converter equipments or thyristor drives form part of the load.

26.5.6 Switchgear and controls

26.5.6.1 Planning

Factors affecting the choice and design of switchgear will be:

(1) The size and nature of initial and future loads. For example, spare panel positions should be considered in an initial switchgear layout, to cater for future extensions.
(2) In-service condition. These relate to such aspects as temperature, humidity, air conditioning and ventilation, dust, corrosion or pollution conditions and affect not only choice of components but also enclosure design.
(3) Foreknowledge of standard equipment available in the market. Custom-built equipment is not justified when standard and competitive units may be readily adapted.
(4) Requirements for reliability and security of supply. For example, the degree of security required influences the choice of bus-bar systems both in terms of numbers installed and their sectionalising.
(5) Requirements for maintenance and safety, consistent with the skills of operating personnel.

26.5.6.2 Fault considerations

Overcurrent protective devices must operate to isolate short-circuit faults safely, minimise damage to circuit elements and avoid, if possible, shutdown of plant. An accurate knowledge of prospective fault currents throughout the system is essential for the correct application of protective devices and the design of bus-bars and terminal arrangements to withstand consequential mechanical and thermal stresses.

Generator short-circuit fault current decreases from a high initial value determined by the subtransient reactance X''_d of the machine, through a lower value determined by the transient reactance X'_d, settling after 0.6–2.0 s to a steady-state level determined by the synchronous reactance X_d. Circuit-breakers and fuses should operate before the steady-short-circuit condition is reached.

System faults are also fed by synchronous and induction motors, which may generate by release of kinetic energy by their rotating masses. The fault contribution from an induction motor ceases after a few periods. Typical values of subtransient and transient reactance, in per-unit on a machine-rating base, are:

	$x''_{\mathrm{d}}{\Leftarrow}$	$x'_{\mathrm{d}}{\Leftarrow}$
Salient-pole generators:		
up to 12-pole	0.16	0.33
14 poles and upwards	0.21	0.33
Synchronous motors:		
4 and 6 poles	0.15	0.25
Induction motors (low-voltage)	0.20	—

In the calculation of system fault levels it is sufficient to employ reactances only, except that in low-voltage (l.v.) systems the resistance of cables cannot be ignored. Arcing impedances should be included in l.v. system calculations. Typical values of per-unit arcing-fault current on 0.415 kV three-phase systems are: 0.70 for a three-phase arc, 0.57 for line-line and 0.14 for line-neutral.

26.5.6.3 Instrumentation and metering

Single running sets require an ammeter in each line (or an ammeter and a selector switch), a frequency indicator and a wattmeter. Optionally, a watt-hour meter and a power recorder may be fitted.

The instrumentation on each of any parallel running sets should include: an ammeter in each line (or a single ammeter with selector switch) a wattmeter and a reactive volt-ampere (var) meter. Power-factor meters may sometimes be substituted for var meters but as they develop low torque at low load they are prone to reading errors below 25% of rated current. The power factor calculated from kilowatt and kilovar values provides a more reliable and accurate indicator. Additionally, a set of synchronising instruments, perhaps mounted on a swivelling frame, is required, comprising a synchroscope, a double movement voltmeter and a double movement frequency indicator. Any incoming generator before being paralleled to the bus-bar should be connected by plug or switch to the 'incomer' movements of this voltmeter and frequency indicator, the second movements of both instruments being permanently connected to the bus-bars.

Recording instruments provide useful information for reconstruction of events, for registering operational patterns and indicating trends. In their multi-channelled form they give a time-sequenced record of events during disturbances and faults, particularly when they simultaneously monitor closing and tripping of key circuit breakers in the system.

26.5.6.4 Protection

The parameters to be monitored on the prime-mover have been dealt with in the preceding text. Protection related to the generator for which provision should be made, might in addition to system-fitted protective devices safeguarding against external short circuits and overcurrent, include the following:

(1) Restricted-earth-fault protection, combined if necessary with overcurrent protection in the one relay.
(2) Differential (circulating current) protection, where access is possible to the stator windings before the star point. Where a generator and step-up transformer are directly connected, an overall-biased differential protection scheme could be considered.
(3) Stator earth-fault protection. It is possible to combine this with differential protection in the one relay.
(4) Rotor earth-fault protection, by field or stator monitoring.
(5) Generated overvoltage and overfrequency protection.
(6) Unbalanced loading protection, with negative-phase sequence protective gear.
(7) Load-shedding protection.

In addition, for generators running in parallel:

(8) Reverse-power protection.
(9) Relays should be set to operate at about 10% reverse and be time delayed.
(10) Check synchronising relays.

Voltage surge protection should be considered where vacuum breakers or contactors are used, particularly with larger and higher voltage rated generators, if surges induced by lightning discharges on lines external to the diesel generator plant or initiated by switching are likely.

26.5.6.5 Control gear

Control gear should be designed to give a comprehensive indication of the state of a generator plant at all times and provide the means for modifying that state. Equipment for this purpose would include: measuring instruments, condition indicators, alarm annunciators, prime-mover and generator regulating controls and command signalling devices to engines and switchgear.

On the simplest single-generator installations, instrumentation and logic controls are usually incorporated within the switchgear cubicle(s) to give convenient operation of the plant from one location. On the more complex multi-generator plants, since adjustments on one machine affect all others in parallel with it, simultaneous observation of all the effects of any one action is essential. Where only two or three generators are involved, individual switchboards arranged as for the single generator may suffice, provided that they are placed as close together as possible and preferably in a quiet location away from the noise of running machinery.

But where adequate observation of a greater number of switchboards is difficult from one vantage point, it is advisable to use a control desk placed some distance in front of the composite switchboard. This may be used to house the generator regulating and control equipments whilst instrumentation and switchgear indicators are retained on the individual switchboards. Where size and cost of an installation warrants, a mimic diagram, representative of the complete generator plant and its feeder networks, may be incorporated into this control desk. Alternatively, it may be fitted into a separate diagram panel surmounting the composite switchboards. The diagram should incorporate pilot lights and semaphores to represent circuit-breakers, bus couplers, isolators and selector switches with miniature indicating instruments to cover feeders and generator output conditions.

26.5.7 Operational aspects

26.5.7.1 Engine ratings

Since the i.c. engine is an air-aspirating machine its output is affected by changes in the temperature and the pressure of the air it breathes. Power ratings should be quoted to the Standard Reference Conditions (S.R.C.) contained in the applicable National Standards of the engine maker concerned. For British built engines this is BS 5514: Part 1, which is equivalent to ISO 3046/1. Although corresponding German (DIN 6270), American (SAE J-243) and Japanese National Standards cover the same technical subject matter, their treatment of it differs, particularly with regard to the standard reference conditions adopted and definitions of kinds of power.

BS 5514, which covers reciprocating i.c. engines using both liquid and gaseous fuels, applies the following standard reference conditions:

Total barometric pressure	100 kPa (750 mmHg)
Air temperature	298 K (25°C)
Relative humidity	30%
Charge air coolant temperature	298 K (25°C)

The ISO standard power rating of the engine has to be adjusted for ambient conditions falling outside the standard reference condition, to arrive at a predicted on-site or service power rating. Section 10 of BS 5514: Part 1:1987 sets out the method for re-rating RIC engines. It lists the formulae to be applied for this purpose; and Annexes B to G offer tables and examples to help simplify and explain the methods of calculation. Since the calculations are fairly complex (see Annex G of BS 5514: Part 1), most engine manufacturers reduce the formulae to simplified forms stating: the percentage reduction in power for a certain increase in temperature and altitude, above the BS 5514 SRC levels. As it is rare, in any part of the world, for high humidity to be combined with very high temperature, a de-rating exceeding 6% for humidity is seldom, if ever, warranted.

The power categories of importance to the diesel generator user are 'continuous power; and 'overload power', as defined in BS 5514: a footnote states that it is customary to permit an overload of 110% power 'for periods and speed corresponding to the engine application'. Engine makers almost universally permit 10% overload for 1 h in any 12 of continuous operation on generator application. In so doing they rightly perpetuate a requirement of BS 649, which was withdrawn on publication of BS 5514 in 1977.

Prospective engine users should be wary of rating classifications other than those above, for generator applications, e.g. stand-by rating, continuous duty with time-limitation rating, reserve rating, intermittent rating and maximum rating, to name but a few of those cited. American, European and Japanese manufacturers will usually declare ISO standard power ratings, when asked.

26.5.7.2 Fuel oils

Liquid fuels used in compression-ignition engines are categorised as either distillate fuels or residual (blended) fuels. These are further subdivided into classes based upon viscosity, measured in seconds, using either the Redwood or the Saybolt systems. A fuel with a viscosity greater than 150 s Redwood No. 1 (expressed as 150 SRI) at 37.8°C is normally considered to be a residual oil.

BS 2869 lists nine classes of fuel oil, two of which, classes A and B, are specifically produced as engine fuels. Of the remaining seven, classes E, F, G and H are classified as industrial and marine fuels: E and F are light residuals or boiler fuels, G and H are heavy residuals or bunker fuels. The distillates, class A (gas oil) and class B (diesel oil), are the most widely used fuels on medium- and high-speed engines. Their American ASTM diesel fuel classification equivalents are D 975-66T Nos 1-D and 2-D. Grade 4-D of ASTM.D 975-66T, although considered a distillate oil, is intolerable to some high-speed engines and is best suited to low- and medium-speed units.

Almost all medium-speed engines will operate on the class E and F light residual fuels, whilst crosshead two-stroke low-speed engines and some of the slower running medium-speed four-stroke engines can successfully operate on class G and, to a lesser extent, class H heavy residual fuels. Residual fuels must be pre-treated, i.e. settled, heated, separated by centrifuge and filtered or waterwashed before being transferred to an engine fuel system. Heavy fuels should not be mixed with gas, distillate or light residual fuels.

The effects that fuel constituents have on engine performance and their influence on maintenance may be briefly summarised as follows:

Viscosity The lubricating property of the fuel falls as the viscosity reduces. Fuel injection equipment is most affected by any reduction in fuel lubricity or excessive water content in the fuel. Special precuations are necessary when 'dry' fuels such as Avtar, Avtag and Avcat are used with high-speed engines.

Sulphur content Sulphur in a fuel forms a corrosive acid when it combines with exhaust gas condensates. This is wear-inducing and may be minimised by ensuring that high jacket-water temperatures are maintained.

Canradson carbon residue value This is a measure of the tendency of a fuel to form 'coke' when heated. A fuel with high Conradson value reduces combustion efficiency whilst increasing the rate of carbon build-up in the engine. This, if coincident with a high sulphur content, greatly increases wear rate.

Ash content The effect on the engine is similar to that described for the Conradson value.

Vanadium or sodium content Their presence in any appreciable quantity affects engine exhaust valves and turbochargers. Contents below 50 ppm are recommended if frequent incidences of valve seat burning and turbine blade failure are to be avoided.

26.5.7.3 Lubricating oil

Selection of lubricant depends on whether the engine has separately lubricated bearings and cylinders (low-speed and some medium-speed engines) or whether it has combined lubrication of bearings and cylinders (all high-speed and most medium-speed engines).

Where separate lubrication is used, specially formulated and refined mineral oils containing anti-oxidant additives are specified for bearings. Some engine operators favour detergent additives despite their tendency to thicken relatively rapidly. Cylinder lubrication calls for a heavier oil than that suitable for bearings, but on many engines the same viscosity grade is satisfactorily applied to both.

For combined lubrication, oils need: a viscosity applicable to both cylinders and bearings; high oxidation stability; sufficient detergency and corrosion inhibition properties. A big proportion of medium- and high-speed engines use a straight mineral oil of the type applied to separately lubricated bearings; but others, especially the highly rated high-speed engines, require heavy duty oils with detergent and contaminant dispersive properties, e.g. MIL-L-46152 specification for naturally aspirated engines and MIL-L-2104C for turbocharged units. If there is any possibility of an engine being subjected to prolonged light-load running, such as on many telecommunication transmitter applications, it is advisable to employ a heavy duty oil even though the engine might not normally require it.

Engine makers will always specify the oil to be used consistent with the duty and rating of an application and the fuel which it is intended to use. Mono-grade oils are normally recommended but makers will countenance the use of multi-grade types, provided that the operator obtains the oil supplier's assurance that the type proposed meets the certified performance level of the equivalent and acceptable mono-grade oil.

26.5.7.4 Maintenance

The diesel engine and the gas turbine are both i.c. heat engines but since the former employs reciprocating motion it is mechanically the more complex machine and demands a greater skill in its maintenance. Also its wear rate is higher. However, repair services for diesel engines are more readily available world-wide than they are for gas turbines.

Plant operating personnel should be fully familiar with all its component parts and be capable of maintaining it in its optimum condition. Operation must be continuously monitored and accurate records kept so that incipient faults are detected and repair effected before major, unscheduled and costly stoppages occur.

The starting point of any preventive maintenance programme must be the manufacturer's operator and service manuals. They indicate what needs to be checked and how frequently. Planned schedules should at first rigidly adhere to the maker's recommended frequencies for inspection checks and maintenance tasks. Only after sufficient operational experience has been accumulated should one contemplate any modification to fit the particular installation and its operating conditions, and certainly not before the first general overhaul (stripping to crankshaft bearing level) has been undertaken. This gives the opportunity to assess achievement against the maker's wear and renewal limit schedules.

Whilst the service intervals recommended by manufacturers are conservative and based on average experience and temperate conditions, periodicity of inspection and service may have to be increased where, for instance, the quality of fuel is in question or where corrosive and very dusty environments pertain, or where less than 50% loading conditions are initially expected. Over the last decade, engine and component manufacturers have achieved impressive improvements in times between overhauls. Typically, major overhauls on well-maintained high-speed engines may now occur at 15 000 h intervals, whilst those on medium-speed engines could be at between 20 000 and 30 000 h—the longer interval applying to the slower running engines.

Careful consideration must be given to plant spares stocks. A more comprehensive holding of both non-consumable and consumable items would be necessary at remote sites or in certain developing countries, where there are no local accredited spares stockists. At locations where there is ready access to good stockists, a much-reduced inventory on consumables is justified.

Even with the most vigilant monitoring, problems can arise at any time. Running plants are usually most vulnerable immediately after commissioning and after general overhauls. *Table 26.7* based on 4 years of analyses of the excellent *Annual Working Cost and Operational Reports* published by the Institution of Diesel and Gas Turbine Engineers, shows the major classes of failure (expressed as rounded-off percentages of total unscheduled stoppages) reported by over 100 diesel and gas engined plants, world-wide.

Table 26.7 The major classes of failure at diesel and gas engined plants

Class of failure	*Percentage of total stoppages*
Fuel injection equipment and fuel supply	26
Water leakages and cooling	16
Valve systems and seatings	13
Bearings	7
Governor gear	6
Turbocharger/lubrication/piston assemblies/gearing and drives	4 each

26.5.8 Plant layout

The size of a plant room is largely determined by the number and rating of the generators installed and by their requirements for ancillary equipment. In addition to the engine-generator assemblies there will be switchgear, distribution and control gear, engine starting equipment, fuel service tanks, provisions for fuel and lubricating oil storage, engine cooling systems, exhaust silencing equipments, and (on the larger engines) lubrication and fuel systems external to the engine.

The internal layout should be such that the basic requirement is to construct a station building around the machinery. It is prudent to provide for future expansion. Growth may be in the form of a larger generator unit to replace the original or additional units, in multi-generator plant, to cater for increased load demand. A removable end wall offers one way of providing for future plant room expansion.

A minimum space of 2 m should be allowed around each set to facilitate maintenance. On multi-engined stations there should be sufficient headroom for overhead, installation and servicing cranes—one for large lifts of 10–30 tonne capacity and a smaller unit of about 2–3 tonne. Height to the underside of the common crane rail girder should be such that the distance between the bottom of the hook on the larger crane, when fully raised, and the floor level, is about 6 m for the biggest engines installed.

Most generator-set manufacturers provide an advisory service on foundation requirements. Where sets are not supplied with anti-vibration mountings a concrete foundation block, preferably 'isolated' from the main building structure to minimise vibration (and therefore noise) transmission, is necessary. A good empirical estimate for foundation mass is that it should be at least 1.5 times the dynamic mass of the associated diesel alternator. On new civil works a basement can be provided at little extra cost to house engine auxiliaries. Similarly, a gallery can be constructed to accommodate fuel service and water make-up tanks. This reduces the need for pipework trenches within the generator hall itself. Access to basement auxiliaries for replacement or maintenance could be by removable open-mesh gratings.

Where ratiators are fitted to generating sets they should have pusher fans and be installed near to and facing an outside wall, with air ductwork and control louvres to regulate the plant room temperatures. Space between the rear wall and the generator end of each set should be sufficient to allow for end-removal of major components, as required.

For ventilation puposes it can be assumed that approximately 8% of a generator nameplate kilowatt rating is radiated as heat from the engine and generator carcasses. The combustion air required by an engine may be taken to be approximately 9.5 m^3/h per nameplate kilowatt. Plant-room air exchange required is then the sum of the ventilation air and engine combustion air requirements. Exhaust gases should be piped to atmosphere through insulant-caulked apertures in an outside wall and silencers should be mounted external to the plant room, if at all possible.

Figure 26.40 gives floor area and height requirements, recommended by the Building Services Research and Information Association (BSRIA) in their *Technical Note TN4/79*, for individual stand-by generators in the range 50–625 kV-A.

Where noise is likely to be a community problem it is necessary to identify all the noise contributors in the plant and obtain octave-band frequency analyses for each so as to calculate the total generated noise level, in worst-case conditions. This must then be related to any noise level limits imposed by communal interests or local legislation in order to determine the noise reduction required. The appropriate noise control treatments should then be selected to restrict the noise transmitted and radiated from the plant room to a value below the promulgated level. Acoustic barriers, partial enclosures, vibration damping materials, vibration isolation, inertial blocks, lined duct work and splitter silencers in ventilation inlets and hot air discharge outlets are but some of the noise control techniques that could be considered. Any treatment applied should not prejudice the operation, maintenance and safety of the plant.

26.5.9 Economic factors

When considering the installation of private generating plant it is important that proposals are not only technically suitable for purpose but are also economically defensible. The actual or predicted costs of purchased power should be compared with the projected costs of privately generated power for clearly defined electrical and heat load cycles, before any management decision is sought.

Factors to be considered are:

(1) Capital costs, embracing land; site preparation and access; foundations; buildings; workshops and tools; fuel storage; the power plant and its ancillary equipments; cranes; stores and non-consumable spares and heat recovery equipment (if applicable).
(2) Installation costs.
(3) Operating costs; fuel; lubricating oil; service spares; wages of operating and maintenance staff; insurance; depreciation, interest on capital and rates.
(4) Costs of any consequent supply outages.
(5) The size of the installation and its mode of operation:

    (a) *Base load*: independent of utility supply, supplemented by utility, with utility as stand-by, and total energy.
    (b) *Peak lopping*: independent of utility, and in parallel with utility.
    (c) *Stand-by to utility supply*.

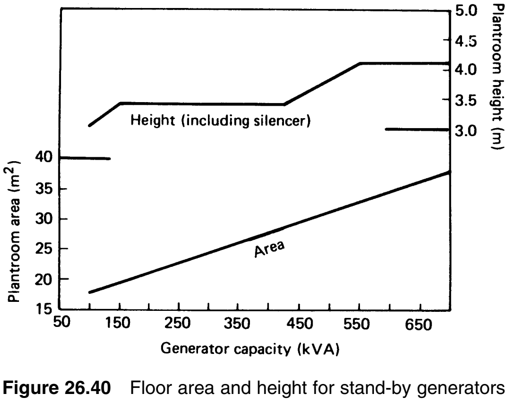


Figure 26.40 Floor area and height for stand-by generators

Running costs are related to the pattern of operation, the number, rating and type of engines, the fuel type used, the overhaul intervals and the hours of maintenance per year. The following cost indicators will enable reasonable estimates to be made at the first stage of any appraisal, when alternative schemes are being investigated.

26.5.9.1 Fuel costs

The cost of fuel to generate electricity only can be estimated as one-eighth of the cost of purchased electricity expressed in unit currency per kilocalorie (e.g. p/kal, where 1 kW-h =⇐ 0.863 kcal). Assume a consumer purchased 150 000 kW-h of electricity over a given period at a total price of 1.5 p/kW-h (=⇐1.74 p/kcal). This would equate to a cost of purchased power of £2610 over the period. The cost of fuel to privately generate 150 000 kW-h would then be £326.

26.5.9.2 Lubricant costs

Lubricating oil consumption may be taken to be 1.5% of the fuel-oil consumption at full load. To this must be added the quantities representative of any oil changes at routine service intervals. This may vary from between 250 h to every 5000 h of running, depending upon the size and speed of the engine. The sump capacity of a 1.2 MW engine is of the order of 800 l. For first estimates of lubricating oil costs, one may work on the basis of 5% of fuel costs for the same period.

26.5.9.3 Maintenance costs

Estimate on the basis of 5% of the combined costs of fuel and lubricating oil over the given period. This covers spares and labour.

26.5.9.4 Depreciation, interest, insurance and rates

Engineers do not always appreciate how substantial these costs can be. A recent economic appraisal for a private generating scheme, employing two 1.2 MW sets (with a third in reserve) operating at a 92% load factor and written down over a 7-year period, revealed that the annual sum of these costs was 70% of the combined total of the other running costs, i.e. those attributable to fuel, lubricant, maintenance, spares and labour.

26.5.9.5 Capital costs

Table 26.8 offers a very rough guide for first estimates on capital costs (including installation), using the cost of the engine generator and its auxiliaries as base 100.

Table 26.8 Estimates of capital costs

Cost item	*Size of unit* (MW)		
	1	*2*	*4*
Engine-generator and auxiliaries	100	100	100
Civil works	20	19.5	16.5
Cranes and services in station	15.5	13.5	10
Installation	18.5	16.5	12.5

26.5.10 Cogeneration/CHP (see also Section 26.3.4)

A total energy system implies on-site power generation in which the energy input from either liquid fuel (diesel engines) or a combination of gaseous and liquid fuels (dual-fuel engines) is maximised by recovering the waste heat from the generating process. By so doing the overall thermal efficiency of generation may be raised from 37% to about 80%.

Compared with gas turbines the quality of waste heat is of a relatively low grade. Nevertheless, sizeable combined heat and power installations have been commissioned both in the UK and in Europe in recent years, most favouring multi-engined systems employing 1.5–2 MW unit sizes with dual-fuel operation.

Table 26.9 shows typical heat balances for a 2 MW 750 rev/min diesel-generator giving 1.8 MW in the dual-fuel mode.

Engines of this type and size give about 1 MW of recoverable heat from jacket water and lubricating oil in the form of low pressure water at 80 °C and another 1 MW recovered from the exhaust gases to give about 1400 kg/h of steam at 850 kPa.

Recoverable heat from a diesel engine is of the order of 250 000 kcal/h/MW from the exhaust gases and 350 000 kcal/h/MW from the jacket water.

Using exhaust heat recovery alone, 0.5 kg of steam at 850 kPa/kW-h is possible. With full jacket, oil and exhaust heat recovery this figure doubles to 1 kg of steam per kilowatt-hour and can be raised still further to 2.5 kg/kW-h if pre-heating of the jacket water into an automatic boiler, using the same fuel as the engine, is employed.

Whilst it was once considered uneconomical to employ heat-recovery systems on generators below 500 kW, the Electricity Act of 1989 has stimulated interest in small-scale combined heat and power (CHP). A UK potential of 3.5 GW of small-scale co-generation has been estimated, for the next 10 years. The majority of these plants will use engines operating on gaseous fuels; and applications will include:

(1) hotels;
(2) hospitals;
(3) leisure centres;
(4) small-scale district heating;
(5) sewage treatment farms; and
(6) industrial processes.

Recommendations for the electrical, mechanical and thermal protection of plants of this type are contained in publications such as:

(1) The Electricity Council Engineering Recommendations G 59;
(2) The Electricity Council's publication *ET 113—Notes on guidance for the protection of private generating sets up to 5 MW for operation in parallel with Electricity Boards' distribution networks*;
(3) British Gas publication *IM 17—Code of Practice for natural gas fuelled spark ignition and dual-fuel engines*; and
(4) H.M. Department of Energy, Energy Efficiency Office's *Good Practice Guide No. 1—Guidance notes for the implementation of small-scale packaged combined heat and power*.

Table 26.9 Heat balances for a 2 MW 750 rev/min diesel generator

Heat balance at full load	*Diesel* (%)	*Dual fuel* (%)
To electricity	38.5	39
To exhaust	36	34
To jacket water	11	10
To lubricating oil	4.5	4
To charge air	4	2.5
To radiation, etc.	6	6
To unmixed gases	—	4.5
Total	100	100

References

ARMSTEAD, H. C. H., 'The undervaluation of hydropower potential—a statistical pitfall', *J. Mech. Eng.* (March 1985)

BERNSTEIN, L. B., *Tidal Energy for Electric Power Plants*, Israel Program for Scientific Translation, Jerusalem (1965)

BOVET, G. A., 'Modern trends in hydraulic turbine design in Europe', *Trans. ASME*, **75**(6) (August 1953)

BRAIKEVITCH, M., 'Fluid engineering & development of water power', Fourth Fluid Science Lecture, British Hydromechanics Association (BHRA) (October 1972)

BRAIKEVITCH, M., HARTLAND, D. and STRUB, R. A., 'Development of reversible pump turbines and pump storage equipment', *J. ASCF* (October 1961)

BUCHI, G., *La Moderne Turbine Idrauliche e d'Regulatori di Velocita*, U. Hoepli, Milan (1957)

CREAGER, P. and JUSTIN, J. D., *Hydro-electric Handbook*, Wiley, New York (1950)

DANEL, P., 'The hydraulic turbine in evolution', James Clayton Lecture, Institute of Mechanical Engineers (1958)

DAVIS, C. V. (Ed.), *Hand Book of Applied Hydraulics*, McGraw Hill, New York (1942)

DERIAZ, P. E. and WARNOCK, J. G., 'Reversible pump turbines for Sir Adam Beck–Niagara Pumping–Generating Station', *Trans. ASME, J. Bas. Eng* (December 1959)

FRAU, O. P. and LEROZE, P. V., 'Pumping in a tidal power plant—experience at La Rance and main aspects of the turbine design', *Pumped Storage: Proc. of I.C.E. Conference*, T. Telford, London (April 1990)

GIBRAT, R., *L'Energie des Marees*, Presses Universitaries de France, Paris (1953)

GOLDWAG, E., 'On the influence of water turbine characteristic on stability & response', *ASME J. Bas. Eng.*, (December 1971)

GOLDWAG, E. and POTTS, R., *Energy Production Developments in Tidal Energy*, I.C.E./T. Telford, London (1990)

GUTHRIE-BROWN, J. (Ed.), *Hydroelectric Engineering Practice*, 2nd edn, Blackie & Son, London (1970)

HAMMON, N. W. and WOOD, P., *Tidal Power from the Mersey; History & Prospects, Developments in Tidal Energy*, I.C.E./T. Telford, London (1990)
HAWS, E. T., WILSON, E. A. and GIBSON, H. R., *Pumped Storage in the Proposed Mersey Tidal Project, Pumped Storage*, I.C.E./T. Telford, London (1990)
HILARET, P. and WEISROCK, G., 'Optimising production from the Rance tidal scheme', *Symposium on Wave, Tidal, OTEC & Small Scale Hydro Energy*, British Hydromechanics Association (BHRA) (May 1986)
HOVEY, L. M., 'Optimum adjustment of governors in hydro generating stations', *Eng. J.* (November 1960)
JAEGER, C., 'Present trends in surge tank design', *Proc. Inst. Mech. Eng.*, **168** (1960)
JAEGER, C., *Engineering Fluid Dynamics*, Blackie & Son, London (1956)
KERR, D., SEVERN, B. E. and DRIVER, S. J., *Severn Barrage: Civil Engineering Aspects Developments in Tidal Energy*, I.C.E./T. Telford, London (1990)
KOVALEV, N. N., *Hydroturbines Design and Construction*, Israel Program for Scientific Translations, Jerusalem (1963)
'La Rance tidal power scheme', *Revue Francaise de Energie*, 1983 (September/October 1966)
MASONYI, E., *Water Power Development*, Hungarian Academy of Science, Budapest, Vol. 1, 2nd edition (1963) and Vol. 2 (1960)
MEIER, W., MULLER, J., GREIN, H. and JAQUET, M., 'Pump-turbines and storage pumps', *Escher Wyss News*, 44(2) (1971)
MERMEL, T. W., 'The world's major dams and hydro plant', *Water Power & Dam Construction* (May 1990)
MILLER, H. and FOCAS, D., 'The Annapolis tidal power plant', *Proc. I.E.E.* (March 1984)
NACHLEBA, M., *Hydraulic Turbines, their Design and Equipment*, Artia, Prague (1957)
PETTY, D. J. and MCDONALD, A., *Turbine Generators for the Severn Barrage, Developments in Tidal Energy*, I.C.E./T. Telford, London (1990)
'Power from water, a special report', *Power*, **124**(4) (1980)
SEONI, R. M., SHADED, E. N., SIMPSON, R. J. and WARNOCK, J. G., 'Review of trends of large hydroelectric generating equipment', *Proc. I.E.E.*, **123** (October 1976)
TAMATSUKURI, T., 'Trends in hydroelectric generating equipment technology', *Hitachi Rev.*, **37**(2) (1988)
The Severn Barrage Project General Report (Energy Paper No. 57), HMSO, London (1989)
'The world's largest turbines', *Water Power & Dam Construction Handbook*, Reed Business Publishing Group, Sutton (1991)
'The world's pumped storage plants—survey', *Water Power & Dam Construction*, Reed Business publishing Group, Sutton (April 1990)
WISLICENUS, G. F., *Fluid Mechanics of Turbomachinery*, McGraw Hill, New York (1947)
VIVIER, L., *Turbines Hydraliques et leur Regulation*, Albin Michel, Paris (1966)
ZIENKIEVICZ, O. C., 'Transmission of water hammer pressures through surge tanks', *Proc. Inst. Mech. Eng.*, **168** (1954)

Codes

INTERNATIONAL ELECTROTECHNICAL COMMISSION:
Publication 41 International code for field acceptance test of hydraulic turbines
Publication 193 International code for model acceptance tests of hydraulic turbines
Publication 193A Amendment No 1
Publication 308 International code for testing of speed governing systems for hydraulic turbines
Publication 609 Cavitation pitting evaluation in hydraulic turbines, storage pumps and pump-turbines

27

Alternative Energy Sources

M A Laughton BASc, PhD, DSc(Eng), FREng, CEng, FIEE
Formerly of Queen Mary and Westfield College, University of London
(Sections 27.1 to 27.5, 27.7, and 27.9)

G W Brundrett BEng, PhD, FIMechE, FCIBSE
School of Architecture and Building, University of Liverpool
(Section 27.10)

P L Surman MA, MSc, PhD, CChem, CEng
The Energy Workshop, Harrogate
(Section 27.6)

R H Taylor PhD, DIC, FIEE, FInstP, CEng
Magnox Electric plc
(Section 27.8)

Contents

27.1 Introduction 27/3
  27.1.1 Embedded generation 27/3
  27.1.2 Energy costs 27/3

27.2 Solar 27/4
  27.2.1 Photovoltaic systems 27/4
  27.2.2 Solar thermal applications 27/5

27.3 Marine energy 27/6
  27.3.1 Tidal energy 27/6
  27.3.2 Wave energy 27/8

27.4 Hydro 27/9

27.5 Wind 27/10
  27.5.1 Wind energy 27/10
  27.5.2 Wind turbines 27/10
  27.5.3 Wind generators 27/10
  27.5.4 Economics 27/11
  27.5.5 Environmental issues 27/12

27.6 Geothermal energy 27/12
  27.6.1 Hydrothermal sources 27/13
  27.6.2 Hot dry rocks 27/13

27.7 Biofuels 27/13
  27.7.1 Introduction 27/13
  27.7.2 Biomass technologies 27/13
  27.7.3 Major biomass sources 27/15

27.8 Direct conversion 27/16
  27.8.1 Thermoelectric generators 27/16
  27.8.2 Thermionic generators 27/16
  27.8.3 Magnetohydrodynamic generators 27/16

27.9 Fuel cells 27/17
  27.9.1 Introduction 27/17
  27.9.2 Fuel cell types 27/17
  27.9.3 Fuel cell structure 27/19
  27.9.4 Fuel cell plant 27/19
  27.9.5 Regenerative fuel cells 27/20

27.10 Heat pumps 27/21
  27.10.1 Introduction 27/21
  27.10.2 Thermodynamics 27/21
  27.10.3 Practical cycles 27/22
  27.10.4 Scale 27/24
  27.10.5 Conclusions 27/29
  27.10.6 Professional guides 27/30

27.1 Introduction

The development of electricity generation from energy sources which are alternatives to hydrocarbons has become a major objective in Europe and elsewhere both for environmental reasons, principally the limitation of carbon dioxide emissions and, latterly, for the longer-term sustainability of energy supplies. The advent of major contributions from such sources is not without significant consequences, however, for the design and operation of electricity supply systems. Secure and low cost power supplies of high quality have been achieved based on historic energy supply patterns, network infrastructures, economies afforded by large-scale generating plant, system protection and control practices all of which would be affected by substantial levels of new distributed generation.

27.1.1 Embedded generation issues

At present virtually all small generating plants such as arising in most CHP and renewable power stations are connected to low voltage distribution networks not high voltage transmission grids. Individually most small embedded generating stations are not subject to central system control, partly because the high cost of conventional telemetry means that in the UK, for example, system control and data acquisition (SCADA) has only low penetration at voltages below 132 kV. At these levels, therefore, distribution networks operate largely with incomplete and uncertain information. In the main, any network problems caused by renewable generators arise either because of the intermittent nature of their output or their point of connection to the network, and not particularly because they are renewable energy sources. In fact some (e.g. biomass, landfill gas, etc.) are indistinguishable from conventional generators. Others (such as windfarms, tidal schemes, wave-power, etc.) can cause problems to the network operator, ranging from:

- System stability problems when a significant proportion of system demand is supplied by randomly intermittent generators, especially at light load. Small generators with low inertia do not provide the same buffering capacity as larger plant under the power swing conditions associated with system disturbances. If a generating plant fails, it is normally disconnected from the network to maintain system stability despite increased demand from the remaining generating sources. Only once the disconnected plant is running in a synchronised manner can it be reconnected to the grid.
- Voltage control and quality problems when generators embedded within the distribution network start/stop generating. This can cause other network users to suffer voltage fluctuation, dips and steps outside of the statutory limits and inject unwanted harmonics into the voltage waveform. The cure requires much more active operation of low-voltage tap-changers on transformers than these have been designed for leading to increased risk of failure, higher maintenance and replacement costs.
- Frequency control problems—many new forms of generation (CCGTs included) don't provide the expected response to low frequency, exposing users to more severe and more frequent excursions in system frequency.
- Large-scale renewable sources (tidal barrages, offshore wind and wave power, even new-generation nuclear) are likely to be situated remote from load centres. Exploitation of these sources requires major investment in reinforcing and expanding the transmission grid; however gaining rights of way to build new circuits can face determined and prolonged opposition from environmental lobbyists.
- The traditional network infrastructure causes additional difficulties in the siting of new generators. Distribution networks are generally designed and operated in radial configurations and are not designed to accommodate active sources of energy. This practice is based on the well established aim of minimising infrastructure costs associated with the number of conductors, protection equipment and switchgear size, apart from obviating the need for power flow control. Embedded generation adds to local fault levels and hence, sooner or later leads to the need for larger switchgear as well as the redesign of the protective systems. In addition and, as a consequence of this design practice, the distribution networks are often tapered in power flow capacity from the bulk supply point down to the customer in much the same way as a road or water network. Much renewable generation, e.g. wind, is sited away from the bulk supply points and nearer to the ends of the network, hence the difficulties in finding suitable connection points on the joint grounds of limited power flow and switchgear capacity.

With suitable engineering redesign and investment these problems can be overcome in time to a certain degree, but the costs of reinforcing the existing distribution network, upgrading control and protection systems, switchgear, transformers and reactive plant to cater for substantial embedded generation have to be considered. It is not clear how these costs will be apportioned in the privatised electricity supply industries now in existence. Furthermore it is not clear either how much randomly intermittent power from wind, wave or solar sources an electricity supply system can tolerate. According to the International Energy Agency as the contributions of these sources approaches 12% of power supplied and electricity not being capable of storage, policy makers need to start thinking about creating an energy buffer. At a 20% level it is contended that a buffer is absolutely necessary. There is, therefore, a technical limit to the development of renewable energy supplies if their output is geared solely to direct connection to the electricity supply system. Such a buffer can be supplied by pumped storage schemes in principle if enough storage capacity is installed, but the big development foreseen is the eventual move to a hydrogen economy with hydrogen from electrolysis providing the buffer and energy storage needed. In this context the development of fuel cells is a complementary activity to the development of renewable energy sources.

27.1.2 Energy costs

Two major methods for evaluating the unit costs of generation are:

- as annual costs per unit of kWh output relating to the costs appropriate to any particular year including a share of capital expenditure, and
- as a levelised cost per unit of kWh output which allocates lifetime production costs over the lifetime of the investment.

The latter method takes into account the time value of money and is the average present value cost in constant money terms per unit of energy produced at which the revenue from total lifetime output equals the total lifetime costs including capital, O and M and any other plant costs. This levelised cost method is recommended for use by the IEA[1] for comparison of generating plant costs under equivalent conditions because it covers the whole investment life of the system under consideration.

Ideally the costs should be evaluated at the user end point of consumption which would then include not only generation but transmission, distribution, system security, environmental, administration and other overhead costs; however beyond the generation plant boundaries the allocation of these extra costs is complex and maybe arbitrary. In electricity analysis, therefore, the bus-bar cost at the point of production is the only cost considered.

To relate capital cost to energy supplied the levelised cost per unit of kWh is found using a sinking fund method for uniform capital recovery whereby a constant capital charge rate $\boldsymbol{A}$, or annual payment, is made into the fund, which then accumulates interest and totals to the desired sum at the end of the period of, say, N years. Since no capital is redeemed interest is payable on the full capital amount $\boldsymbol{P}$ and the capital charge rate is

$$\boldsymbol{A} = \boldsymbol{P}\frac{i(1+i)^{\mathrm{N}}}{i(1+i)^{\mathrm{N}} - 1}$$

where i is the interest rate. Here the period N relates to the period for capital recovery rather than to the longer period of plant technical lifetime. The capital amount $\boldsymbol{P}$ is not usually the total capital cost of the project, because projects financed with private equity as well as loan, say in a 20%: 80% ratio, would be able to sustain lower energy production costs to cover the loan with returns to investors from whatever revenues from prices surplus to costs are possible.

Defining the factor $\boldsymbol{R}$ as $\boldsymbol{R} = i\,(1+i\,)^{\mathrm{N}}/i\,(1+i\,)^{\mathrm{N}} - 1$ then values of $\boldsymbol{R}$ are given in *Table 27.1* for periods of 3 to 25 years and for interest rates of 8 to 14%.

Given the rated plant full-load power generating capacity in kW, the load or capacity factor which is the equivalent proportion of the year that the plant would have to be generating at rated output to produce the actual energy output achieved and the total number of hours in a year (8760), then the energy $\boldsymbol{E}$ produced per annum in kWh is

$$\boldsymbol{E} = (\text{rated power full-load capacity}) \Leftarrow \times (\text{capacity factor}) \times (8760) \Leftarrow$$

The cost of generation $\boldsymbol{C}_{\mathrm{cr}}$ attributable to capital recovery is then $\boldsymbol{C}_{\mathrm{cr}} = \boldsymbol{A}/\boldsymbol{E}$

To this cost must be added the annual operating costs including maintenance and, in the case of conventional plant, the annual fuel costs.

Comparisons of costs of energy produced from different renewable sources and from conventional generation plant involve a number of factors, which are inevitably site specific; therefore calculated costs should be considered only as indicative. The costs of any energy system could fall within a wide range depending on the specific conditions found at each site, the maturity of the technology, the economies of scale, the financing arrangements and the market conditions in which the electrical energy produced is traded.

Table 27.1 Values of factor *R* for capital recovery

Periods *N years*	*Interest rate* 8%	*Interest rate* 11%	*Interest rate* 14%
3	0.388	0.409	0.431
5	0.251	0.271	0.291
10	0.149	0.170	0.192
15	0.117	0.139	0.163
20	0.102	0.126	0.151
25	0.094	0.119	0.146

27.2 Solar[2–9]

27.2.1 Photovoltaic systems

The present day market in the application of photovoltaic (PV) cells is defined by three sectors

- the *professional sector* which includes the powering of orbiting satellites or telecommunication repeaters in remote locations where the requirement for reliable, maintenance-free, independent power is essential,
- the *off-grid sector* which includes applications in developing countries, rural health, water pumping, irrigation and rural lighting and education, and
- the *grid-connected sector* where the PV electricity flows into a national grid. This sector is sub-divided again into two, that which forms part of centralised generation, e.g. for powering pumped storage schemes, and that which forms the larger sub-section of building integrated photovoltaic systems (BIPV).

27.2.1.1 Principles

When light strikes certain semiconducting materials such as silicon, gallium arsenide or cadmium sulphide, fabricated in the form of a p–n junction, an electric current can flow through an externally connected circuit. About 80% of PV cells manufactured are based on crystalline silicon. The silicon semiconductor band-gap energy of 1.1 eV, equivalent to a photon wavelength of 1100 nm, is the minimum threshold photon energy required to excite an electron-hole pair. Light of wavelength greater than 1100 nm, about one third of the spectral content of sunlight, does not have sufficient energy to create an electron-hole pair, whereas light of wavelength less than 1100 nm has sufficient energy to create an electron-hole pair, but only the band-gap energy can contribute to the cell voltage. The excess energy is lost in phonon transitions which heat up the cell. The theoretical efficiency of the silicon cell is thus around 30%. At present because of material imperfections the best laboratory cells are 24% efficient.

In bright sunlight, an irradiance of 1 kW/m^2, a 10 cm square cell will give an output of about 0.5 V and 3 A, i.e. about 1.5 W of power, or 150 W/m^2. Manufacturers quote the output of their cells for a sunlight intensity of 1.5 kW/m^2 (similar to that of the Sahara Desert at noon). This standard output is labelled 'peak watts ' or '$\mathrm{W_p}$', and is measured at a standard temperature of 25°C. The power output of a solar cell varies with the intensity of light falling on it, which includes the angle of the sun to the plane of the module. The current output is halved if the light intensity is halved, but the voltage will drop by only a few per cent. The voltage output also depends on the temperature of the cell and decreases by about 0.5% for every degree Celsius rise in temperature above 25°C. Such cells need to be connected in series/parallel to provide a current-voltage relationship suitable for the load.

27.2.1.2 Technology

Three feasible broad groups of cell modules comprise crystalline flat plate, thin film flat plate and concentrator designs. Currently available commercial flat plate modules are based on either crystalline silicon or thin film amorphous silicon cells. There are also a number of concentrator installations in different countries based on single crystal silicon cells. Crystalline silicon flat plate modules dominate the power market, defined as arrays larger than 50 $\mathrm{W_p}$. Cells are made from single crystal or polycrystalline material with individual cells series-strung and then encapsulated behind low iron-content tempered glass. The best commercially available product at present is 16–17% efficient. Thin film cells based on materials having better photon absorption properties such as amorphous silicon, cadmium telluride and copper indium have had considerable development for the consumer market and also several large demonstration plants are in operation around the world. Low cost manufacturing of thin film solar cells has yet to be realised and, although having lower efficiencies than silicon cells, they hold the promise of being eventually cheaper to manufacture in high volume. The use of more expensive cells is possible by using optical concentration and high efficiency cells based both on gallium arsenide and silicon.

27.2.1.3 Photovoltaic systems

Typical modules produce about 14 V d.c. and about 3A under an irradiance of 1 kW/m^2. Higher power levels are obtained by series and/or parallel connections. Various module sizes rated from 30 W to 120 W are presently available. The module must also be mechanically supported to withstand wind loads and be correctly orientated to obtain maximum irradiance. The power output must be delivered in an acceptable form such as a constant d.c. voltage or 240 V, 50 Hz a.c. and this usually needs power-conditioning equipment. Presently available solar modules are manufactured to withstand extremes of climate and carry a twenty-year warranty to back up claims of trouble-free reliability.

The cost of a PV system comprises the manufacturing costs of silicon wafer production and assembly of modules plus the costs of installing the system. The silicon wafer cost is approximately 25% of the system cost at present of which 5% is the silicon feed-stock material cost, the wafer-to-module conversion cost represents a further 25% with the installation of the system comprising the mechanical support structure, cabling and inverters making up the remaining 50% of the total installed cost.

The PV industry is growing rapidly at about 30% per annum with significant urban markets developing in Germany, Holland, Japan and the USA stimulated by attractive subsidies. In the USA, for example, in 1998, there were 8500 solar buildings, up from only 2000 the previous year. The projected numbers of solar buildings for 2000, 2005, and 2010 are 51 000, 376 000 and 1 014 000 respectively. By way of illustration the Long Island Power Authority is providing a $4.5 million grant to help underwrite the capital cost of a photovoltaic system requiring 7 800 PV panels of varying sizes on the roofs of three buildings, having a capital cost of $9.3million (2001). The system will generate 1.5 MW of power during peak periods.

Guidelines and standards applicable to PV systems in various countries cover the requirements for safe installation of PV building-integrated systems applicable to PV modules and arrays, e.g. critical temperatures, voltage ratings, cable and insulation types, sizing for safe design, overcurrent protection, manual disconnects, grounding, anti-islanding protection, and in-surge and transient protection.

The major barriers to PV becoming a large-scale energy source are generated electricity costs and availability of raw materials. Today's PV systems sell for about \$6–\$10/W$_p$, with an implied electricity price of about \$0.4 to \$0.7/kWh in California and \$1.00/kWh in the UK. This compares with a domestic electricity price of around \$0.10/kWh in the UK and average generation costs of \$0.04 to \$0.05/kWh. Projections of future costs based on progress in PV technologies are consistent with module costs below \$0.5/W$_p$ (compared to \$3/W$_p$ today) with further reduction in installation costs. In the longer-term material availability may have to be addressed because several thin film PV technologies that are expected to have excellent cost potential use a rare raw material: germanium in amorphous silicon; tellurium in cadmium telluride; and indium and gallium in copper indium diselenide.

In the UK PV technology is likely to be grid-connected avoiding the cost of battery storage with integration into building facades or rooftops. Of all the renewable energy technologies contributing to the generation of electricity, photovoltaics is perhaps the one most easily integrated into the existing electricity supply structure, also with minimal environmental impact.

27.2.2 Solar thermal applications

Solar energy can be converted easily into useful heat and provide a significant proportion of space heating and hot water demand for low temperature-rise applications, e.g. domestic hot water.

27.2.2.1 Active solar systems

The most widely used active method for converting solar energy into heat is by the use of flat plate collectors comprising an absorber plate (transparent cover), tubes or channels integral with the collector absorber plate carrying water or other fluid, an absorber plate which is normally metal and with a black surface, insulation to minimise heat losses and a casing for protection against the weather, combined with pumps or fans for the circulation of the heat. Collectors come in a variety of forms, with combinations of flat, grooved and corrugated shapes for transferring the absorbed solar radiation from the surface. Solar ponds about 1 m deep with a blackened bottom can also be used as a collector offering the possibilities of achieving water temperatures up to 100°C with a collection efficiency of 15–20% using salt concentration to ensure that the density of water increases towards the bottom, thus preventing circulation by convection. Heat can be removed by circulating the lower levels of water through a heat exchanger without disturbing the upper layer. Advanced collectors are necessary for temperatures greater than 100°C because of the relatively large heat losses of simple flat plate collectors. Although there is a considerable technical potential market for active solar systems in North European countries such as the UK, for example, and, despite the publication of several Codes or Guides to good practice, the high costs of installations have reduced the development of the market in favour of alternative conventional sources of energy supply.

27.2.2.2 Passive solar systems

In passive systems use is made of the building materials and design for both space heating and cooling. By combining energy efficiency measures with passive solar design techniques for energy collection, storage and distribution energy costs in buildings can be greatly reduced from present day standards.[4,5] Several categories of passive systems exist including

- the 'direct gain' approach using windows as solar collectors,
- the Trombe or thermal storage wall in which the heat is stored in the wall which also absorbs the solar energy passing through the glazing,
- the solar greenhouse combining these two approaches whereby glazing, i.e. a greenhouse, is added to the outside

of a thermal storage wall facing south (in the northern hemisphere),

- the roof pond system in which a shallow pond or tank with moveable insulation is located on and covers the roof affording both heating and cooling options depending on the time of day covered and season and
- the natural convective loop using air.

Factors to be considered in passive solar design schemes include the influence of the local climate on the optimum design, the perceived merits or otherwise measured against energy use and cost, comfort and amenity value, public acceptance and the integration of energy use into the overall design process.

27.2.2.3 Photochemical conversion

In the photochemical conversion process, irradiation of an electrode/electrolyte system results in a current flow in an external circuit. The current may be generated by a photochemical reaction in the electrolyte, or by a photosensitive electrode. Devices based on this effect, the absorption of solar photons in a molecule producing an excited state or alternatively in a semiconductor raising electrons from the valence band to the conduction band, may be used either to produce electric power directly or to produce a chemical product (photo electrolysis). The latter process stores the energy and regenerates the reactants on subsequent conversion to electricity. It is this capacity for energy storage that makes the devices based on the electrochemical effect particularly attractive for solar energy conversion. So far, however, only low efficiencies have been obtained and the process is at present almost entirely experimental. The majority of photochemical reactions are exothermic and not suitable for converting solar radiation into stored chemical energy. The known endothermic (energy storing) reactions that occur with visible light are, in theory, capable of producing valuable chemical fuels. A major problem, however, has been that most of these endothermic reactions reverse too quickly to store the energy of absorbed light.

27.2.2.4 Chromogenic materials

Most of the literature on this subject concerns small-scale electronic information display, but chromogenic materials offer also the possibility of developing advanced glazing which combines variable control of solar gain with efficient thermal insulation. Electrochromic windows are essentially electric cells comprising an electrochromic layer and a counter-electrode, or ion storage layer separated by an ion conductor and sandwiched between two transparent electronic conductors that are deposited onto transparent substrates, e.g. glass or polymeric materials. In operation a d.c. electric field is applied across the transparent conductors and ions are driven either into or out of the electrochromic layer. In addition the electrochromic layer may be caused to bleach or to colour in a reversible way under the influence of the field causing reflectance or absorptance modulation of visible and near infrared electromagnetic radiation and hence changes in the optical properties of the window.

27.2.2.5 Transparent insulation materials

Transparent (or translucent) insulation materials (TIMs) are a relatively new class of materials that combine the uses of glazing and insulation. With these materials high transmission of light and heat from the sun (solar gain), good insulation (U-value), i.e. low conduction of thermal wavelengths, and strong convection suppressant characteristics are possible.

27.3 Marine energy[10–16]

The main potential sources of marine energy are tides, marine currents caused by tidal movement but including other effects as well, waves and ocean thermal energy.

27.3.1 Tidal energy

There are two types of technology involved in extracting energy from the tides, the first capturing and subsequently using the potential energy within a storage basin, the second converting the kinetic energy of tidal streams. The concept of installing turbines in a barrage that encloses inlets or estuaries is well established, whereas the use of tidal streams, although considered for some time, has recently seen a marked increase in interest through comprehensive studies and progress towards demonstration projects.[10]

The sea level varies approximately sinusoidally with a 12.4-hour period, the diurnal ebb and flow cycle, superimposed on a longer sinusoid with a period of 353 h, the spring-neap cycle. When the sun and the moon are almost in line with the earth the tides have their maximum amplitude and are known as the spring tides; when the moon–earth–sun angle is a right angle the tides are at minimum amplitude and are known as the neap tides. The ratio between the amplitudes of the maximum spring tide and the minimum neap tide can be up to as much as 3 to 1. Smaller seasonal variations also occur. The peak-to-peak amplitude of the tidal variation is known as the tidal range. In mid-ocean this range is about 1 m, but it is often amplified in coastal areas by a complex interaction with coastal features. The greatest amplitude occurs in estuaries where the tide is in resonance condition with the advancing tidal wave interacting with the reflected waves from the side of the estuary.

27.3.1.1 Tidal barrages

The simplest method of extracting energy from the tide is to build a dam (barrage) across an estuary or inlet, the dam containing turbines to generate electricity. In its most basic form the rising ‘flood’ tide enters the basin through gated openings or sluices and through the turbines idling in reverse. At high tide all openings are closed until the tide has ebbed sufficiently to develop a useful head across the barrage. The turbines are then opened and generate electricity for several hours until the difference in water level between the emptying basin and the next flood tide has dropped to the minimum at which the turbines can operate. Shortly afterwards the levels will be equal, the sluices are opened and the cycle repeats. There are alternative methods of operation, the main options being generating at ebb tide only with or without using the turbines also as pumps at high tide to raise the level of water in the basin and generating during both flood and ebb tides. Studies have shown that the method of operation that results in the lowest unit cost of energy is either simple ebb generation or ebb generation with pumping at high tide. With ebb generation electricity is produced for 5–6 h during spring tides and about 3 h during neap tides out of a tidal cycle lasting approximately 12.4 h; thus a tidal barrage produces two blocks of energy each day, the size and timing of which follows the lunar cycle. As the generation period is about 1 h later each day, the generation (and pumping if used) needs

Table 27.2 Mean tidal range in selected locations[11]

Location	*Range* (m)
Bay of Fundy Canada	10.8
Severn Estuary, UK	8.8
Rance Estuary, France	8.45
Passamaquoddy bay, USA	5.46
Solway Firth, UK	5.1

to be planned in advance to integrate with the demand and supply of the grid.

Assessments of technical and economic feasibility of tidal barrages are site specific. Some locations are particularly favourable for large tidal schemes because of the focusing and concentrating effect obtained by the bays or estuaries. Typical ranges are shown in *Table 27.2* which includes the world's largest tidal range in the Bay of Fundy and Europe's largest, the Severn Estuary.

Other sites representing important energy resources include Alaska (Cook Inlet), Argentine (San Jose), Australia (north west coast), Brazil (north coast), China (Yellow Sea), France (Iles de Chausee), India (Gulf of Cambay, Gulf of Katchch), South Korea (west coast), Russia (Okhotsk Sea, Jugursk Bay).[12]

The largest scheme in operation is the 240 MW barrage at St Malo in the Rance Estuary France. Work on the Rance site commenced in June 1960, the final closure against the sea in July 1963 and the last of the 24 × 10 MW turbines being commissioned in November 1967. The overall length of the barrage is 700 m. Tides follow a two-week cycle throughout the year. During the first week the tidal range is between 9 m and 12 m and in the second week between 5 m and 9 m. For the lower ranges electricity is generated on the ebb tide with the basin level increased by pumping while for the higher ranges the electricity is generated during both ebb and flood tides, sometimes augmented by pumping. The output is computer controlled and optimised to match the needs of the French national grid. The nominal average output of between 50 and 65 MW is thus not the maximum that could be obtained, but contributes maximum savings to the grid.

While La Rance electricity is the cheapest electricity on the French national grid Electricité de France say that it would be too expensive to build any further power stations. The same conclusion applies to the proposed Severn Barrage scheme in the UK. This larger project would involve a single basin ebb generation scheme with a 13 km barrage. The estimated total installed capacity would be 8640 MW obtained from 216 turbine generators, each 9 m in diameter and rated at 40 MW with an annual output of about 17 TWh. The nominal lifetime is taken to be 120 years, but in practice with maintenance and turbine replacement as necessary is indefinite. Elsewhere in the UK the second largest project would be the Mersey Barrage. Here the installed capacity would be only 600 MW, but there would be substantial benefits to the local economy. Overall in the UK the theoretical tidal barrage capacity is considered to be approximately 25 GW.

Tidal barrage schemes have a large impact on an estuary. With a reduced tidal range above the barrage there are potential changes to land drainage, fish migration, navigation of ships, wading bird activities plus the possible increased sedimentation due to altered channel turbulence and flow in the estuary. Several advantages follow from building a barrage, however, through the provision of a route for road traffic across an estuary and with the reduced tidal range come increased opportunities for recreational sailing and water sports.

27.3.1.2 Marine currents[10]

Relatively rapid marine currents exist at locations where natural marine flows occur through constraining channels such as straits between islands, shallows between open seas and around the ends of headlands. Marine currents are driven primarily by the tides, but also to a lesser extent by coriolis forces due to the earth's rotation, salinity and temperature differences between sea areas.

Studies of the European resources showed 106 locations suitable for exploitation with a total capacity of some 12 GW yielding about 48 TWh per annum. Other large resources exist in SE Asia, Canada, the SE coast of South Africa and elsewhere. Typical velocities at peak spring tides are in the region of 2 to 3 m/s or more. The main requirements are fast flowing water, a relatively uniform seabed to minimise turbulence, sufficient depth of water to allow large enough turbines to be installed, such conditions to extend over as wide an area as possible to allow the installation of enough turbines to make the project cost effective, free from shipping constraints and near enough to a shore-based electricity supply network capable of taking the power delivered.

The extraction of energy from marine currents by means of propeller turbine rotors is governed by the same equations as for wind turbines, thus the power theoretically available from a stream of water through a turbine is

$$P = 0.5 d_{\mathrm{w}} A V^3$$

where P is the power, d_{w} the density of water, A the area swept by the rotor blades and V the stream velocity. The power density of water compared to that of wind may be seen from the *Table 27.3* for different velocities assuming a density for salt water of 1030 kg/m^3 and an air density of 1.2473 kg/m^3 corresponding to air at 10°C.

The maximum amount of energy that can be extracted is 16/27 or 0.59259 of the theoretically available energy (the Betz limit) and, as for wind turbines, this efficiency can only be approached by careful blade design.

In contrast to wind turbines of similar output the high power densities achieved with streams of flowing water at the velocities encountered mean that large thrust forces are applied to marine turbines. *Figure 27.1* illustrates the power output and speed of a marine turbine for different rotor sizes assuming a stream velocity of 2 m/s and a conversion efficiency of 30%[10] with the possible overload torque and thrust loads encountered at the higher velocity of 4 m/s.

Table 27.3 Relative power densities of marine currents and air at different velocities

Velocity m/s	*Power density marine* kW/m^2	*Power density wind* kW/m^2
1	0.52	—
2	4.12	—
3	13.91	0.2
10	—	0.62
15	—	2.10
20	—	4.99

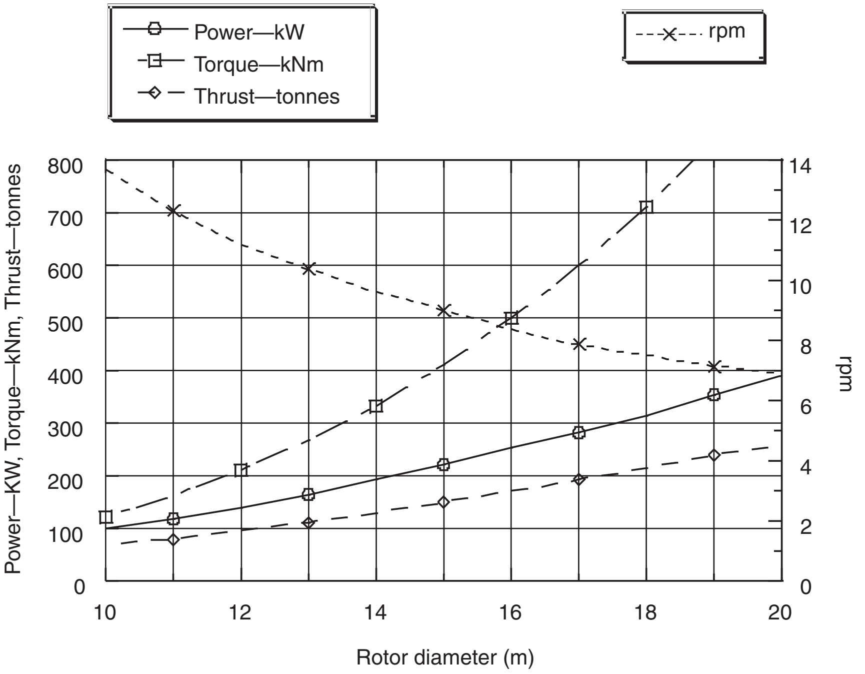


Figure 27.1 Typical performance characteristics of marine turbines in 2 m/s tidal stream and 30% efficiency

The high axial thrust requires the turbine to be attached to a structure which is either anchored firmly to the seabed via gravity based or piled platform structures or floated beneath a vessel held by high tension moorings. Various designs are possible for the power train linking the horizontally mounted turbine to the generator from which output is delivered via a marine cable laid across the seabed to the shore at voltages of 11 or 33 kV.

27.3.2 Wave energy

27.3.2.1 Resources

The oceans act as an integrator of wind energy. Waves arriving at any point can have originated from storms many hundreds of kilometres distant, the 'swell' sea, or from local wind conditions, the 'wind' sea. The distance from the origin of the swell waves is known as the fetch. As a general rule coastlines with an ocean fetch of greater than 400 km are suitable sites for recovering wave energy with the greatest resources being available between latitudes 30° and 60° in the Northern and Southern hemispheres.

In the UK alone it has been estimated that the recoverable wave energy resource exceeds total UK electricity demand. Estimates of the total power available for wave energy systems in the UK in the mid-1970s were based on data gathered from the weathership *India* (50°N, 19°W). Translating this data into wave power likely to be obtained from shoreline devices gave the analysis summarised in *Table 27.4*.

Measurements reported by the then CEGB in 1983 at inshore sites in the UK showed power levels between 40 and 50 kW m^{-1} of wave front in water about 50 m deep

Table 27.4 Achievable wave power resource in the UK[13]

TOTAL RESOURCE Estimated in 1974 by assuming weathership data of 80 kW/m mean annual output applied to 1500 km of UK coastline	120 GW
GEOGRAPHICAL LIMITATIONS Land masses prevent formation of energetic waves from easterly direction and Eire screening part of UK coastline	−72 GW
DEVICE CONFIGURATION LIMITATIONS Waves from different directions are not absorbed with equal efficiency by a line of devices. Directionality factor = 0.76	−12 GW
STATION DESIGN LIMITATIONS Some sites are not suitable: devices must be spaced apart and permit navigation. Device/space ration = 0.75	−9 GW
DEVICE CAPTURE LIMITATIONS Efficiency of power absorption varies with wavelength. Some power is rejected. Overall efficiency = 40%	−16 GW
POWER TRAIN LIMITATIONS Losses due to efficiencies of turbine, generator, transmission system. Overall efficiency = 50%	−5 GW
AVAILABLE RESOURCE May be subject to further limitations both environmental and economic	6 GW

near South Uist and 25 kW m^{-1} off the north-east coasts of England or south-west Wales.

Further assessments made in 1988[14] gave a larger technical potential of 30 GW capable of providing some 50 TWh yr^{-1}, mainly off the Western Isles of Scotland and the coast of Cornwall. Such estimates have to be moderated by the lack of electrical transmission network infrastructure in NW Scotland and between Scotland and England that need to be installed if wave power is to contribute significantly to the UK demand for electrical energy.

27.3.2.2 Device design

Over three hundred ideas have been considered in the UK Department of Energy programme.[13] In general three main approaches to capturing wave energy are as follows:

(i) *Floating or pitching devices* These devices generate electricity from the bobbing or pitching action of a floating object. Examples include tethered buoy structures where the rise and fall is restrained by the mooring with energy extracted from a pump in the mooring. Alternatively the device can be mounted to a floating raft or fixed to the sea floor where energy is extracted from the relative motion between inner and outer parts of converters mounted on a common frame or spine across the wave front or to hinged wave-contour structures perpendicular to the wave front where the energy is extracted from the relative motion of adjacent sections.

(ii) *Oscillating water columns (OWC)* These devices generate electricity from airflow caused by the wave-driven rise and fall of water in wave energy collectors which are in the form of a partially submerged shell into which seawater is free to enter and leave. As the water enters or leaves a column of air, contained above the water level, is alternately compressed and decompressed to generate an alternating stream of high velocity air in an exit blowhole. If this air stream is allowed to flow to and from the atmosphere via a pneumatic turbine, energy can be extracted from the system and used to generate electricity. Designs exist for 2 MW near shore gravity anchored wave stations designed for regional power generation and coastal protection and for larger 3.5 MW near shore combined wave and wind stations which can be constructed also in multiple units when larger quantities of electricity are required.[16]

(iii) *Wave surge or focusing devices* These devices rely either on naturally occurring 'tapered channel' gullies in shorelines or on a shore-mounted structure to channel and concentrate the waves. 'Overtopping' schemes as in the 'Tapchan' system in Norway, use the enhanced height of the wave action to cause water to flow over a dam where it is stored and allowed to run out through a turbine when needed. An alternative approach is to combine the tapered channel approach with the oscillating water column. Such a scheme is the first commercial wave power station on the Scottish island of Islay based on the LIMPET 500, a 0.5 MW shoreline wave power station.

27.3.2.3 Mooring and anchoring

The provision of storm resistant mooring and anchoring for floating wave energy devices especially in deep water presents difficult technical problems which have not been readily solved. In general the cost of mooring and anchoring, or the provision of seabed attachment, together with the cost of the initial device installation would be a significant proportion of the overall capital costs. It could range from 10 to 15% for floating devices and possibly up to 30% for devices fixed to the seabed. The latter systems would be virtually maintenance free whereas moorings require periodic inspection and replacement, thereby incurring an extra operational cost.

27.3.2.4 Power conversion and transmission

Various hydraulic and mechanical systems have been proposed, but the largest numbers of designs use air as a working fluid. Either the flow of air is rectified by valves, or it flows backwards and forwards through a turbine such as the Wells turbine. Wells turbines have the property of turning in the same direction regardless of which way the air is flowing across the turbine blades. No large wave power scheme comprising many devices exists at present, but in principle the generation of electricity by means of a number of small alternators can be aggregated by a d.c. system for transmission to shore and thence inverted by power electronics to the electricity supply network requirements.[12,13]

27.3.2.5 Environmental acceptability

Wave power development has not yet progressed sufficiently to encounter institutionalised environmental opposition. The establishment of large schemes offshore, however, would raise questions for the shipping and fishing industries. In addition changes to shoreline wave environments and noise, particularly from the oscillating water column devices, would have to be considered unless the installations were suitable distances from habitation.

27.3.2.6 Economics

To date the costs of prospective wave power schemes has led to unfavourable comparisons with alternative ways of harnessing renewable energy such as wind. Wave power generators are as yet only likely to be useful at suitable coastal sites, especially on remote islands where the costs of conventional diesel generated power is high. In addition for coastal sites in areas of outstanding natural beauty they will have less environmental impact than large wind turbines and thus may be more readily acceptable.

27.4 Hydro

The technology for the use of hydropower for the direct generation of electricity has developed from wooden water wheels (overshot, undershot and crossflow) through to the Francis, Kaplan, Pelton, Crossflow and Turgo turbines of the 20th century. In developed countries the larger hydro resources were among the first to be exploited. Attention has focused more recently on smaller resources which are generally classified as shown in *Table 27.5*[3] (other size classifications are also found in the literature).

Table 27.5 Definition of hydro scheme size

Large	50 MW and above
Small	5 MW to 50 MW
Mini	500 kW to 5 MW
Micro	500 kW and below

By the 1980s the country with the greatest experience in small-scale hydropower development was the People's Republic of China where nearly 100 000 plants had been constructed in the previous twenty years.

With a mature technology the investigation, design and construction of conventional large-scale hydroelectric schemes is well defined and understood and can be costed with reasonable certainty. For small schemes at the mini and micro scale, however, economic constraints demand an innovative approach. Novel low head designs for prime-movers together with a trend towards the use of off-the shelf components and plastics for small impeller-type turbines, the use of micro and power electronics in generation and again the use of plastics for pipelines in the civil engineering works all contribute to extending the boundaries of economic viability.

The distinctive features of small hydropower plants include being usually run-of-river type, construction in a relatively short period of time, using well developed technology with an overall efficiency of over 80% and having automatic operating systems with low operation and maintenance costs.

As for all renewable resources the size of an exploitable resource depends on both technical and economic factors. For hydro the broad categories are

- the gross river potential which is approximately the summation of (annual run-off × potential head),
- the exploitable technical potential which is the gross river potential less the potential which is technically impossible to develop,
- the economic potential which is the technical potential less the potential which is uneconomic to develop, and more recently in the history of this technology,
- the environmentally acceptable potential which is the economic potential less the potential which is considered environmentally unacceptable to develop.

Without significant storage capacity large variations in available water flow may be experienced. In the UK the capacity factor for hydro, i.e. the ratio of actual annual energy generated to energy produced at rated output over twelve months, is approximately 30% which is nearly the same as for wind energy. Small-scale hydro-schemes with turbines having sufficient rotational velocity can employ either induction or synchronous generators. Low-head run-of-river turbines run more slowly and so need either a gearbox or a large multiple pole generator for energy conversion.

A more detailed discussion of hydroelectric plant is given in Section 26.4.

27.5 Wind[17–25]

27.5.1 Wind energy

The annual energy available from a wind turbine in any particular location depends on the wind speed at hub height, which in turn depends on the shape of the local landscape, the height of the turbine above the ground and the annual climatic cycle. An empirical relationship between mean wind velocity V and turbine height H is $V = H^a$ where a has a value of 0.13 in the UK for open, level ground, rising to 0.25 for an urban site and to 0.33 for a city site.[8] An ideal site is a long, gently sloping hill.

Offshore wind speeds are generally higher than on land, e.g. ten kilometres from the shore speeds are typically 1 m/s higher than on land. Although wind/wave interactions exist, turbulence is lower which reduces the fatigue loading on the turbine blades. It is offshore where the very large wind-farms of several tens or even hundreds of MW in size are anticipated in the future.

27.5.2 Wind turbines

The theoretical power in an air stream is 0.5 $d_a A V^3$ where d_a is the density of air, A the cross-sectional area and V the velocity. The actual power P extracted by a wind turbine, however, is of the same form as for water turbines

$$P = C(0.5 d_a A V^3)$$

where C is a coefficient of performance or power coefficient. The German Engineer Betz showed in 1927 that the maximum power extracted from a moving air stream to be 16/27 or 0.59259 of the theoretically available power. This efficiency can only be approached by careful blade design with blade tip speeds a factor of six times wind velocity and is known as the Betz limit. Modern designs of wind turbines for electricity generation operate with a power coefficient (C) of about 0.4, with the major losses caused by drag on the blades and the swirl imparted to the air flow by the preceding rotating blades. Any wind turbine will operate only between a minimum starting wind velocity value, V_S, and its rated value V_R. Typically the ratio V_R/V_S is between 2 and 3, although if the pitch of the blades can be altered at velocities greater than V_R then the turbine should continue to operate at its rated output, the upper limit being set by design limitations. Depending on the location the wind speed may be less than V_S for 25% of the year when the turbine is shut down and the annual load factor, the ratio of energy produced to the energy that would be produced if run at maximum rated output over the whole year, is typically between 25% and 35%. A typical operating power characteristic is shown in *Figure 27.2*.

In general machines are designed to operate with a peak output in the range 250–500 W/m^2; thus 20 m machines have ratings of around 200 kW, 30 m machines around 300 kW and 50 m machines around 1 MW. Power limitation is accomplished either by using feathered blades in larger machines, with feathering along the whole blade length, or in smaller machines by taking advantage of the natural tendency of blades to stall as the angle of attack increases in high winds. Current commercial wind turbines tend to operate at a fixed speed with tip speeds of around 50–80 m/s for power generation and slower speeds for high torque applications such as pumping.

The technical options available to designers of wind turbines and the interaction of option choices in the determination of machine weight and cost are discussed.[17] This work examines the extensive debate surrounding the optimal size of machines and the engineering implications of a move from heavy, stiff designs to more lightweight, compliant designs. Typical turbine characteristics are summarised in *Table 27.6*.[21]

27.5.3 Wind generators

With wind speed fluctuations being inevitable it is important to damp out the consequent driving torque oscillations in the generator. For network-connected fixed speed turbines synchronous generators do not provide adequate damping which must be supplied elsewhere in the transmission otherwise power fluctuations and blade loads may be unacceptable; hence most use 4- or 6-pole induction generators. Induction generators provide inherent damping where

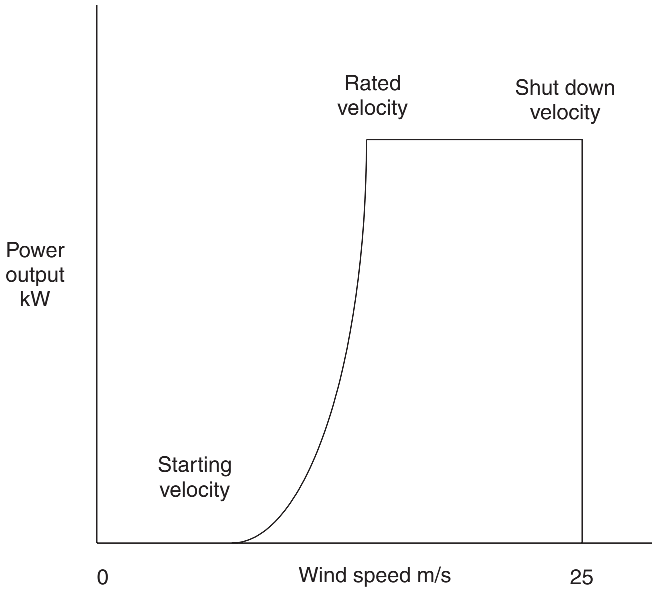


Figure 27.2 Typical wind turbine power/wind-speed operating characteristic

Table 27.6 Typical wind turbine characteristics

Ratings	1 to 2 MW now available and increasing
Availability	98–99%, a mature reliable technology
Rotor diameter	Up to 80 m, larger diameters to follow
Number of blades	Majority now three, reducing percentage of two or even one
Blade material	Glass-reinforced plastic or wood-epoxy
Rotor orientation	Mostly upwind of tower, some downwind
Rotational speed	Usually constant, c. 25 rpm at 52 m diameter, some two speed and variable speed
Power control	Stall control in high winds with fixed blades, pitch control where all or part of blades rotate to limit power
Power train	Step-up gear boxes most common, direct drive without gearbox now also used
Yaw control	Wind direction sensors linked to powered rotor alignment, some passive yaw control
Towers	Cylindrical steel construction, concrete towers used for some large machines

the damping is provided by the slip speed difference between the rotor and stator rotating mmf, but suffer from the disadvantage of drawing reactive power from the system and also from drawing high starting currents which may cause local flicker. Typically the induction generator is wound for 690 V, 1500 or 1000 rpm operation and generally of squirrel-cage construction. Connection to the network is via power-factor correction capacitors and power electronic converters to accommodate starting conditions or, in the case of variable speed operation, to decouple the speed of the rotor from the frequency of the network.

27.5.4 Economics

Onshore wind farms are relatively competitive. Costs have fallen from 8.6 pence per kWh in 1992 to an average of 2.88 pence per kWh for the new farms coming on line now. They are among the most competitive renewable energy plants and are less expensive than some new coal plants. Electricity from offshore wind farms remains uncompetitive, however, at present in comparison with electrical energy provided by conventional power sources.

The estimate of capital costs of wind farms in the USA is \$983/kW (1999 prices).[22] The same source gives a typical capacity factor at 32% although the average for UK wind farms is approximately 24%[23] as shown in *Figure 27.3* which illustrates typical monthly variation in capacity.

European installed costs are quoted at

- Euros 875/kW[24] onshore and Euros 1600/kW offshore,[25]
- 15 year depreciation, 7% discount rate
- O and M costs: Euros 20/kW/yr + Euros 0.004/kWh.

Using these figures *Table 27.7* illustrates the calculation of the generation costs for a 50 MW wind farm assuming a capacity factor of 28% giving a total cost of Euros 0.051/kWh or 3.07 p/kWh.

Figure 27.4 shows the effect of different capital recovery periods where for short pay-back times the costs are prohibitive and the value of longer term guaranteed support over fifteen years can be seen. With significant differences in generation costs being measured in tenths of p/kWh the importance of capacity factors, i.e. choice of site, turbine hub height, etc., can also be appreciated.

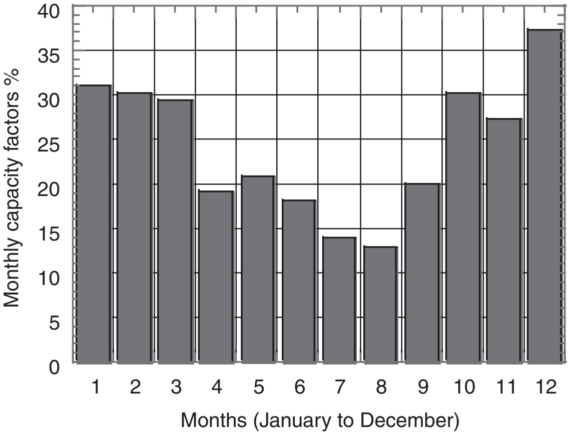


Figure 27.3 Wind farm typical monthly capacity factors (Annual average 24%)

Table 27.7 Generation cost estimates

Capital cost	Euros/kW	875
Total capital cost, ***P***	Euros	875 × 50 000
Annual energy generated, ***E***		
at 28% capacity factor	kWh/yr	122 640 000
Discount rate	%	7
Capital recovery period	years	15
R (see section 27.3)		0.110
Capital cost ***A***/***E*** = ***P*** × ***R***/***E***		
for 24% capacity factor	Euros/kWh	0.039
O & M costs		
maintenance	Euros/kWh	0.008
operation	Euros/kWh	0.004
TOTAL generating cost	Euros/kWh	0.051

27.5.5 Environmental issues

Wind energy is a non-polluting form of energy fulfilling the aims of present governments to replace polluting and non-renewable hydrocarbon-generated electrical power with power generated from cleaner forms of energy. In the course of the development of the industry, however, other forms of pollution have been noted and been the object of much debate and hostility. The principle environmental issues to be considered in deploying wind farms are:

- Visual and landscape impact;
- Noise;
- Electromagnetic interference;
- Effect on birds and wildlife; and
- Land use.

The same considerations apply, however, to all other forms of renewable energy exploitation. Wind energy has been in the forefront of development, deployment and commercialisation so has been the first renewable industry to meet the environmental objections and constraints.

Chief among these effects are visual pollution and noise. Visual pollution, although subjective, is easy to comprehend arising simply from the juxtaposition of industrial plant in areas of outstanding natural beauty, i.e. unspoiled countryside. Other wind-turbine sites in ‘brown field’ sites in or near urban areas where industry has been before have not created the same opposition and indeed have been taken as symbols of regeneration.

Noise is more of a problem inasmuch as there appears to be some evidence of low frequency (infrasound) noise effects near some sites, but the data is sparse. Early problems of noisy gears have largely abated with better design and manufacture. The level of noise of the aerodynamically generated swishing noise of the turbine blades, however, varies from site to site and can be negligible. Particularly strange effects have been reported from Montgomeryshire in Wales where the existence of steep nearby valleys seems to act in some cases as a means of focusing noise at different places away from the wind farm sites.[18,19]

Offshore wind farms will have less of these problems, but for on-shore sites local objections have meant that the majority of planning applications in the UK have been refused so it is vitally important that projects are properly and sensitively integrated into the landscape and developed in consultation with local communities.[20]

27.6 Geothermal energy

The Earth is an almost infinite source of heat with a continuous heat flow of some 10^{13} J/s. Unlike the other renewable

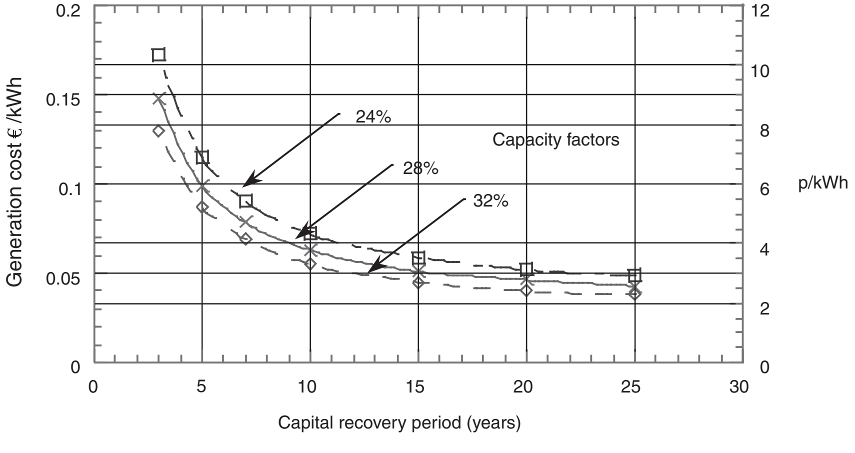


Figure 27.4 The influence of capital recovery periods and capacity factors on wind generation costs

resources, which are intermittent, geothermal energy can deliver a constant source of power until the reservoir is exhausted. There are two possible techniques for exploiting geothermal energy:

(1) use of existing hydrothermal sources;
(2) to attempt to extract heat from hot dry impermeable rocks deep below the surface.

The first technique is now commercially established, whilst the second is still at the research and development stage.

27.6.1 Hydrothermal sources

Although steam and hot water come naturally to the surface of the earth in some locations, for large-scale use boreholes are normally sunk with depths of up to 3 km, releasing steam and water at temperatures of 200–300°C and pressures of up to 3000 kN/m^2.

Flowing well-head steam pressures vary between 200 and 1500 kN/m^2. From the well heads the steam is often transmitted by pipelines of up to 1 m diameter over distances up to about 3 km to the central power station. Water separators are usually required, as superheating the steam to minimise wetness requires fossil fuel, the use of which is usually uneconomical. Care must be taken in the choice of materials for the plant, as the steam may contain solid and gaseous impurities. Steam may come from several bore holes at different pressures, so that the steam layout of the generating station may involve two or three different pressures in cascade. If water quality is satisfactory, direct use of the steam in turbines is possible. In most cases jet condensers are used, with the condensate being discharged to waste along with the cooling water. The only major electrical plant required in the generating station is a cooling-water pump. The thermodynamic efficiency of the power station at these low temperatures and pressures is about 10–15%. Thus large quantities of exhaust heat are available for local use or, alternatively, must be discharged as waste heat.

Heat from aquifers at lower temperatures is unsuitable for electricity generation but may be used for district heating or agricultural purposes. Several such aquifers with temperatures in the range 60–80°C exist.

Recent examples of district heating schemes exist in Southampton where the City Council is exploiting the heat from the Wessex Basin for space heating of Council office buildings. Several schemes are now operating in the Paris and Aquitaine Basins.

Power generation from geothermal aquifers began in 1904 in Lardarello, Italy, where installed capacity now exceeds 400 MW. Similar power-generation developments have taken place in China, El Salvador, Greece, Hawaii, Iceland, Indonesia, Japan, Kenya, Mexico, New Zealand, Nicaragua, the Philippines, Turkey, the CIS and the USA. Overall, it is estimated that some 8 GW of power is now being produced world-wide from geothermal aquifers.

At first sight geothermal energy would appear environmentally benign. However, no system of power generation is without some adverse environmental effect. Conventional geothermal power stations involve surface pipelines bringing steam from widely distributed bore holes to central power stations, and these can be unsightly and lead to visual intrusion. Geothermal aquifers incorporate gaseous components such as carbon dioxide, hydrogen sulphide, methane and radon. Such discharges need to be considered at the planning stage for new geothermal plant, otherwise environmental legislation may inhibit further expansion of the geothermal resource. Power generation from geothermal aquifers in volcanic areas can provide cheap electric power. Indeed, the generation costs are often quoted as being comparable with those of hydroelectric stations.

27.6.2 Hot dry rocks

The average thermal gradient near the surface of the earth is about 25°C/km. This temperature gradient is exceeded in some granites because of the local heat generated by the decay of radioactive elements. For example, in New Mexico granites have been identified with a thermal gradient of 50–60°C/km and in Cornwall gradients of 30–40°C/km have been measured in the Carmenellis granites.

In order to exploit these elevated subsurface temperatures for power generation, a temperature of greater than about 200°C is required. Thus wells of depths between 4 and 7 km are required. Current drilling technology sets a practical limit of about 6 km.

Two sets of field trials have been carried out in an attempt to exploit this enhanced heat source in granites. The first was conducted by the University of California at the Los Alamos Laboratories in New Mexico; the second was done by the Cambourne School of Mines at Rosemanowes quarry in Cornwall. The trials involve drilling a bore hole deep into the granite, fracturing the rock at the base of the bore hole using hydrostatic pressure and explosives, and then drilling a second bore hole to intercept the fractured zone. Water can then be circulated down one bore hole, permeate through the fractured zone, and hot water can rise to the surface through the second bore hole.

There are a number of practical difficulties associated with this scheme. The drilling of a second bore hole that is located sufficiently accurately to intercept the fractured zone is a problem in itself. Having created a circuit for water circulation, it has proved difficult to manage the flow of water such that a large surface area of hot rock is contacted by the circulating fluid. In practice, circulation through larger channels has tended to dominate and, therefore, depress the temperature rise of the circulating fluid. A further problem is the loss of circulating fluid due to subterranean leakage and the requirement for large volumes of cooling water for a process the overall thermodynamic efficiency of which is about 5%. Environmental problems from hot dry rocks include the release of radon gas, visual intrusion from cooling towers, and the stimulation of minor seismic events associated with the creation of the reservoir.

27.7 Biofuels

27.7.1 Introduction

Biomass and biofuels have no strict definition, but include agricultural residues, energy crops, municipal solid waste (MSW) and landfill gas.[26] It is convenient to differentiate between biofuels arising from agricultural sources and those which arise from human, urban and industrial processes. *Table 27.8* compares solid and gaseous fuels used for power generation according to this distinction[27] and *Table 27.9* lists relative energy content of the different materials used as fuel.

27.7.2 Biomass technologies

27.7.2.1 Overview

Energy can be recovered from biomass in a variety of ways including aerobic and anaerobic digestion, combustion with heat recovery and gasification/pyrolysis processes.

Table 27.8 Renewable and waste fuel resources

Agricultural resources	*Waste resources*
Sugar cane waste (bagasse)	Sewage Digester Gas (SDG)
Timber mill waste or sawdust	Landfill Gas (LFG)
Forestry residues	Mines gas
Short-rotation coppicing (SRC)	Coke-oven gas (COG)
Straw	Refinery and process plant flare/off gas
Rice husks and coffee husks	Stripped crude gas
Peanut and other nut shells	Municipal solid waste (MSW)
Palm oil and coconut residues	Hazardous and chemical waste
Meat and bone meal (MBM)	Hospital and clinical waste
Poultry litter	Sewage sludge
Livestock slurry	Vehicle tyres

Table 27.9 Calorific values for different materials used as fuels

Material used as fuel	*Calorific value* (MJ/kg)
Coal	23–32
Fuel oil	40–45
Natural gas	50–55
Plastic	27–34
Municipal solid waste	8.5–11
Hospital and clinical waste	17.5–22.5
Chemical waste	18.5–23
Sewage sludge	7–13 (depending on dryness)
Vehicle tyres	32–40
Sugar cane bagasse	8–12.5 (depending on dryness)
Wood	17–20
Rice husks, Rice straw	12–18
Straw	14–15.5
Meat and bone meal	20–28 (depending on fat content)
Poultry litter	13–14

The main bio-energy conversion routes are represented in *Table 27.10* along with the principal products formed.

With the exception of sewage sludge and animal slurries most biomass materials are not in a form well suited to digestion processes. Commercially available digester plants have a relatively limited capacity, typically up to about 1 MW of biogas; hence for larger commercial applications the thermal treatment technologies for solid biomass are presently of more significance. Much work is underway at present, however, in the development of enzyme treatments to enable fermentation of cheap agricultural feedstock such as corn leaves and wheat straw into ethanol. A demonstration plant consuming 40 t of straw per day is reported as producing up to 4 m litres of ethanol/year. A car fuelled by a mixture containing 85% ethanol produces 91% less greenhouse gas emissions compared with petrol and so, given the present focus on reducing such emissions worldwide, this particular biomass transformation should feature more prominently in the future.

27.7.2.2 Pyrolysis and gasification

Pyrolysis is the thermal degradation of biomass in the absence of oxygen, but this may include partial gasification. Three products result: a solid char residue, gas and a complex, oxygenated hydrocarbon liquid containing water with the relative yields optimised according to the requirements and process control. The pyrolysis liquids can be burned in boilers, dual fuel diesel engine and turbines. Typically the pyrolysis gas is vented because of its low concentration of combustible gases, but high temperature pyrolysis produces a fuel gas that could be used in an engine or turbine.

Thermo-chemical gasification is the conversion by partial oxidation at elevated temperature of a carbonaceous feedstock such as biomass into a gas containing CO, CO_2, H_2, CH_4, trace amounts of higher hydrocarbons such as C_2H_2 and C_2H_6, water, nitrogen (if air is used as the oxidising agent) and various ash, oil and tar contaminants. The partial oxidation can be carried using air, oxygen, steam or a mixture of these. Air gasification produces a low heating value gas (up to 4–7 MJ/Nm^3) suitable for boiler, engine and turbine operation but not for pipeline transportation because of its low energy density; oxygen gasification produces a medium heating value gas (up to 10–18 MJ/Nm^3)

Table 27.10 Bio-conversion processes and products

Process type	*Process*	*Initial product*	*Final product*
Aqueous	Anaerobic digestion	Biogas (2 parts CH_4 to one part CO_2, 22–28 MJ/m^3)	Methane (38 MJ/m^3)
	Alcohol fermentation		Ethanol (19 MJ/litre)
	Chemical reduction		Oils (35–40 MJ/kg)
Dry thermo-chemical	Pyrolysis		Pyrolytic oils (23–30 MJ/kg) Gas (8–15 MJ/m^3) Char (19–31.5 MJ/kg)
	Gasification	Low/Medium energy gas (7–15 MJ/m^3)	Methane (38 MJ/m^3) Methanol (16.9 MJ/litre) Ammonia Electricity (3.6 MJ/kWh)
	Hydrogasification		Methane (38 MJ/m^3) Ethane (70.5 MJ/m^3) Char (19–31.5 MJ/kg)
Direct combustion	of various fuels, e.g. wood chips (17–20 MJ/kg dry weight)	High pressure steam	High pressure steam Electricity (3.6 MJ/kWh)

suitable for limited pipeline distribution and as synthesis gas for conversion to methanol and gasoline.[34]

27.7.2.3 Direct combustion

The direct combustion of organic matter to produce steam or electricity is the most advanced of these conversion processes and, when carried out under controlled conditions is probably the most efficient. Waterwall incineration whereby water pipes within the walls of the incinerator are heated produces steam at high efficiencies for both electricity generation and CHP purposes. The moving grate combustion technology is well proven for coal firing and can accept many forms of biomass including relatively coarse solid materials, but has limited ability to handle wet sludges. Fluidised bed (FB) technology is also well established and proven for biomass combustion. It features intimate mixing of gases and high-temperature solids and allows very close control of temperature and reaction stoichiometry. The major advantages of fluidised beds are:

- Greater fuel flexibility—the ability to combust a variety of fuels with a wide ranging characteristics such as very-high ash fuels, high moisture fuels such as bark and sludges and high grade coal fuels.
- Lower temperature combustion without decrease in combustion efficiency thus inhibiting the formation of NO_x emissions.
- The ability to fire fuels with low melting point ash.

There are two types of fluidised bed technology in common use in Europe, the bubbling fluidised bed (BFB) and the circulating fluidised bed (CFB). BFB units are available up to 100 MWe and have been proven on biomass and waste materials with homogeneous characteristics, especially lower capacity units starting from 5 MWth using well pre-prepared fuel. CFB units are available up to 400–600 MWe and have been demonstrated on up to seventy different fuel types both as single fuels and co-combusted providing increasingly cost-effective plant for combusting low-grade fuels and different wastes with low environmental impact.

27.7.3 Major biomass sources

27.7.3.1 Short rotation coppicing and forestry residues

Short rotation coppicing (SRC) consists of shrub willow planted at high density. At the end of the first year after planting the shoots are cut back to ground level to encourage a multi-stemmed form and the crop is then harvested on a three-year rotation in the winter. Fuel consumed in the harvesting represents a significant proportion of final electricity costs and care is needed to optimise fuel supply chains both from SRC and forestry residues. Studies[28,29] showed that for a base case transport distance of 56 km the energy ratio of energy content of delivered wood to energy expended in production was 26:1, with even a transport distance of 320 km showing a positive ratio of 8:1.

27.7.3.2 Municipal solid waste

Household waste (municipal solid waste) is processed by either recycling and/or composting or disposed of in landfill sites. For the UK the amount generated at present totals about 27 mt/y and increasing at 3% per year with some 85% of this total being sent for landfill (more than most other EC countries). The EC landfill directive will enforce major changes in landfill practices, however, requiring reductions in the biodegradable domestic waste sent to landfill by 2016 to be no more than 35% of its level in 1995. The development of more energy from waste schemes (EfW) is a virtual certainty.[35]

Table 27.11 EC emission control requirements for energy from waste plant

Component	*Emission to air in ng/Nm^3—dioxins in ng/Nm^3—dry gas 11% O_2*
Particulates	10
HCl	10
HF	1
SO_2	50
NO_x as NO_2	200 (plant >ς3 tph)
CO	50
VOC	10
Hg	0.05
Cd	0.05 (Cd and TI)
7 HM (heavy metal summation)	0.5
Dioxin I-TEQ ng/Nm^3	0.1

With the limitations imposed on the landfill option the major route for EfW is by mass burn incinerators. Gasification and pyrolysis of municipal solid waste has reached the large-scale pilot plant stage but their implementation is expected to be gradual and not to supersede conventional combustion processes.[35] Although the prime purpose of EfW plants is to dispose of waste, the typical plant size of 30–50 MWe output allows close integration into CHP schemes as is common in Austria, Denmark, France, Germany, Sweden, etc.

A further EC requirement is expressed in emission standards as shown in *Table 27.11*. All UK plants are being retrofitted and new ones designed to meet these requirements.

27.7.3.3 Landfill gas

Landfill gas is the general name given to the gaseous products of bacterial decomposition of organic material within landfill sites. Such sites are constructed according to strict controls to limit their environmental impact with impermeable capping and internal lining, drainage control, in situ refuse compaction and ventilation. Organic matter decomposes aerobically in the presence of oxygen to produce carbon dioxide, but exclusion of air by means of the impermeable cap ensures anaerobic decomposition to produce water, and a flammable mixture of carbon dioxide and methane. The composition of landfill gas is variable depending on many factors, the rate of production depending on the rate of refuse decomposition with a half-life for the refuse-to-gas decomposition typically in the range of 3 to 10 years depending on the physical and chemical conditions within the waste, especially moisture levels.

It should be noted that not all landfill gas can be collected, but 25–50% is currently typical of a well-engineered site. A landfill site taking around 300 tonnes of degradable waste per day for ten years could theoretically generate as much as 4.5 billion cubic metres of gas over a period of forty years.

Spark ignited gas engines have provided the most popular means for the generation of electricity from landfill gas with the turbo-charged engine offering the best compromise in terms of capital cost, efficiency, performance and maintenance. Modern gas engines have a thermal efficiency of up to 39% and, with low compressor energy consumption, about 77% more electricity can be generated from the same volume of gas than by using gas turbines, which are also not very suitable for several reasons. Capital costs of a 1 MW installation (excluding gas collection and grid connection would be about £400 K (£400/kW) with engine maintenance about 0.75 p/kWh. The most cost-effective packages comprise 1–1.3 MW units and installing multiple units provides the flexibility to cope with the gas supply variations associated with landfill site operation.

27.8 Direct conversion

Some methods of extracting electrical energy from renewable sources do not rely on a heat stage, and the limitations of Carnot efficiency are avoided. Some other methods, still largely small scale and experimental, also eliminate machinery, relying instead on direct conversion processes. Like solar cells, the direct conversion processes described below produce electricity at low direct voltage; many units need to be interconnected and coupled to inverter systems to give a.c. output. Direct conversion processes generally need to be operated at high temperatures, and difficulties are encountered in finding suitable materials to withstand such temperatures over a long lifetime.

27.8.1 Thermoelectric generators

27.8.1.1 Principles

If two dissimilar materials are joined in a loop with the two junctions maintained at different temperatures, an e.m.f. $E = \alpha\theta$ is set up around the loop, where θ is the temperature difference and α is the Seebeck coefficient (itself depending to some extent on temperature). The phenomenon, long used in thermocouples, enables generators with semiconductor junctions to supply up to 5 kW for radionavigation beacons and satellites. A useful figure of merit is $Z = \alpha^2\sigma/K$, where σ is the electrical conductivity (which should be high, to reduce I^2R loss) and K is the thermal conductivity (which should be low, to limit heat transfer between junctions).

If the thermoelectric generator works between absolute temperatures T_1 and T_2, the efficiency as a fraction of the Carnot efficiency is

$$\eta = [(1 + ZT)^{1/2} - 1]/[(1 + ZT)^{1/2} + T_2/T_1]$$

where $T = \frac{1}{2}(T_1 + T_2)$. Thus the efficiency depends on the product ZT, which is a convenient assessment for possible thermoelectric materials. In practice, no single combination maintains a high ZT over a wide temperature range: most practical designs use a series of stages with n- and p-type semiconductor junctions that have a high ZT over their relevant temperature differences.

Taking into account mechanical characteristics, stability under operating conditions and ease of fabrication, bismuth telluride appears to be one of the most suitable materials; it can be alloyed with such materials as bismuth selenide, antimony telluride, lead selenide and tin telluride to give improved properties, is suitable for temperatures up to about 180°C, and can give efficiencies up to about 5%. A silicon–germanium alloy with phosphorus and boron impurities can be used up to 1000°C and might give efficiencies up to 10%.

27.8.1.2 Practical developments

A typical thermoelectric couple could be designed to give about 0.1 V and 2 A (i.e. about 0.2 W), so that a 10 W device suitable for a navigational beacon or unattended weather station would require about 50 couples in series. Various methods have been used to provide a source of heat for the hot junction. These include small oil or gas burners, isotopic heating and solar radiation. Although some progress has been made in developing suitable materials, theoretical studies seem to show that the scope for improvement in ZT values is rather limited. For this reason, interest in thermoelectric devices has declined over the last decade.

27.8.2 Thermionic generators

In its simplest form the thermionic converter comprises a heated cathode (electron emitter) and an anode (electron collector) separated in a vacuum, the electrical output circuit being connected between the two. Heat supplied to the cathode raises the energy of its electrons to a level enabling them to escape from the surface and flow to the anode: at the anode their energy appears partially as heat (removed by cooling) and partially as electrical energy delivered to the circuit. Although the distance between anode and cathode is only about 1 mm, the negative space charge with such an arrangement hinders the passage of the electrons and must be reduced—e.g. by introducing positive ions into the inter-electrode space, caesium vapour being a valuable source of such ions. Anode materials should have a low work function (e.g. barium oxide and strontium oxide), while that of the cathode should be considerably higher, tungsten impregnated with a barium compound being a suitable material. With these materials temperatures up to 2000°C will be needed to secure, for the generator itself, efficiencies of 30–35%, although higher overall efficiency can be obtained by using the heat from the coolant. Electrical outputs of about 6 W/cm^2 of anode surface have been suggested.

Developments of thermionic generators using radioactive isotopes as the heat source have taken place for space applications. Thermionic devices, in general, do not appear to offer significant potential as power sources.

27.8.3 Magnetohydrodynamic generators

In the magnetohydrodynamic generator a partially conducting gas is heated by a fuel fired or nuclear reactor, allowed to expand through a nozzle to convert the heat energy to kinetic energy, and then passed between the poles of an electromagnet, the field of which converts some of the kinetic energy to electrical energy which can be collected from electrodes situated in the gas channel. The generator is thus not quite a direct heat-to-electricity device, as are the thermoelectric and thermionic devices, for there is an intermediate kinetic energy stage; also, largely owing to the power required for the electromagnets and other losses, it is unlikely to give a useful output unless built in sizes of 50 MW or more.

Provided that the gas is conducting and moving at right angles to the magnetic field, an e.m.f. will be set up at right angles to the direction of motion and of the magnetic field, being proportional to the velocity of the gas and to the magnetic flux density. This e.m.f. can be collected from suitable electrodes located in the gas stream and can supply power to an external circuit. The power output is proportional to the square of the velocity, to the square of the flux density and to the conductivity of the gas between the electrodes.

The field density should therefore be as high as possible, making superconducting magnets to give fields of 4–5 T, a great advantage over conventional magnets. Gas velocities up to 1000 m/s are practicable. The electrical conductivity of gases even at temperatures of 2000–3000°C is too low to give practicable powers. Ionisation must therefore be artificially increased by *seeding* the gas with an easily ionisable element such as caesium or potassium. Either must be recycled for economical operation, but caesium is so expensive and corrosive that a closed-cycle system is essential.

The possibility of using a liquid metal, sodium or potassium, is being investigated; such a fluid would have a much higher electrical conductivity but a lower velocity, the major problem being that of producing a high velocity with sufficient liquid density to give an adequate conductivity.

With gaseous conductors a complication is introduced by the Hall effect—i.e. by the fact that the current flow between the electrodes is not in the same direction as the field: the Hall angle may reach 80°,⇐ increasing with low pressures, high magnetic fields and high electron mobilities. The resulting axial component of current flow leads to inefficiency. To counter the Hall effect the number and configuration of the electrodes is more complex. In the Faraday generator a single pair of electrodes, or several pairs connected to separate load circuits, are used, but these arrangements are not appropriate for Hall angles of more than 45°.⇐ In the Hall generator use is made of the axial component by a more complex electrode arrangement in which current is collected from axially spaced electrodes; Hall angles up to 80°⇐ are appropriate with this type.

Extensive programmes of work are in progress in the USA and the CIS to develop the open-cycle fossil fired system. In this the fuel-combustion products at temperatures over 2000°C, achieved by preheating the combustion air, are seeded with potassium carbonate and passed through a magnetohydrodynamic duct. The waste gases are then used to heat a conventional steam cycle. The magnetohydrodynamic process is therefore a topping unit increasing the overall efficiency of power generation to about 45% and potentially over 50%. The potassium carbonate also combines with the sulphur in the coal to form potassium sulphate which is removed from the boiler with the ash, the sulphur being removed and the potassium recycled. The system therefore has an added advantage where sulphur emissions must be controlled, as is required in the USA. The main problems are the development and cost of a suitable gas duct and electrode system to withstand the high temperature and corrosive effect of the gas for long periods, and the effective recovery and recycling of the seed material. In addition, large air heaters and superconducting magnets are needed. The CIS has a 20 MW prototype station which has operated continuously for up to 250 h at 10 MW on natural gas. A larger unit is proposed in which the duct-life problem is avoided by having two ducts which are used alternately and regularly refurbished. The US programme is also being directed towards the evaluation of large-scale components for open-cycle magnetohydrodynamic generation, but using coal as the fuel.

27.9 Fuel cells[36–40]

27.9.1 Introduction

Fuel cells are energy conversion devices, which by combining hydrogen and oxygen into water convert chemical energy into electricity and heat. A fuel cell works much like a battery. In both batteries and fuel cells two electrodes, an anode and a cathode, are separated by an electrolyte. Whereas a storage battery contains all the substances in the electrochemical oxidation-reduction reactions involved and has, therefore, a limited capacity, a fuel cell is supplied with its reactants externally and operates continuously as long as it is supplied with fuel.[36,37]

27.9.2 Fuel cell types

There are five basic types of fuel cell being commercially developed, a classification being based on the electrolytes used. Low temperature types include the alkaline fuel cell (AFC) and the solid polymer fuel cell (SPFC), to which belong the proton exchange membrane fuel cell (PEMFC) and the direct methanol fuel cell (DMFC), the medium temperature type is the phosphoric acid fuel cell (PAFC) and the two high temperature types are the molten carbonate fuel cell (MCFC) and the solid oxide fuel cell (SOFC). An idealised schematic diagram illustrating the structure, electron and ion flow for the various types of fuel cell is shown in *Figure 27.5*.

Phosphoric acid (PAFC) and proton exchange (PEMFC) fuel cells both use acid electrolytes, alkaline (AFC) and molten carbonate (MCFC) fuel cells use liquid alkaline-based electrolytes while solid oxide (SOFC) fuel cells use a zirconia-based ceramic. The direction of the ion flow depends on whether the ion is positively or negatively charged and also determines the site of water formation and subsequent removal. *Table 27.12* summarises the different characteristics of these fuel cells

Alkaline fuel cells (AFC) The application of AFCs in space has been especially noteworthy. They are ideally suited to closed environments containing their own supplies of hydrogen and oxygen and have also been demonstrated in a variety of automotive applications. AFC performance is particularly sensitive to contaminants in the gas supplies, notably carbon dioxide, which reacts with the electrolyte to form a carbonate and reduces the conductivity. With the chemical reaction occurring at the cathode and the low operating temperature the start-up time of the AFC is very fast and the cell yields high power generation efficiencies as seen in *Table 27.12.*

Proton exchange membrane fuel cells (PEMFC) The low operating temperatures and the solid electrolyte—an acid-based ion conducting plastic membrane—also make PEMFCs suited to a wide array of uses from low and medium to high power applications. Such a list would include power tools, compressors, recreational applications in camping and boats, heat and electricity to dwellings and electricity to commercial buildings, schools and hospitals. They are most widely known at present for their potential in automotive applications where much evaluation is underway. The low temperature electrolyte requires platinum as the catalyst applied to either side of the membrane to accelerate the dissociation of hydrogen and oxygen. The hydrogen fuel stream should contain less than 10 ppm of carbon monoxide, preferably none, because carbon monoxide will bond to the platinum and poison its catalytic property leading to significantly degradation in the cell performance.

Direct methanol fuel cell (DMFC) This cell is similar to the PEMFC except that hydrogen is extracted from a methanol/water solution. This gives it an advantage inasmuch as the hydrogen is extracted by the catalyst and not by the addition of complex reforming plant. In addition

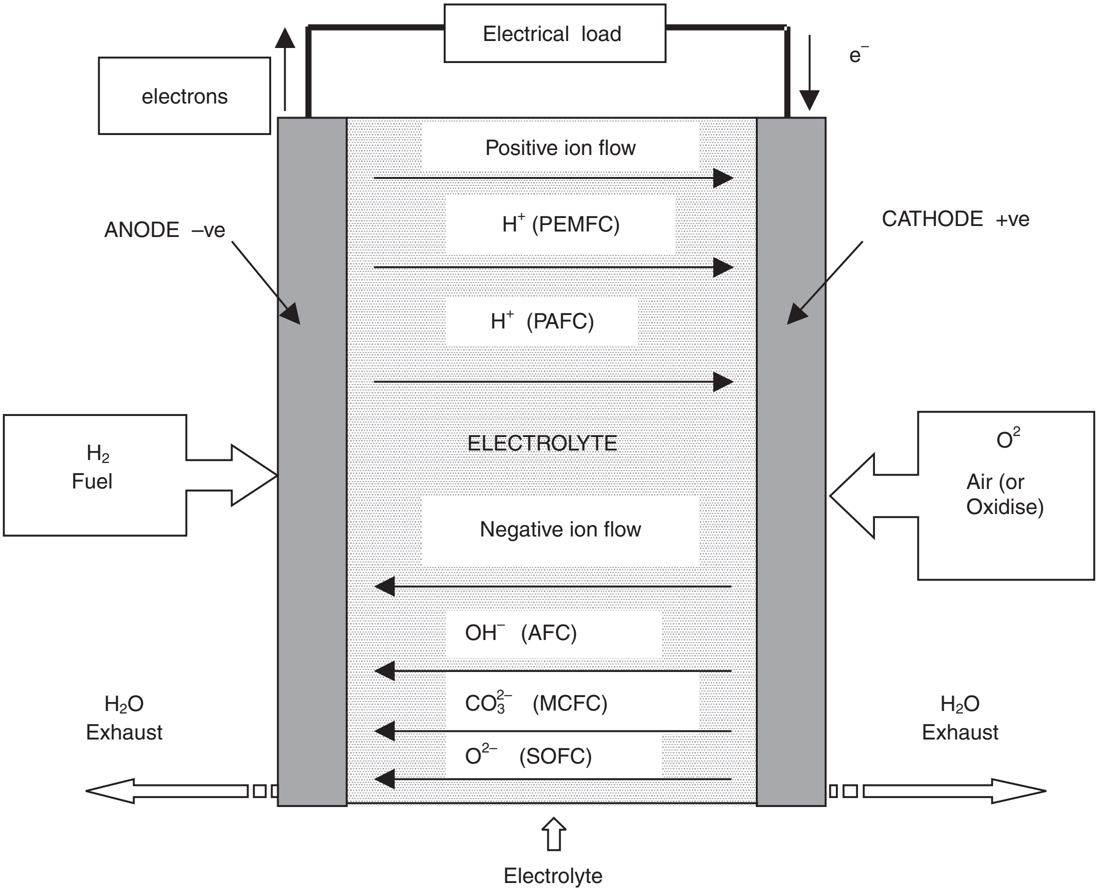


Figure 27.5 Principle of fuel cell operation

the use of methanol would not require such major forecourt engineering for automotive purposes, as would the input of pure of hydrogen. The initial markets for DMFCs are considered to be in the small and medium power applications such as occupied by batteries in consumer and military electronic products. The power to weight ration is theoretically of the order of 10:1 compared with batteries. The main outstanding problem is the 'cross-over' of methanol from the anode to the cathode through the membrane drastically reducing the cell performance. In addition significantly more platinum is required at the anode than in the PEMFC to catalyse the water/methanol mix.

Phosphoric acid fuel cell (PAFC) These fuel cells have given much durable service over many years with a very stable electrolyte in medium to high power stationary applications. Typically 200 kW units are found in hospitals, schools, hotels and various military installations. With its higher temperature carbon monoxide poisoning is not such a severe problem as in the PEMFCs, but with poor ionic conduction this comparatively large and weighty cell technology appears to be relatively expensive.

Molten carbonate fuel cell (MCFC) Multi-megawatt plants have been demonstrated with obvious applications in commercial buildings, especially those requiring high quality heat such as in commercial buildings, hospitals and hotels. Developments include cells for use in natural gas and coal-based power plants in heavy industries. The high operating temperatures allow the use of fuels such as natural gas and coal gas without the need of noble metal catalysts and allow the fuel to be internally reformed without the need of complex reforming equipment. The quality of the heat allows for the possibility of higher efficiencies when coupled to CHP or combined cycle plant.

Solid oxide fuel cell (SOFC) Here the biggest problem in the development of these cells is coping with the high operating temperatures above 850°C where significant material problems occur. Below this temperature ionic conduction is a problem and the cell's performance deteriorates rapidly. Designs and demonstrations are being carried out to supply a wide variety of stationary needs from a few kilowatts to multi-megawatt industrial and power system plants as well as automotive power units. The high temperatures confer the same advantages onto this cell as for the MCFC design with the high temperatures adding further to CHP and combined cycle plants. As a completely solid state device, however, the management problems inherent in liquid electrolyte designs are avoided and, in principle, there are no constraints on cell configuration allowing flexibility in design.

Table 27.12 Fuel cell types and characteristics

Type	*AFC*	*PEMFC*	*DMFC*	*PAFC*	*MCFC*	*SOFC*
Electrolyte	Aqueous potassium hydroxide (30–40%)	Sulphonated organic polymer (hydrated during operation)	Sulphonated organic polymer (hydrated during operation)	Phosphoric acid	Molten Lithium/ Sodium/ Potassium carbonate	Yttria-stabilised Zirconia
Operating Temp.°C	60–90°C	70–100°C	90°C	150–220°C	600–700°C	650–1000°C
Anode	Nickel (Ni) or precious metal	Platinum (Pt)	Platinum–Ruthenium (Pt, Ru)	Platinum (Pt)	Nickel/ Chromium oxide	Nickel/Yttria-stabilised Zirconia
Cathode	Platinum (Pt) or lithiated NiO	Platinum (Pt)	Platinum–Ruthenium (Pt, Ru)	Platinum (Pt)	Nickel oxide (NiO)	Strontium (Sr) doped Lanthanum manganite
Charge Carrier	$OH^{-\Leftarrow}$	$H^{+\Leftarrow}$	$H^{+\Leftarrow}$	$H^{+\Leftarrow}$	$CO_3^{=\Leftarrow}$	$O^{=\Leftarrow}$
Anode reaction	$2H_2 + 4OH^{-\Leftarrow} \rightarrow 4H_2O + 4e^{-\Leftarrow}$	$2H_2 \rightarrow \Leftarrow 4H^+ + 4e^{-\Leftarrow}$	$CH_3OH + \Leftarrow H_2O \rightarrow CO_2 + 6H^+ + 6e^{-\Leftarrow}$	$2H_2 \rightarrow \Leftarrow 4H^+ + 4e^{-\Leftarrow}$	$2H_2 + 2CO_3^{=\Leftarrow} \rightarrow 2H_2O + \Leftarrow 2CO_2 + 4e^{-\Leftarrow}$	$2H_2 + 2O^{=\Leftarrow} \rightarrow 2H_2O + 4e^{-\Leftarrow}$
Cathode reaction	$O_2 + 2H_2O + 4e^- \rightarrow 4OH^{-\Leftarrow}$	$O_2 + 4H^+ + \Leftarrow 4e^- \rightarrow 2H_2O$	$3/2\ O_2 + 6H^{+\Leftarrow} + 6e^- \rightarrow \Leftarrow 3H_2O$	$O_2 + 4H^+ + \Leftarrow 4e^- \rightarrow 2H_2O$	$O_2 + 2CO_2 + 4e^- \rightarrow 2\ CO_3^{=\Leftarrow}$	$O_2 + 4e^- \rightarrow 2O^{=\Leftarrow}$
Heat application	Space +⇐ Water	Space +⇐ Water	Space +⇐ Water	Space +⇐ Water	Combined cycle, CHP	Combined cycle, CHP
Electrical efficiency %	60–70	40–45	30–35	40–45	50–60	50–60
Fuel sources	H_2 removal of CO_2 from both gas streams necessary	H_2 reformate with less than 10 ppm CO	Water/ Methanol Solution	H_2 reformate	H_2, CO, Natural gas	H_2, CO, Natural gas

27.9.3 Fuel cell structure

The theoretical limit on voltage developed by a single cell is about 1.23 V with typical operation being at about 0.7 V. Current (d.c.) delivered is approximately 0.5 amps/cm^2 of cell surface area giving an output power of about 0.35 watts/cm^2. To generate more power the cells are connected together in stacks. *Figure 27.6* shows the structure of a phosphoric acid fuel cell stack together with the flow of gas. These fuel cells can be divided into two groups according to their structure: ribbed separators or ribbed electrodes. A thin layer, or matrix electrode, containing phosphoric acid is sandwiched between the electrodes. The electrodes, or the layers in contact with them, are ribbed to provide a manifold, which ensures that each cell is fed uniformly with air and fuel. The cells can be stacked horizontally or vertically and held together by endplates. The structure is much the same for all types of fuel cell.

27.9.4 Fuel cell plant

A complete fuel cell plant comprises a fuel delivery system, a stack, various controls over the plant operation and output power conditioning equipment. The fuel delivery system can range from a simple flow control unit to a fuel pre-processing unit. In practice hydrogen always occurs in combination with other elements. It is necessary, therefore, to produce it either by electrolysis in the case of water or by separation in a reformer if a hydrocarbon fuel is used as a primary source. A fuel cell system, which includes a fuel reformer, can utilise, in principle, the hydrogen from any hydrocarbon fuel from natural gas to methanol. The two primary types of reformers being developed for transportation are steam reformers and partial oxidation reformers. Steam reformers have higher efficiency but partial oxidation reformers are simpler.

Plant control includes water management control and appropriate temperature control. As exhaust product water is seen as providing a potentially important added benefit in fuel cell operation in future, given concerns about the adequate local supplies of pure water for human consumption and industrial purposes. With regard to temperature control, a preheating stage particularly for high temperature fuel cells may be required for start-up. Heat exchangers are required to ensure the reactants enter the cells at appropriate temperatures for operation, again a particularly important requirement for high temperature cells. The flow rate of the oxidant generally controls the stack temperature.

Finally the power conditioning equipment converts the electricity generated, which is in the form of direct current, into the form required for use, usually alternating current at specified voltage and frequency.

Figure 27.7 illustrates the structure of a fuel cell plant with heat reformer to convert fossil fuel such as natural gas or methanol into hydrogen, or a hydrogen rich gas for supply to the cells and *Table 27.13* provides an outline specification of a 2.8 kW alkaline fuel cell.

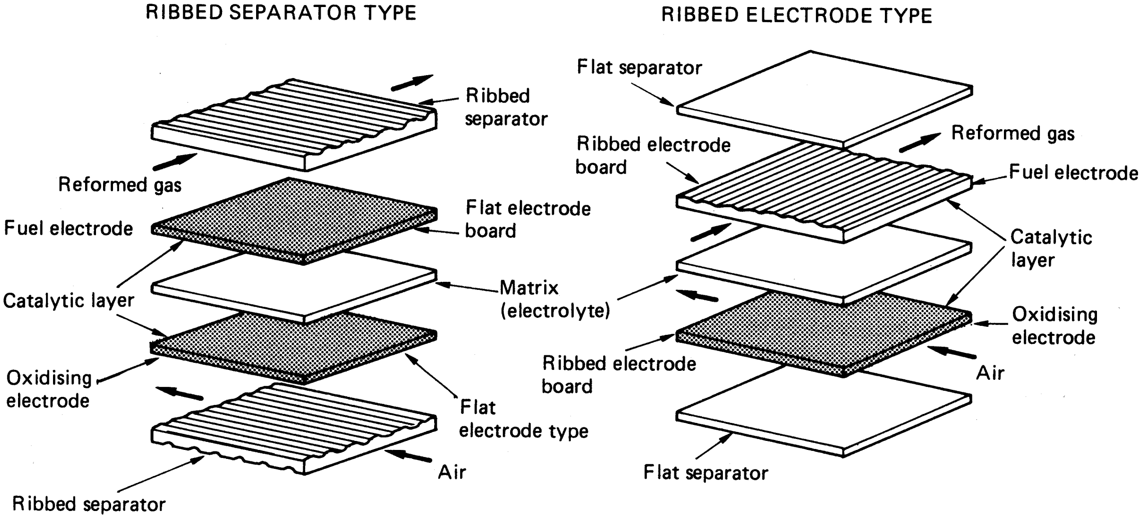

RIBBED SEPARATOR TYPE
Ribbed separator
Reformed gas
Fuel electrode
Flat electrode board
Catalytic layer
Matrix (electrolyte)
Oxidising electrode
Flat electrode type
Air
Ribbed separator
RIBBED ELECTRODE TYPE
Flat separator
Ribbed electrode board
Reformed gas
Fuel electrode
Catalytic layer
Oxidising electrode
Ribbed electrode board
Air
Flat separator

Figure 27.6 Stack cell structure for a phosphoric acid fuel cell

27.9.5 Regenerative fuel cells[38–40]

Fuel cell systems incorporating the collection of the water exhaust, its dissociation by electrolysis using an external source of electricity with storage and subsequent use of the hydrogen and oxygen produced, form a regenerative system. Such systems, for example belonging to the PEM group, could be independent of established fuel/energy infrastructures. The pure forms of the gases would be a further benefit in cells susceptible to carbon monoxide or carbon dioxide. Applications are seen initially in uninterruptible power supplies and in remote power requirements.

A different type of regenerative fuel cell has been developed known as the Regenysis® system which could also be considered as a flow cell type of battery using fuel cell technology. Its chemistry is not based on combining hydrogen and oxygen into water and converting chemical energy into electricity and heat, but instead stores or releases energy by means of a reversible electrochemical reaction between two electrolyte solutions. The electrolytes are the respective salts sodium bromide and a sodium polysulphide and are physically separated by a permeable ion-exchange membrane.

The simplified overall reaction for the cell is given by

$$\underset{\text{uncharged state}}{3\,\mathrm{NaBr} + \mathrm{Na_2S_4}} \Longleftrightarrow \underset{\text{charged state}}{2\mathrm{Na_2S_2} + \mathrm{NaBr_3}}$$

The electrolytes are pumped through two separate electrolyte circuits and transformed electrochemically inside the cell. The charge capacity is limited by the quantities of electrolytes stored in external tanks from which they flow into and out of the cell through separate manifolds via a controlled pumping supply plant. The input of electrical energy from an external source charges the cell with a cation selective membrane preventing the sulphur anions reacting

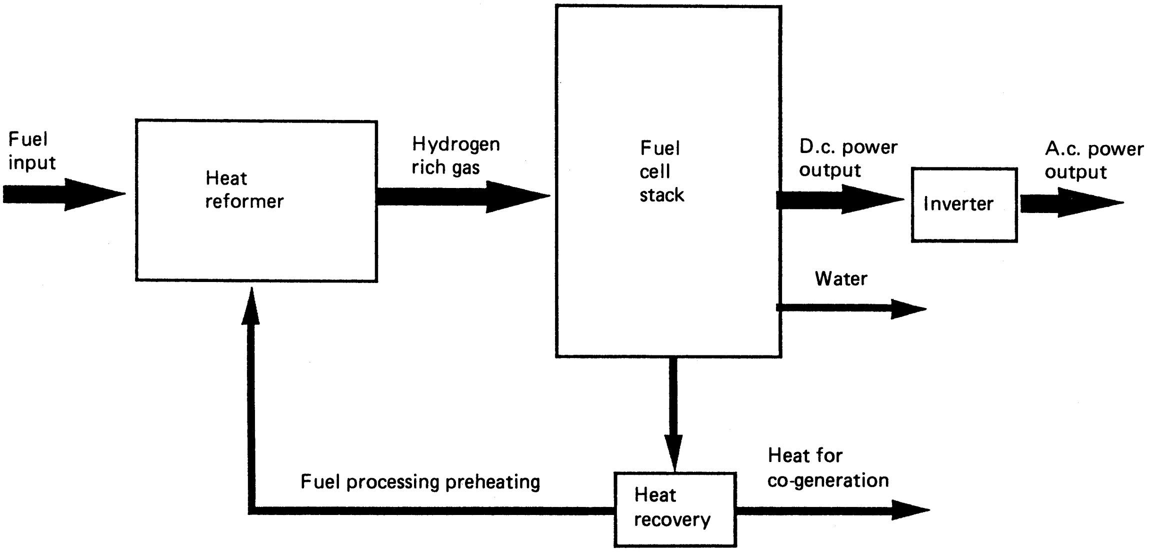

Fuel input
Heat reformer
Hydrogen rich gas
Fuel cell stack
D.c. power output
Inverter
A.c. power output
Water
Fuel processing preheating
Heat recovery
Heat for co-generation

Figure 27.7 Fuel cell power-generation system

Table 27.13 Specification of a low power alkaline fuel cell

Stack configuration		*Nominal operating conditions*	
8 modules connected in series		Reaction temperature	70°C
Nominal dimensions		Reaction pressure	Atmospheric
Length	895 mm	Hydrogen quality	Industrial grade (99.95%)
Width	250 mm	Air quality	Max. 50 ppm CO_2
Height	310 mm	Nitrogen quality	Industrial grade (99.998%)
Weight (excluding electrolyte)	38 kg	Electrolyte	Potassium hydroxide
Weight (including electrolyte)	48 kg	Electrolyte concentration	6.6 mol/l (30% by weight)
Volume	69 litres	Hydrogen supply pressure	Ambient + 40 mbar
		Air supply pressure	Ambient + 40 mbar
Environmental operating conditions		Electrolyte supply pressure	Ambient + 60 mbar
Temperature range	− 10°C to 55°C	Stack lifetime (10% degradation in output power)	5000 h
Humidity range	50% to 90%		
Pressure range	Ambient ±10%	*Fluid flows at full power*	
		Hydrogen consumption	2.15 Nm^3/h
Electrical output at nominal operating conditions		Air consumption	22.5 Nm^3/h
Open circuit voltage	Min. 44 Vdc +10%	Nitrogen consumption	0.16 Nm^3 per on/off cycle
Nominal voltage	32 Vdc ±5%	Electrolyte flow rate	400 l/h at ΔT of 2.5°C
Nominal current	90 A ±5%	Max. water production rate	1.93 l/h
Nominal power	2800 W ±5%		
Maximum allowable current	100 A	*Electrical efficiency*	45%–55% (subject to load, excluding heat recovery)

directly with the bromine separating the solutions. Electrical balance is achieved by the transport of the sodium ions across the membrane.

The overall electrical efficiency of such a system comprising cell, control system and power conversion unit is about 65%, being restricted by membrane performance in transferring the sodium ion Na^+, but with a target efficiency of 80%. Such a fuel cell operates at ambient temperatures and pressures and is, therefore, not suitable for CHP schemes. One design comprising one hundred 100 kW stacks or modules provides 120 MWh of energy storage capable of being released at a rate of 12 MW for 10 h with a peak output of 14.75 MW. In common with all electrochemical systems, maximum efficiency is achieved below the maximum power rating. Power response is fast with zero to full discharge being achieved in 10–15 ms, being limited by the performance of the power conversion unit.

27.10 Heat pumps

27.10.1 Introduction

Heat pumping is the use of a thermodynamic cycle to extract heat from a lower temperature source and supply it to a higher temperature sink where it is useful. In doing so, the purchased energy needed to drive the cycle is less than that usefully supplied. The ratio between useful heat delivered and the energy purchased is the coefficient of performance (COP). It should always be greater than unity.

Thermodynamic cycles have been well developed over the last century for refrigeration, but it is only in recent years that the heating application has developed. They were first applied in the USA where the coastal regions required air conditioned cooling in summer and space heating in winter.[41–47]

The energy crisis in 1974 produced a resurgence in interest in heat pumps, particularly in Japan, France, Germany and Scandinavia. The high capital cost of the units and the stabilisation of fuel prices led to a reduction in sales until the environmental problems of CO_2 were recognised in the 1980s. This provoked a renewed interest and led to newer more reliable, quieter compressors and more efficient cycles. The Japanese are leading this quiet revolution. The applications have widened from simple heat pumping in winter to industrial heat recovery and dehumidification.

27.10.2 Thermodynamics

The ideal thermodynamic cycle for heat pumps, developed by Carnot, assumes a perfect working fluid operating in perfect conditions:

$$\text{Ideal COP} = \frac{T_{\text{hot}}}{T_{\text{hot}} - T_{\text{cold}}}$$

where the heat is extracted from a cold source and supplied to a hot sink. The temperatures are absolute temperature (i.e. degrees Celsius + 273).

This ideal cycle shows that the usefulness of a heat pump cycle depends mainly on the temperature difference between the heat source and the heat sink. The heat pump performance improves as this temperature difference narrows. The heat pump also improves slightly with increasing temperature.

$$\text{COP} = \frac{\text{Heat out}}{\text{Compressor} + \text{Fan energy in}}$$

In practice, with real fluids and real equipment the best performance is obtained from the vapour compression cycle, which achieves about one-third of the ideal Carnot efficiency (see *Figure 27.8*).

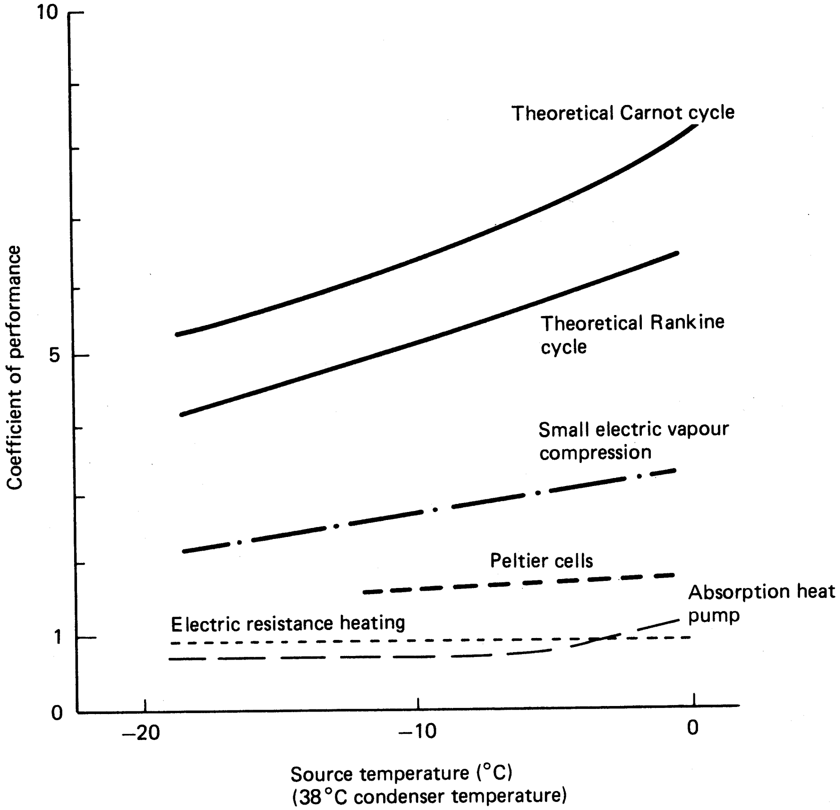


Figure 27.8 Coefficients of performance in theory and practice

27.10.3 Practical cycles

The four basic cycles are described below.

27.10.3.1 Air cycle

When air is compressed, it becomes warmer. Heat can be extracted and the cooled pressurised air expanded down to its original pressure. The expander can be a turbine which drives the compressor. This open cycle can be used for heating buildings. Unfortunately, the equipment is very bulky and its efficiency is very sensitive to the inefficiencies of both the compressor and the expander.[48,49] It is not commercially attractive except for special applications where compressed air is readily available, such as aircraft air conditioning.

27.10.3.2 Vapour compression cycle

The vapour compression cycle relies on the condensation temperature increasing with increase in pressure. A vapour from the evaporator when compressed will condense at a higher temperature, corresponding to the new higher pressure. Successful working fluids must have a high latent heat of condensation so that the bulk of the heat can be extracted at the highest possible temperature. This principle applies to all conventional heat pumps (*Figure 27.9*).

The change in pressure between the evaporator and the condenser can be created by any mechanically driven compressor. Almost all the equipment in use is mechanically driven and electric motors are the driving units. They are favoured because of their cost, simplicity, silence, efficiency, long life and reliability. However, there is an increasing use of fossil-fuel-driven engines because the waste heat from such an engine can often be incorporated into the heating scheme[50] (*Figure 27.10*).

The selection of heat pump working fluids is complex, but a critical factor is the range of permissible condensing temperatures. The upper-temperature limits depend on the compressor. Piston compressors, where the lubricating oil is in

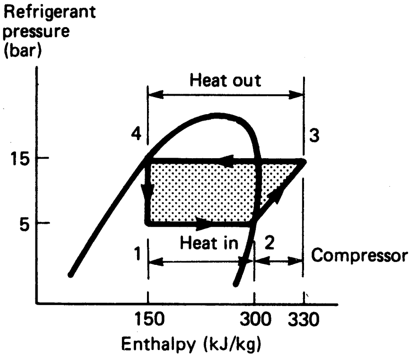


Figure 27.9 The vapour compression cycle (R22). From 1 to 2 the refrigerant vapour absorbs heat. From 2 to 3 the compressor compresses the gas. From 3 to 4 the gas is condensed and its latent heat released. From 4 to 1 the liquid expands to a vapour at the lower pressure

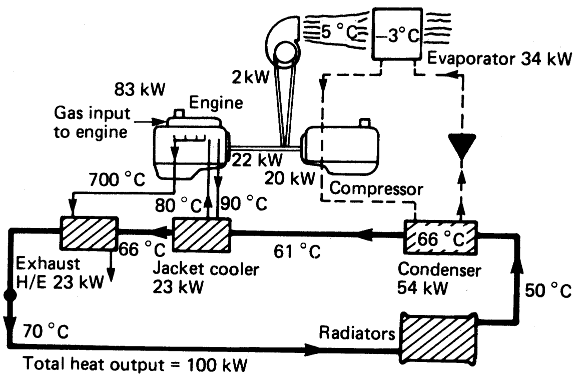


Figure 27.10 A gas engine driven air-to-water heat pump with heat recovery from the engine (coefficient of performance 2.7)

intimate contact with the working fluid, are constrained by oil degradation. Dry compressors, where the bearings are sealed from the working fluid, can operate at higher temperatures. However, the chemical and physical stability of the working fluid itself then provides the working-temperature limits.[51,52]

In 1985 environmentalists discovered that certain refrigerants are attacking the protective stratospheric ozone layer around the earth and destroying it. These refrigerants are the chlorofluorocarbons (CFCs). Government action in Montreal in 1985 produced the first international agreement on restrictions necessary to protect the environment.[53] The two popular fluids R11 and R12 were amongst the most damaging chemicals and are progressively being phased out. Consumption in 1998 was down to 50% of the production in 1986. New ozone friendly refrigerants, in particular R134a, as a substitute for R12, are being developed.[54,55] A range of operating temperatures for different working fluids is illustrated in *Figure 27.11*.

Present-day refrigerants are normally pure single halocarbon compounds. They have simple properties and boil and condense at constant temperatures. They are azeotropic. Non-azeotropic mixtures have advantages in heat transfer and can be used to optimise cycle efficiencies. An illustrative enthalpy diagram is shown in *Figure 27.12*.[56]

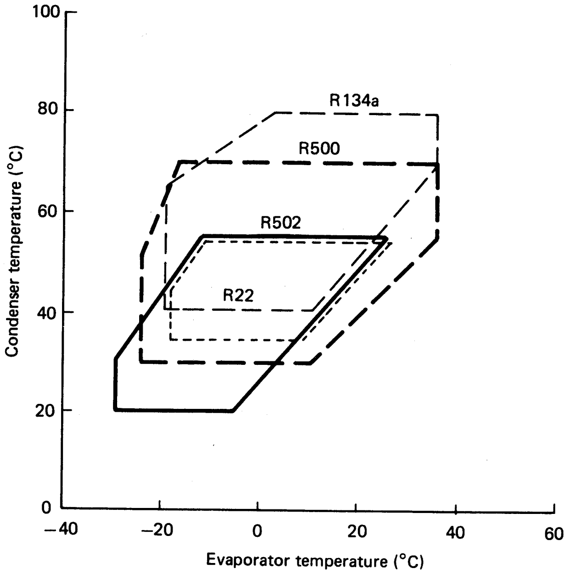


Figure 27.11 The operating-temperature range for different working fluids

This concept is now being used to control externally the refrigerant in the heat pump circuit by storing the refrigerant in a heated rectifier.[57–59] The composition of the refrigerant in the circuit is then controlled by the rectifier temperature (*Figure 27.13*).

27.10.3.3 Absorption heat pump

The absorption cycle works on the same principle as the vapour compression cycle, except that the change in pressure is brought about by adding heat to a solution and releasing the absorbed refrigerant at a higher temperature and pressure.[45,60] The circuit is illustrated in *Figure 27.14*.

The refrigerant leaving the evaporator is physically absorbed in the absorbent, releasing heat. In this mixture the evaporator pressure becomes the partial pressure of the refrigerant in solution. This mixture is then brought up to the pressure of the generator by a liquid solution pump. Heat is applied to the generator and part of the refrigerant boils off and passes to the condenser when heat is released. The refrigerant then passes through an expansion valve to the evaporator. Meanwhile, within the generator, the solution, now depleted of much of the refrigerant, is brought down to the absorber pressure to restart the cycle.

The working fluids are usually ammonia and water or lithium bromide and water. Coefficients of performance are typically 1.2–1.4 on full load.[59] While this is modest, such units are usually gas fired with efficiencies of around 60–70%. The absorption cycle is, therefore, able to double the effectiveness of such energy use.

Part-load performance falls rapidly below 40% heating duty (*Figure 27.15*) and, therefore, the heat pump units are

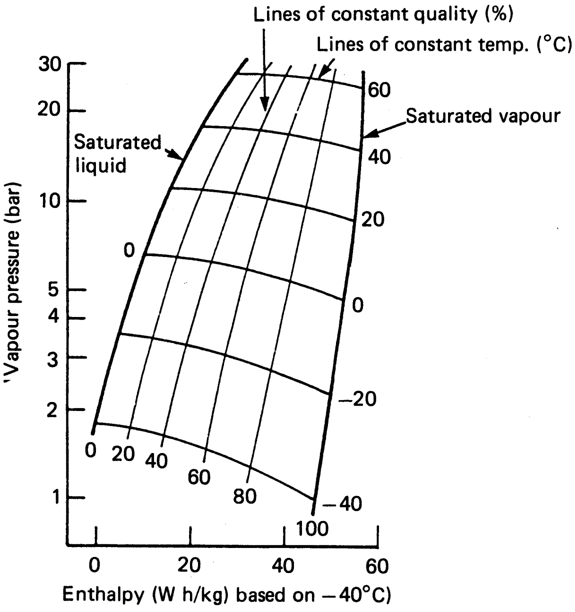


Figure 27.12 Non-azeotropic mixtures do not boil and condense at constant temperature. The enthalpy diagram illustrates the saturated liquid line (bubble point) and the saturated vapour line (dew-point) of a 70%/30% mixture of R13B1/R152A

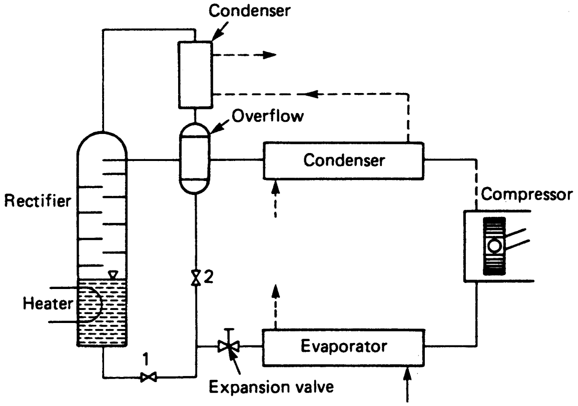


Figure 27.13 The refrigerant composition can be varied at will with a heated rectifier in the non-azeotropic refrigerant circuit

normally underrated for the design condition so that they spend most of their time at full load. The performance characteristics are much less sensitive to source temperature than vapour compression cycles[61] (*Figure 27.16*).

27.10.3.4 Thermoelectric heat pump: Peltier device

When a direct electric current passes round a circuit incorporating two different metals, one junction of the two metals is heated and the other cooled.[62,63] To be effective, these Peltier couples must have a high thermoelectric coefficient α, a low thermal conductivity κ, and low electrical resistivity ρ. The high thermal conductivity of metals normally makes the units very inefficient (coefficient of performance 1.01). However, recent progress in semiconductors has improved α, enabling much more effective units to be made.

The Peltier effectiveness (z) is given be

$$z = \frac{\alpha^2}{\kappa\rho\varsigma}$$

Present-day materials have $z = 0.003/\kappa$.

The overall performance of such devices is still short of that achieved by vapour compression cycles but the small size, reliability and ease of making low-capacity modules gives them a special market. A typical module layout is illustrated in *Figure 27.17*.

27.10.4 Scale

The size and complexity of heat pump application is very wide, with appropriate specialist techniques for each application. For convenience, we examine applications by size.

27.10.4.1 1–10 MW (thermal) schemes

Large-scale heat pump investments are attractive when the running time is long each year. Two types of application meet this requirement. These are base load space and water heating for district heating schemes[64,65] and heat-recovery techniques in large continuous industrial processes.[66,67] The compressors are usually of the high-speed centrifugal type or screw compressors.

Groundwater, sea-water, lakes or sewage treatment can provide the heat source for district heating schemes. Results from the Swedish Sala Municipal district heating network show that a screw compressor can provide 3.2 MW thermal energy at an annual COP of 2.7 (*Figure 27.18*). Availability in its first year was 80%. This heat pump operates throughout the year, providing the base load in winter and the hot water heating in summer.

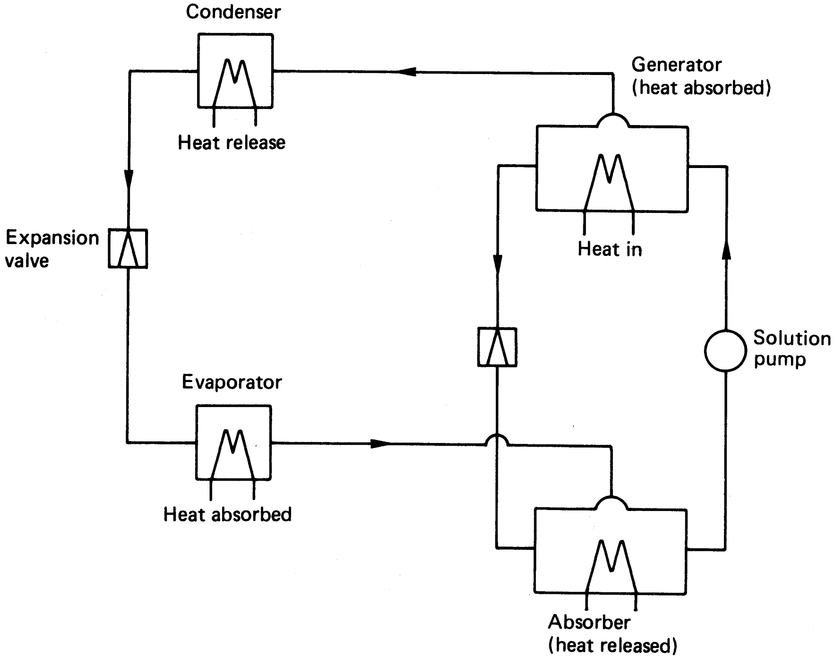


Figure 27.14 Schematic diagram of an absorption heat pump

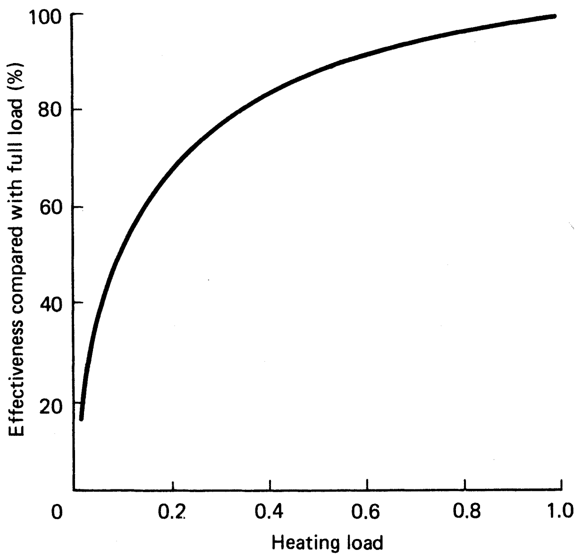


Figure 27.15 The effectiveness of the part-load condition for an absorption heat pump (ammonia/water pair, 12 kW input). (Courtesy of McLinden[61])

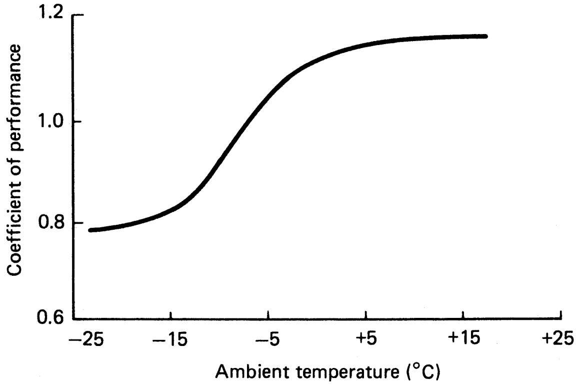


Figure 27.16 The performance characteristics of an absorption heat pump (ammonia/water pair, 12 kW input, 70°C water supply temperature returning at 50°C)

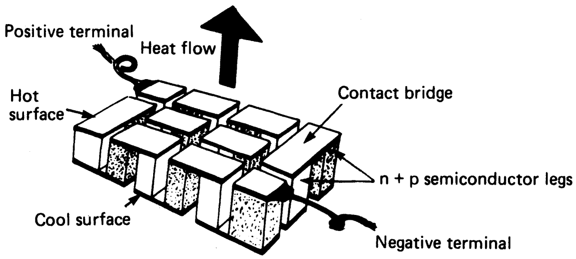

Positive terminal
Heat flow
Hot surface
Contact bridge
n + p semiconductor legs
Cool surface
Negative terminal

Figure 27.17 The Peltier thermoelectric module (d.d. electric)

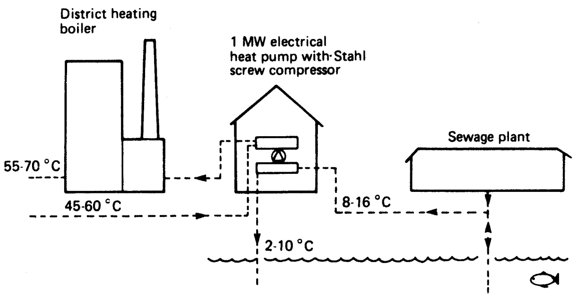

District heating boiler
1 MW electrical heat pump with Stahl screw compressor
Sewage plant
55-70 °C
45-60 °C
8-16 °C
2-10 °C

Figure 27.18 Town sewage provides the heat for Sala, central Sweden, with an annual coefficient of performance of 2.7

Supplementary heating which is needed in the depth of winter is provided by a conventional oil-burning boiler.

Process heat treatment is illustrated by the 1 MW Milk Marketing Board dairy plant at Bamber Bridge, England. This dairy is a bottling and cartoning deport serving retail outlets. Two McQuay Templifier centrifugal compressors are employed in series (*Figure 27.19*). Recycled effluent from the bottle washers is stored and then pumped through both the evaporator and the condenser sections of the first heat pump. The outlet water from the evaporator at 7°C is directed into the dairy supply tank as chilled water for dairy services. The water leaving the condenser passes into the condenser of the second heat pump, where it is heated to 60°C and provides a boiler feed preheat and a crate washing unit. The overall coefficient of performance is 5.5.

27.10.4.2 100 kW to 1 MW (thermal) schemes

The three main applications are commercial buildings, small industrial batch drying plant and swimming pools.[68,69] The compressors are of the multicylinder piston or rotating vane types.

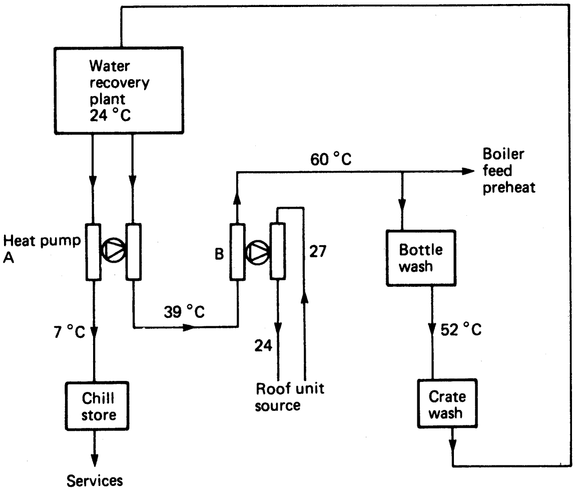

Water recovery plant 24 °C
60 °C
Boiler feed preheat
Heat pump A
B
27
Bottle wash
39 °C
7 °C
24
52 °C
Chill store
Roof unit source
Crate wash
Services

Figure 27.19 Two McQuay centrifugal compressors operate up to 70°C to provide 1 MW of heat recovery in the British Milk Marketing Board's dairy at Bamber Bridge. The coefficient of performance under these conditions is 5.5

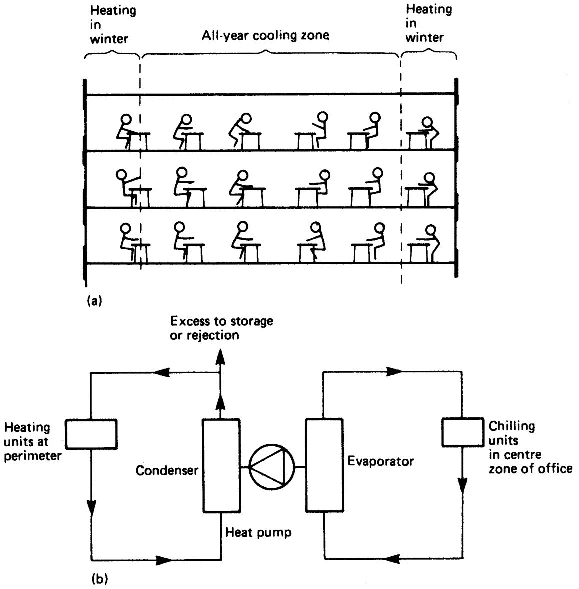

Heating in winter
All-year cooling zone
Heating in winter
(a)
Excess to storage or rejection
Heating units at perimeter
Condenser
Evaporator
Chilling units in centre zone of office
Heat pump
(b)

Figure 27.20 Heat pump recovery in deep-plan offices: (a) section of a deep-plan office; (b) the cooling circuit provides the perimeter heating in winter

Deep open-plan offices require winter heating at the perimeter to combat the outdoor climate, but require permanent cooling of the central core. The heat pump enables the energy to be redistributed within the building (*Figure 27.20*). Such installations halved the energy cost of air-conditioned buildings in Britain, and formed the new concept of integrated environmental design. Coefficients of performances are 3–4.

The second batch of machines is usually a factory packages system for batch industrial drying.[70–72] Compact dehumidifiers which operate up to 80°C are now available for timber drying. Such machines are particularly suitable for the controlled drying required for hardwood to avoid timber splitting (*Figure 27.21*).

Swimming pools are particularly energy intensive. Internal design conditions must not exceed 70% relative humidity of the pool hall air if condensation and mould growth are to be avoided. The conventional technique is to ventilate at a high rate and lower the moisture content of the air by dilution with air from outside. Heat pump dehumidification enables the moisture in the pool air to be controlled without losing great amounts of heat with high ventilation rates. The latent heat recovered from the moisture is used to heat up the pool water (*Figure 27.22*). The warm, moist air conditions mean that the heat pump can operate at a coefficient of performance of 5–6.

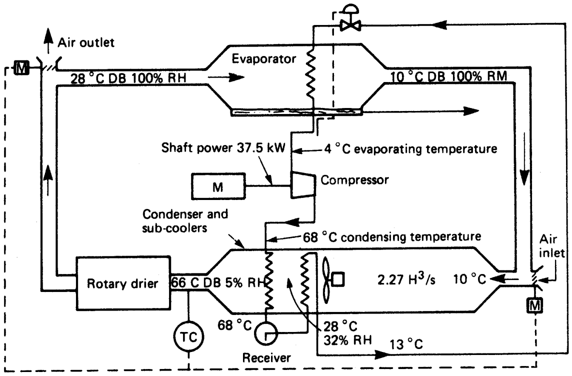

Air outlet
M
Evaporator
28 °C DB 100% RH
10 °C DB 100% RM
Shaft power 37.5 kW
4 °C evaporating temperature
M
Compressor
Condenser and sub-coolers
68 °C condensing temperature
Air inlet
Rotary drier
66 C DB 5% RH
2.27 H³/s
10 °C
M
68 °C
TC
28 °C
32% RH
13 °C
Receiver

Figure 27.21 A single-stage heat pump dryer with subcooler has a coefficient of performance of 4.4[71]

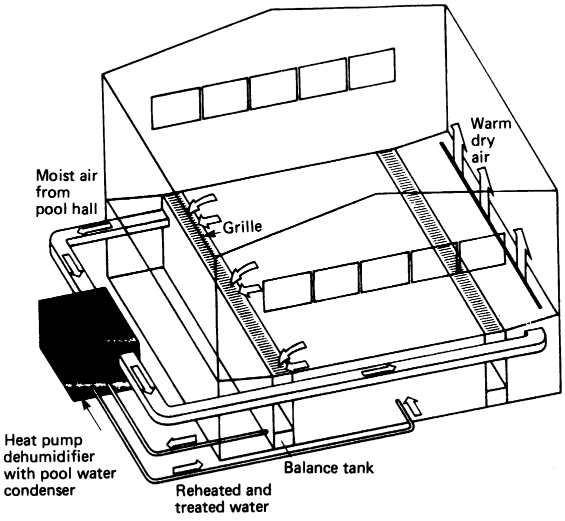

Warm dry air
Moist air from pool hall
Grille
Heat pump dehumidifier with pool water condenser
Balance tank
Reheated and treated water

Figure 27.22 Heat pump dehumidifiers can recover much of the sensible and latent heat from the warm moist air and return it to the pool water. Not only does this recover energy, but it also enables the conventional ventilation rate to be reduced[68]

27.10.4.3 10–100 kW (thermal) packaged units

Two types of machine of this size are factory packages. The most common is the reversible air-to-air space conditioner. It is reversible because, by an arrangement of valves, the unit can interchange evaporator and condenser by cooling in summer or heating in winter. The equipment is usually installed ‘through the wall’ in offices and shops, with each unit controlling a small zone within the buildings (*Figure 27.23*). It is also commonly sited on the flat roof of shops.

A modification of this principle is applied to large buildings, particularly older office blocks where the glazing area is large. In such buildings the individual heat pumps are attached to a ring main of recirculating tepid water. For those parts of the building needing cooling, the local heat pumps reject the heat to the ring main (*Figure 27.24*). Those parts of the building needing heating use the heat from the ring main as their heat source. Any net heating or cooling is provided from the central boiler or the central chiller.

The second and less common type is the air/water space heating heat pump.[73–75] They usually use outside air as the heat source and supply the heat to the house through conventional water radiators (*Figure 27.25*). They are often used in conjunction with supplementary heating because both the effectiveness and the output of such machines fall when the outdoor temperature falls and heating need is greatest (*Figure 27.26*). They also have a maximum water temperature of 55°C, which is lower than the figure for the conventional boiler of 80°C. Care must be taken, therefore,

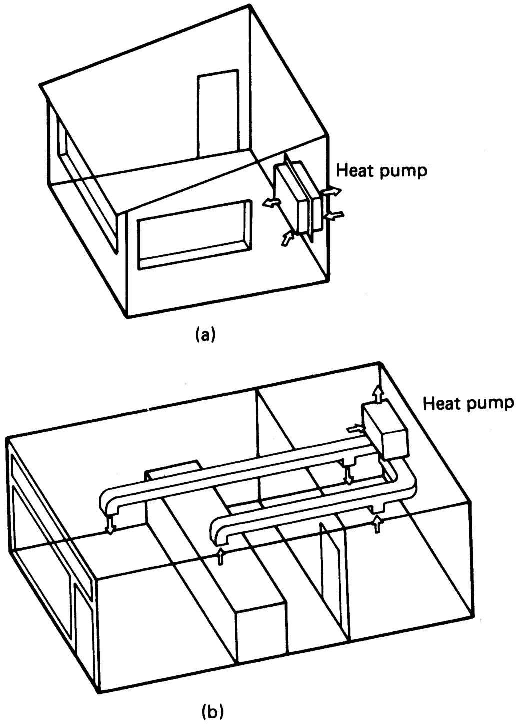

Heat pump
(a)
Heat pump
(b)

Figure 27.23 Reversible air-to-air room units can either heat or cool: (a) reversible air-to-air room unit; (b) self-contained ducted roof mounted heat pump

to ensure that the area of radiator is sufficient when operating at 55°C. There are two other cautionary points—noise and starting current surges. Air source heat pumps can be noisy and must, therefore, be selected and sited so that noise levels immediately outside the bedroom window are below 45 dB(A). Compressors with single-phase electrical drives greater than 1 kW rating must be checked with the local electricity company to see whether the electrical network would be unduly disturbed by the connection of such a pump. Soft-start units are now available on single-phase domestic units. Such devices are particularly helpful if the heat pump is switching on and off frequently.

While ambient air is the usual energy source, groundwater can be used. Even the earth around the building can supply the energy, provided that a sufficient area of brine filled pipes is buried in it (*Figure 27.27*).

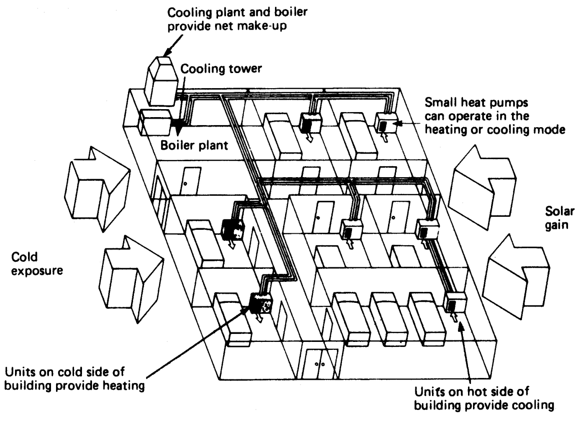

Cooling plant and boiler provide net make-up
Cooling tower
Small heat pumps can operate in the heating or cooling mode
Boiler plant
Solar gain
Cold exposure
Units on cold side of building provide heating
Units on hot side of building provide cooling

Figure 27.24 Linked room units can provide heating in some areas cooling in others

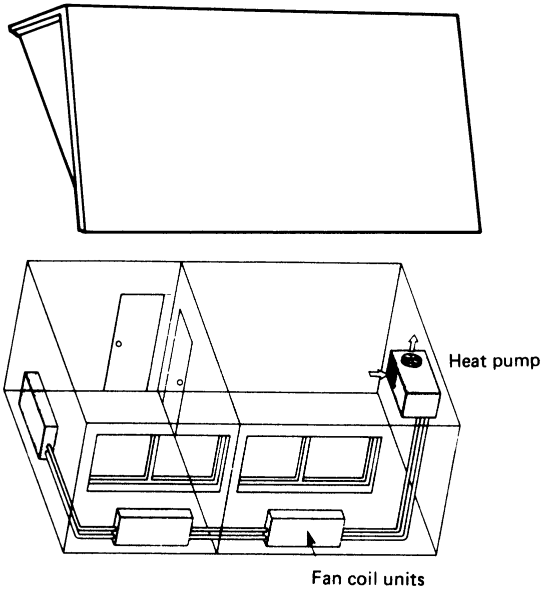

Heat pump
Fan coil units

Figure 27.25 An air-to-water domestic heat pump. The heat pump is shown inside the roof space but it can be sited in the garden

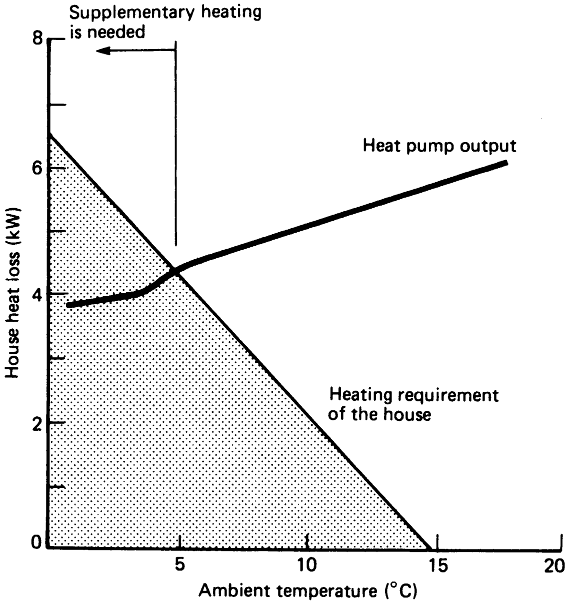


Figure 27.26 The output of an air source heat pump declines with colder conditions. The small step decline around 5°C is due to de-icing the evaporator

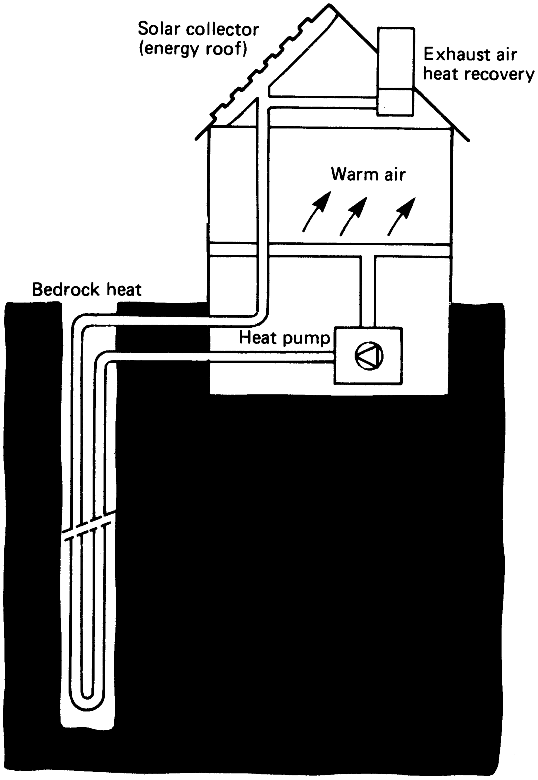

Solar collector
(energy roof)
Exhaust air
heat recovery
Warm air
Bedrock heat
Heat pump

Figure 27.27 Advanced house design uses the heat pump in conjunction with heat recovery. The 'chimney' is now a ventilator, the roof a solar collector and the floor a low-temperature heat emitter. The ground provides the extra energy[65]

27.10.4.4 1–10 kW (thermal) packaged units

Most recent technical advances have occurred in these packaged units. New types of compressor, notably the scroll orbiting (not rotating) compressor, reduce noise levels, are small in size and low in cost and are combined with improved ribbed heat-transfer surfaces and more efficient electric motors. The operating cycle of the scroll compressor is shown in *Figure 27.28* and shown in section in *Figure 27.29*.[76]

New variable-speed drive motors are available to minimise frequent on/off cycles with their consequent wear and intermittent performance. This means that proportional control becomes readily available. The compressor speed range can vary from 1800 rev/min (40% load) to 7200 rev/min (12% design load).[77]

There are three types of application. The first domestic use is a hot water heater. This is usually combined with a hot water cylinder (*Figure 27.30*). The heat source comes from inside the building. Such application is very attractive if interior cooling is needed simultaneously, e.g. cooling a beer cellar and using the reclaimed heat to provide hot water for washing the glasses. Conventional equipment has a maximum temperature of 55°C and, therefore, the volume of stored hot water has to be slightly larger than normal.[78]

The second domestic application is in a mechanical ventilation heat-recovery system. The cycle is illustrated in *Figure 27.31*[79] In a well-insulated house such units can provide 90% of the heating requirements. The heat pump can approach a coefficient of performance of 3.[79]

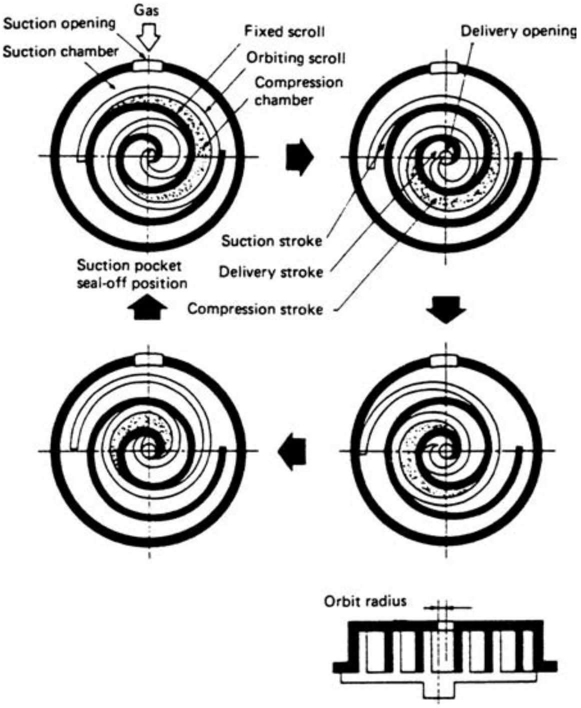

Gas
Suction opening
Suction chamber
Fixed scroll
Orbiting scroll
Compression chamber
Delivery opening
Suction stroke
Suction pocket seal-off position
Delivery stroke
Compression stroke
Orbit radius

Figure 27.28 The operating cycle of the scroll compressor

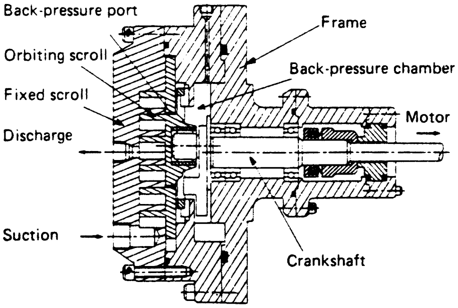

Back-pressure port
Frame
Orbiting scroll
Back-pressure chamber
Fixed scroll
Motor
Discharge
Suction
Crankshaft

Figure 27.29 Section of an orbiting scroll compressor

The industrial application is dehumidification. Portable units are available to dry out damp or newly built buildings. Fixed units are now being used to maintain low humidities in warehouses, particularly as warehouses are becoming automated and no longer require heating for the occupants.

27.10.4.5 100 W to 1 kW (thermal) units

Damp, cold conditions characterise the winter climate in the UK. Small heat pump dehumidifiers extract moisture and translate the latent heat into sensible heat. Cool, damp air enters the evaporator and is chilled, depositing much of its moisture. The same air is then reheated over the condenser and returned to the room. Present equipment has a coefficient of performance which varies from 1.1 to 2.0, the higher value being associated with warmer and damper conditions[80] (*Figure 27.32*).

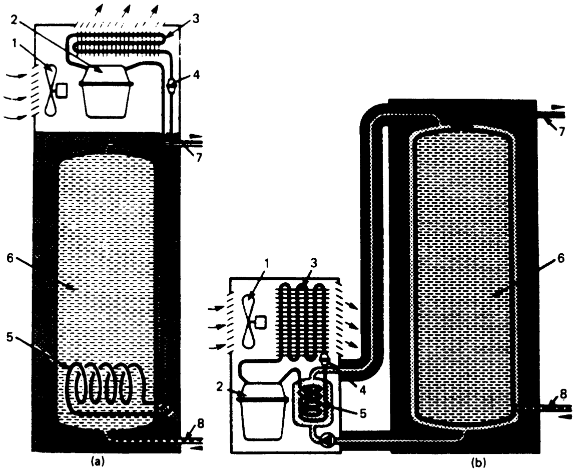

1
2
3
4
5
6
7
8
(a)
(b)

Figure 27.30 A heat pump domestic water heater. Local water regulations may prohibit the direct immersion of a refrigerant heat exchanger into the water cylinder.[78] 1, Fan; 2, compressor; 3, evaporator; 4, expansion valve; 5, condenser; 6, storage; 7, hot water; 8, cold water

27.10.4.6 10–100 W (thermal) modules

Peltier modules are effective ways of providing very small heat flows which can heat or cool. Their main application is to provide stable reference temperatures in scientific equipment.

27.10.5 Conclusions

Heat pump applications are very diverse. The expertise required varies widely with the different types of application. The very large plants (ca. 1 MW) are tailor-made for specific tasks which need much design analysis for successful integration. As the plant size decreases, the heat pump technology is built into packaged units and the application skills needed become those of a building service engineer. Illustrations of the performance and size of different applications are summarised in *Figure 27.33*.

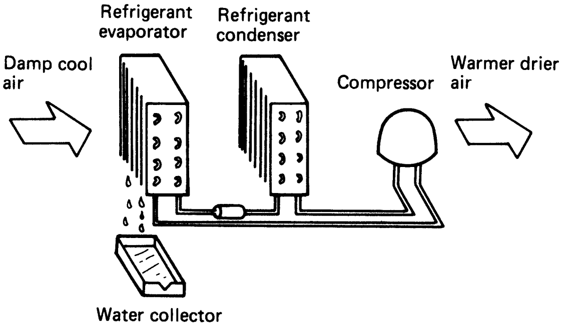

Refrigerant evaporator
Refrigerant condenser
Damp cool air
Compressor
Warmer drier air
Water collector

Figure 27.32 A domestic dehumidifier removes water vapour and provides heat

In general, the heat pump is an advanced piece of engineering which saves energy by extracting it from a low temperature source and making it available at a higher and more useful temperature. The three key factors for its successful use are:

(1) where both heating and cooling are required, preferably simultaneously;
(2) where moisture must be removed, and preferably where some heating is needed simultaneously; and
(3) where the hours of use are long each year, so that the revenue savings can justify the increased initial cost which the heat pump incurs.

It also provides a dramatic reduction in carbon dioxide release and its associated greenhouse effect.

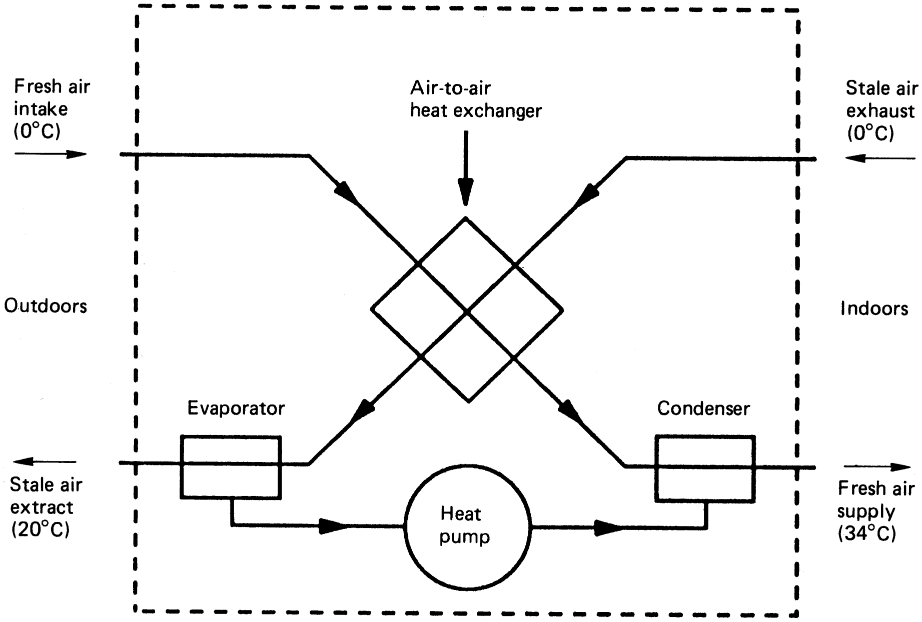

Fresh air intake (0°C)
Air-to-air heat exchanger
Stale air exhaust (0°C)
Outdoors
Indoors
Evaporator
Condenser
Stale air extract (20°C)
Heat pump
Fresh air supply (34°C)

Figure 27.31 Schematic diagram of a mechanical ventilation heat recovery pump

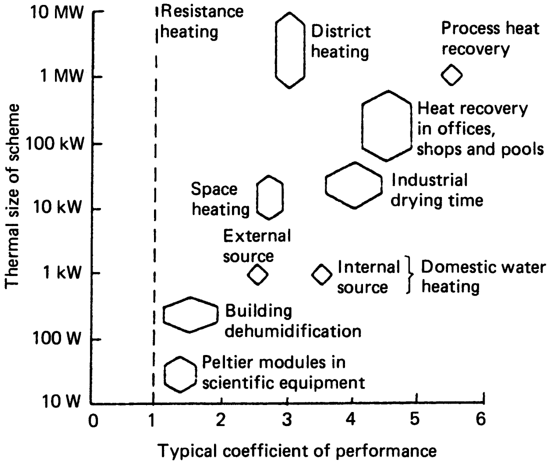


Figure 27.33 Some coefficients of performance for heat pumps in different applications. Coefficient of performance = Heat out/Purchased energy

27.10.6 Professional guides

Air Conditioning and Refrigeration Institute (ARI), USA
ARI 240-77 *Air source unitary heat pump equipment*
ARI 260-75 *Application, installation and servicing of unitary systems*
ARI 320-76 *Water source heat pumps*
ARI 340-80 *Commercial and industrial heat pump equipment*
ARI 270-80 *Sound rating of outdoor unitary equipment*
ANSI/ARI 310-76 *Packaged terminal heat pumps*

Underwriters Laboratories Inc. (UL), USA
ANSI/UL—559-1976 *Heat pumps (a standard for safety)*

National Bureau of Standards (NBS), USA
NBSIR 76-1029 *Unitary heat pump specification for military family housing,* by C. W. Phillips, B. A. Peavy and W. J. Milroy, Prepared for Family Housing Division, US Air Force (1976)
NBSIR 80-2002 *Method of testing, rating and estimating the seasonal performance of heat pumps,* by W. H. Parken, G. E. Kelly and D. A. Didion (1980)

American Society of Heating Refrigerating and Air Conditioning Engineers (ASHRAE), USA
ANSI/ASHRAE 37-38 *Method of testing for rating unitary air conditioning and heat pump equipment*

German Standards (Deutsche Normen)
DIN 8900 *Heat pumps; heat pump units with electric driven compressors*
Part 1 *Concepts* (April 1980)
Part 2 *Rating conditions, extent of testing, marking* (October 1980)
Part 3 *Testing of water/water and brine/water heat pumps* (August 1979)
Part 4 *Testing of air/water heat pumps* (June 1982)

British Air Conditioning Approvals Authority, 30 Millbank, London, SW1P 4RD, UK
Interim standard for performance and rating of air to air and air to water heat pumps up to 15 kW capacity, by B. J. Hough (1982)

Electricity Council, 30 Millbank, London, SW1P 4RD, UK
Heat pumps and air conditioning: a guide to packages systems (1982)

http://www.heatpumpnet.org.uk/

References (Sections 27.1–27.5, 27.7, 27.9, 27.10)

1 'Renewable Sources of Energy', International Energy Agency, OECD, Paris (March 1987)
2 Photovoltaic Government—Industry Group Final Report (187 Kb) http://www.dti.gov.uk/renewable/photovoltaic/index.htm/ May, 2001
3 British Standards Institute, '*Code of Practice for Solar Water Heating Systems for Domestic Hot Water*', BS 5918:1980, London (rev. 1998)
4 British Standards Institute, '*Methods of Test for the Thermal Performance of Solar Water Collectors*', BS 6757:1986, London
5 British Standards Institute, '*Code of Practice for Solar Heating Systems for Swimming Pools*', BS 6785:1986 London
6 ACHARD, P. and GICQUEL, R. (Eds.) '*European Passive Solar Handbook*', Commission of the European Communities, Brussels (1986)
7 '*Gains all round through passive solar design*', Review 2, Dept. of Energy, London (1988)
8 McVEIGH, C. C., '*Alternative energy sources*', Butterworth-Heinemann, Ch. 12, *Mechanical Engineers Reference Book* (1994)
9 LAMPERT, C. M. '*Electrochromic materials and devices for energy efficient windows*', Solar Energy Materials', 11, 1 (1984)
10 FRAENKEL, P., '*Marine currents—a promising large clean energy resource*', IMechE Seminar Publication *Power Generation by Renewables*, pp. 221–233. Bury St Edmunds, UK (2000)
11 '*Tidal Power barrages in the Severn Estuary*', Energy Paper No. 23, Department of Energy, HMSO, London (1977)
12 LAUGHTON, M. A. (Ed.) '*Renewable Energy Sources*', Report No. 22, The Watt Committee on Energy, Elsevier Applied Science, London and New York (1990)
13 '*Wave Energy—the Department of Energy's R & D Programme 1974–83*', Davies, P. G. (Ed.), ETSU Report R26, Harwell, Oxfordshire (March 1985)
14 '*Renewable Energy in the UK: the Way Forward*', Energy paper 55, Department of Energy, HMSO London (1988)
15 '*Wave Energy Review; Interim Report*', ETSU Report R60 for the Department of Energy, Harwell, Oxfordshire (October 1991)
16 http://www.wavegen.co.uk/
17 HARRISON, RHAU, E. and SNEL, H., '*Large Wind Turbines Design and Economics*', John Wiley and Sons (November 2000)
18 'Alternative Energy Sources', Session 1987–88, 16th Report of the Select Committee on the European Communities, House of Lords paper 88, London 21 (June 1988)
19 Wind Energy, Second Report of the Welsh Affairs Committee, Session 1993–94, House of Commons Report, London 13 (July 1994)
20 MAYS, I., 'The current status of experience with wind energy worldwide', Paper S717/017/2000, IMechE Seminar Publication, *Power Generation by Renewables*, pp. 177–183. Bury St Edmunds, UK (2000)

21 MILBORROW, D. J., 'Wind energy technology—the state of the art', IMechE Seminar Publication *Power Generation by Renewables*, pp. 185–193. Bury St Edmunds, UK (2000)

22 US Energy Information Administration Annual Energy Outlook (www.eia.doe.gov) (2001)

23 JENKINS, N., ALLAN, R., CROSSLEY, P., KIRSCHEN, D. and STRBAC G., 'Embedded Generation', IEE Power and Energy Series 31, (2000)

24 ISET, 'Development of wind turbine size', http://euwinet.iset.uni-kassel.de/euwinet/owa/statistics.dispatch (2000)

25 ETSU, 'New and renewable energy: prospects in the UK for the 21st century; supporting analysis', ETSU R-122 (1999)

26 Department of Trade and Industry (1999), 'New and Renewable Energy—prospects for the 21st Century', DTI Publication URN 99/744. London

27 ARBON, I. M. and BOWELL, B. P., 'Experiences with power generation from biomass', IMechE Seminar Publication *Power Generation by Renewables*, pp. 263–290. Bury St Edmunds, UK (2000)

28 'Modelling of Carbon and Energy Budgets of Wood Fuel Coppice Systems', MATTHEWS, R., ETSU, B/W5/REP (1994)

29 PITCHER, K., PATERSON, B. and WEEKES, A., '*ARBRE—a case study of the development of a Biomass Integrated Gasification-combined Cycle Power plant*', IMechE Seminar Publication, *Power Generation by Renewables*, pp. 41–52. Bury St Edmunds, UK (2000)

30 www.ott.doe.gov/biofuels/biofuels.html

31 www.eren.doe.gov/RE/bio_resources.html

32 www.press.detr.gov.uk/0103/0127.htm

33 www.ieabioenergy.com

34 PEACOCKE, G. V. C. and BRIDGWATER, T., '*Pyrolysis and gasification of biomass: status of the UK industry*', IMechE Seminar Publication, *Power Generation by Renewables*, pp. 77–92, Bury St Edmunds, UK (2000)

35 PORTEOUS, A., 'Municipal solid waste recovery—a comparison between mass burn incineration and gasification options', IMechE Seminar Publication *Power Generation by Renewables*, pp. 113–130, Bury St Edmunds, UK (2000)

36 LARMINIE, J. and DICKS, A., '*Fuel Cell Systems Explained*', Wiley (April 2000)

37 HIRCHENHOFER, J. H., STAUFFER, D. B., ENGLEMAN, R. R. and KLETT, M. G., '*Fuel Cell Handbook*', 4th Edition, US Department of Energy, Morgantown, WV, USA (1998)

38 PRICE, A. and McCARTHY, L., '*Power generation using renewables and the regenesys energy storage system*', IMechE Seminar Publication, *Power Generation by Renewables*, pp. 195–206, Bury St Edmunds, UK (2000)

39 BURKE, K. A., '*High energy density regenerative fuel cell systems for terrestrial application*', IEEE AES Systems magazine, pp. 23–24, (December 1999)

40 MITLITSKY, F., MYERS, B. and WEISBERG, A. H., '*Regenerative fuel cell systems*', Energy and Fuels, Vol. 12, pp. 56–71 (1998)

41 AMBROSE, E. R., *Heat Pumps and Electric Heating*, Wiley, New York (1966)

42 HEAP, R. D., *Heat Pumps*, E & F Spon, London (1979)

43 BERNIER, J., *La Pompe de Chaleur*, Pyc Edition, Paris (1979)

44 REAY, D. A. and MACMICHAEL, D. B. A., *Heat Pumps—Design and Application*, Pergamon Press, Oxford (1979)

45 VON CUBE, H. L. and STEIMLE, F., *Heat Pump Technology*, Butterworths, London (1981)

46 GROFF, G. C. (Ed.), *Heat Pump and Space Conditioning Systems for the 1990s*, International Symposium by Carrier Corporation, USA (1979)

47 LOPEZ-CACICEDO, C. L., Electrically driven heat pumps: current research and future prospects, *Energy World Heat Pump Supplement* (19 October 1981)

48 COOPER, K. W. and SUMNER, L. E., An open cycle study using moist air thermodynamics, *ASHRAE J.*, 68–71 (January 1978)

49 VAUTH, R., A new heat pump operating on the cold air principle, *Heat Pump Conference, Essen* (1977). (Electricity Council OA Trans 1775)

50 MASTERS, J. and PEARSON, J., Automotive gas engines power air conditioning systems, heat pump installations and heat and power units, *Int. Gas Eng.* (January 1981)

51 JENSEN, W., Criteria for the use of compression heat pumps in industry, *ASVE*, **6**, 22–28 (1981)

52 PAUL, J., Schrauben und Kolbenverdichter in Vergleich, *Kalte Klimatechnik,* **12**, 1–6 (1981)

53 Montreal protocol on substances that deplete the ozone layer, Montreal, 16 September 1987. Presented to Parliament January 1988 *CM283 Miscellaneous 1*, HMSO, London (1988)

54 SAND, J. R., VINEYARD, E. A. and NOWAK, R. J., Experimental performance of ozone safe alternative refrigerants, *ASHRAE Trans.*, *Part 2*, **96**, 173–182 (1990)

55 KAMEL, A., PIAO, C. C., SATO, H. and WATANABE, K., Thermodynamic charts and cycle performance of FC 134a and FC 152a, *ASHRAE Trans., Part 1*, **96**, 141–148 (1990)

56 CONNON, H. A. and DREW, D. W., Estimation and application of thermodynamic properties for a non azeotropic refrigerant mixture, presented at Int. Inst. Refrig. Conference on Heat Pumps and Air Circulation in Conditioned Spaces, Essen, Germany, *Proc. of the Meetings of Commissions B1, B2, E1, E2*, 91–100 (September 1981)

57 KRUSE, H., The advantages of non azeotropic refrigerant mixtures for heat pump applications, *Int. J. Refrig.*, **4**(3), 119–125 (1981)

58 KRUSE, H., Current status and future potential of non azeotropic mixed refrigerants, *Proc. 1987 IEA Heat Pump Conference, Orlando*, pp. 173–94, Lewis Publishers, New York (1988)

59 MUHLMANN, H. P. and WESSING, W., Test rig data, process data and operating data for a prototype air to water absorption heat pump, presented at Absorption Heat Pumps Congress, Paris, March 1985, *Report EUR 1007 EN*, 452–467, Commission of European Communities, Luxembourg (1985)

60 JESINGHAUS, J., Development trends in absorption heat pumps, *Sonnenergie Warmepumpe*, **6**(5), 31–34 (September/October 1981). *Translation WH 738*, British Gas, Watson House

61 McLINDEN, M., Part load performance of absorption heat pumps, presented at Absorption Heat Pumps Congress, Paris, March 1985, *Report EUR 10007 EN*, 503–516, Commission of European Communities, Luxembourg (1985)

62 IOFFE, A. F., *Semiconductor Thermoelements and Thermoelectric Cooling*, Infosearch Ltd, London (1957)

63 SPANKE, D., Air conditioning using heat pumps and Peltier cells, *Elektrowarme Int.*, **26**(6), 220–227 (June 1968). *OA Translation 1152*, Electricity Council

64 YANKOV, V. S. and FILKOV, V. M., Soviet Research on Large Heat Pump Stations for Centralised Heat Supply, *World Energy Conference Working Party Report*, London (1979)

65 SWEDISH STATE POWER BOARD, *Heat Pumps and Solar Energy*, Stockholm (1982)

66 NEI PROJECTS LTD, *The Templifier Heat Pump* (*Publication NP2*), Newcastle (1983)

67 LAWTON, J., MACLAREN, J. E. T. and FRESHWATER, D. C., Heat pumps in industrial processes, in *The Rational Use of Energy*, 47–56, Watt Committee, London (1977)

68 BRAHAM, D., The energy factor, *51st Annual Conference on Baths & Recreational Management*, 1–27 (September 1981)

69 FESSEL, E., Public indoor swimming pool with heat recovery, *Elektrowarme Int*., **33**(5), 230–234 (1975)

70 HODGETT, D. L., Dehumidifying evaporators for high temperature heat pumps, *UF-Int. Inst. Refrigeration Conference*, Belgrade (1977)

71 PERRY, E. J., *Drying by Cascaded Heat Pumps*, Institution of Refrigeration, London (1981)

72 GEERAERT, B., Air drying by heat pumps with special reference to timber drying, in Camatini, E. and Kester, T. (Eds), *Heat Pumps and their Contribution to Energy Conservation*, Noordhoff, Brussels (1982)

73 SETA, P., BERHONDO, P. and ROBIN, P., Heat pump tests in new individual homes and block apartments: first results and observations, *Comité Français d'Electrothermie*, *Versailles Symposium* (1979)

74 JACKSON, A. and STERLINI, P. A., The performance of air to water heat pumps in domestic premises, *Domestic Heating*, **14**, 12–14 (1981)

75 KALISCHER, P., Operating experience with dual fuel heat pump systems, *Elektrowarme Inst*., **37**, 266–271 (1979)

76 IKEGAWA, M., SATO, E., TOJO, K., ARAI, A. and ARAI, N., Scroll compressor with self adjusting back pressure measurement, *ASHRAE Trans*., **90**(2A), 314–326 (1984)

77 TOSHIBA, *Toshiba Technical Leaflet: Inverter Controller Air Conditioning* (1986)

78 KALISCHER, P., The heat pump for hot water supply in the residential sector, *Unipede Workshop on Domestic Electric Hot Water Supply*, EBES, Antwerp, (April 1982)

79 BERTINAT, M. P., Heat pump ventilation units: their potential in UK houses, *ERDC Memo 2586*, Capenhurst (June 1991)

80 BRUNDRETT, G. W. and BLUNDELL, C. F., An advanced dehumidifier for Britain, *Heating Vent. Eng*., 6–9 (November 1980)

28

Alternating Current Generators

F Parker MBA, PgD, BSc (Hons), CEng, MIEE
Newage International Ltd

Contents

28.1 Introduction 28/3

28.2 Airgap flux and open-circuit e.m.f. 28/4
  28.2.1 Airgap flux waveforms 28/4
  28.2.2 Open-circuit e.m.f.: integral slots per pole 28/4
  28.2.3 Open-circuit e.m.f.: fractional slot windings 28/7
  28.2.4 Slot ripple e.m.f.s 28/7

28.3 Alternating current windings 28/9
  28.3.1 Choice of slot number 28/10
  28.3.2 Integral-slot windings 28/10
  28.3.3 Fractional-slot windings 28/10
  28.3.4 Parallel circuits 28/11

28.4 Coils and insulation 28/11
  28.4.1 Service conditions 28/11
  28.4.2 Stator coils 28/12
  28.4.3 High voltage insulation systems 28/14
  28.4.4 Insulation testing 28/15
  28.4.5 Rotor coils 28/16

28.5 Temperature rise 28/16

28.6 Output equation 28/17
  28.6.1 Some design parameters 28/18

28.7 Armature reaction 28/20
  28.7.1 Cylindrical-rotor machine 28/20
  28.7.2 Salient-pole rotor machine 28/20
  28.7.3 Magnitudes and equivalence of stator and rotor m.m.f. 28/21

28.8 Reactances and time constants 28/22
  28.8.1 Armature leakage reactance 28/22
  28.8.2 Magnetisation (armature reaction) reactances 28/22
  28.8.3 Synchronous reactances 28/22
  28.8.4 Transient and subtransient reactances 28/22
  28.8.5 Negative-sequence reactance 28/24
  28.8.6 Zero-sequence reactance 28/24
  28.8.7 Reactance values 28/24
  28.8.8 Reactances and time constants 28/24
  28.8.9 Potier reactance 28/25
  28.8.10 Frequency-response tests 28/25

28.9 Steady-state operation 28/25
  28.9.1 Open- and short-circuit characteristics 28/25
  28.9.2 Phasor diagram and power output 28/26

28.10 Synchronising 28/27
  28.10.1 Synchronising procedure 28/27
  28.10.2 Synchronising power and torque 28/28
  28.10.3 Rotor oscillation 28/28

28.11 Operating charts 28/29
  28.11.1 Cylindrical-rotor generator 28/29
  28.11.2 Salient-pole generator 28/30

28.12 On-load excitation 28/31
  28.12.1 M.m.f. phasor diagram 28/31
  28.12.2 The ANSI Potier reactance method 28/31
  28.12.3 Use of design calculation 28/32

28.13 Sudden three-phase short circuit 28/32

28.14 Excitation systems 28/34
  28.14.1 D.c. exciters 28/36
  28.14.2 A.c exciters with static rectifiers 28/36
  28.14.3 Brushless excitation 28/36
  28.14.4 Thyristor excitation 28/37
  28.14.5 Excitation systems circuits 28/37
  28.14.6 Excitation control 28/37
  28.14.7 Basic principles of voltage control 28/38
  28.14.8 Additional control features 28/40
  28.14.9 Overall voltage response 28/40
  28.14.10 Digital control 28/41

28.15 Turbogenerators 28/41
  28.15.1 Main dimensions 28/41
  28.15.2 Rotor body 28/42
  28.15.3 Rotor winding 28/42
  28.15.4 Stator core 28/43
  28.15.5 Stator casing 28/43
  28.15.6 Stator winding 28/43
  28.15.7 Cooling 28/43

28.16 Generator–transformer connection 28/46

28.17 Hydrogenerators 28/46
  28.17.1 Introduction 28/46
  28.17.2 Construction 28/47
  28.17.3 Cooling 28/49
  28.17.4 Excitation 28/49
  28.17.5 Pumped storage units 28/49

28.18 Salient-pole generators other than hydrogenerators 28/50
  28.18.1 Applications 28/50
  28.18.2 Construction 28/51
  28.18.3 Ventilation and cooling 28/51
  28.18.4 Particular design requirements 28/52

28.19 Synchronous compensators 28/52

28.20 Induction generators 28/53

28.21 Standards 28/53

28.1 Introduction

For the generation, transmission, distribution and use of electrical power the three-phase system has large economic and practical advantages over single-phase or two-phase systems. Hence the great majority of alternating-current (a.c.) generators are three-phase machines, operating at one of the standard frequencies, 50 or 60 Hz. Some generators operate at other frequencies; examples are: (a) generators up to several megawatts in output operating at $16\frac{2}{3}$ Hz have been used for years to supply rail traction systems, mostly in Europe, (b) high frequency generators, at 500–10 000 Hz, were used extensively for induction heating in industrial processes, (c) small shaft mounted generators with more than three phases are used as exciters for synchronous generators and some hydrogenerators, (d) aircraft ground power supplies operating at 400 Hz, and (e) a growing trend for machines to operate at other variable frequencies with power electronics being used to convert their output to 50 or 60 Hz. However, due to the cost of the electronic conditioning equipment, these are largely limited to less than 150 kW in output and under 1000 V.

The form of construction of the generator depends on its output power and its speed: these are determined by the prime mover that drives it. To generate at a frequency of f hertz when driven at n rev/min the generator must have $2p$ poles, where

$$2p = \frac{120f}{n}$$

n is the synchronous speed, being the same as the speed of rotation of the magnetic field produced by currents in a three-phase winding when connected to a supply of frequency f Hz. The maximum permissible diameter of the rotor will be determined by the rotational stresses acting on it and thus the speed of rotation determines the shape of the generator. For a given output power, a high-speed machine will have a smaller diameter, and longer length than a low-speed machine.

Generators may be classified as follows.

Synchronous generator This type of generator requires a winding carrying direct current (or in small sizes a series of permanent magnets) to establish the magnetic flux. In nearly all machines this excitation winding (known as the field winding) is carried on the rotor, which for a 50 or 60 Hz output must rotate at the synchronous speed.

There are two sorts of synchronous generator which are differentiated by the type of rotor used; rotors are either cylindrical or of the salient pole type.

The first type of synchronous generators are known as *turbogenerators*. This family of machines use a cylindrical rotor in which the field winding is housed in axial slots. They are invariably driven by a steam turbine or a gas turbine. At ratings below 60 MW a gear box may be used to provide a rotational speed of 3600 (2 pole) or 1800 rev/min (4 pole) to provide power at 60 Hz, or 3000 rev/min (2 pole) or 1500 rev/min (4 pole) to provide power at 50 Hz. Alternatively, and especially at higher power ratings the generator is directly driven by the steam or gas turbine. The rotors will thus have either two or four poles. Smaller machines may use a laminated construction for the rotor while larger machines will use a forged rotor. A feature of these machines is that their length is several times their diameter. Power outputs range from a few megawatts up to about 1500 MW. The machine is cooled by circulating air or hydrogen over the active parts or water through the windings. Hydrogen was commonly used for outputs greater than about 50 MW; water was, and still is, used for the stator winding with outputs exceeding about 200 MW. Air-cooled machines are now available up to almost 200 MW.

The second type of synchronous generators use a salient pole rotor. Types of salient pole machines are:

Hydrogenerators This is a family of salient pole machines driven by water turbines at a speed in the range 50–1000 rev/min. The speed depends on the type of turbine, which in turn depends on the head and the flow rate of the water available. At low speeds, the permissible rotor diameter will be several times its active length. Generally, the largest allowable diameter of rotor is used to maximise the machine's inertia which is an important part of governing the water turbine. Outputs up to 800 MW have been achieved. A small high-speed unit will have a horizontal shaft, but for reasons of mechanical construction and stability larger machines have vertical shafts.

Reciprocating gas, diesel or petrol engine-driven generators For this application, the generator may be coupled directly to the internal combustion engine and the generator will invariably be of the salient pole construction. Many combinations of output power and speed are available, from a few kilovoltamperes, usually at four-pole speed, up to 45 MW or more at 100 rev/min, using a 2 stroke diesel engine of the type used for ship propulsion.

Steam turbine or gas turbine driven generators Salient pole generators connected to these prime movers are driven through a gearbox, usually at 4-pole speed. Ratings are limited to below 60 MW both by the gearbox capacity and by the difficulty of holding large salient pole rotors together.

Synchronous compensators (also known as synchronous capacitors) These machines draw current from the system at zero power factor, lagging or leading as required to control the voltage of the system. They also draw a small amount of real power to maintain the synchronous rotation. They will usually have six or eight salient poles, and a rating in the range from a few megavolt-amperes up to say 350 MVAR. For machines larger than say 50 MVAR, hydrogen cooling is used to reduce the windage loss. The synchronous compensator has been rendered obsolete by the development of static volt-ampere (VAR) control equipment.

Asynchronous (induction) generators In construction these machines resemble induction motors, and similarly draw their magnetising current from the power system, to which they deliver power when driven at very slightly above synchronous speed. Ratings are usually less than 3 MW, at speeds up to 1000 rev/min. Much smaller units, operating with capacitor excitation circuits, can provide isolated power supplies.

For generators of all types the practicable and economic voltage increases with the increase of rated output in order to limit the amount of current to be handled to a manageable level. Some standards specify preferred voltages, and some are normally mandatory. Typical ranges of output and voltage are:

MVA	1–6	4–16	4–100	100–400	400–800	800–1500
kV	1–4	5–7	10–13	14–16	18–23	26–30

However, a number of manufacturers have found significant markets by going away from these norms. Low voltage, high current machines together with an appropriate transformer have found favour where regulations restrict the installation of higher voltage generators. Conversely, high voltage generators, even at very modest outputs have proved attractive as the cost of a step-up transformer can be avoided.

Operating generators singly or in a group When a generating set runs alone to supply power to a load, the prime mover supplies the active power demanded. As the demand changes, small changes of speed cause the governor to adjust the mechanical input power to maintain as nearly as possible the correct speed and frequency. The generator supplies the active power and the reactive power required by the load, and this determines the power factor. Changing the generator's field current changes the terminal voltage, and the consequent changes in active and reactive power depend on the nature of the load, e.g. the proportion of motor load to static load. Usually, as the load changes, the voltage is held almost constant by adjustments to the excitation made by the automatic voltage regulator.

Most generators, however, are synchronous machines connected to an extensive supply network, and changing the excitation of one machine does not affect the system voltage, the speed or the power output of the generator. It changes only the reactive power, and a compensating change in reactive power is shared by other generators on the same or neighbouring bus-bars. Thus adjusting the generator field current when it is connected to a system results only in a change in the power factor at which the generator operates.

System engineers find it convenient to regard active power in watts (P) and reactive power in vars (Q) (the two components of the apparent power in volt-amperes (VA)) as separate but related entities to be generated, transmitted and absorbed.

Induction motors, chokes, transformers and underexcited synchronous machines all draw magnetising current, lagging the supply voltage by 90°.⇐ By convention, this is regarded as a demand for positive vars from the system. Capacitors and overexcited synchronous machines draw leading current, and so supply positive vars to the system. A synchronous machine is overexcited if its field current is more than is needed to make it operate at unity power factor.

28.2 Airgap flux and open-circuit e.m.f.[1,6,7,8,11,17,19,23]

28.2.1 Airgap flux waveforms[7,16,23,44–47]

Figure 28.1 shows the flux pattern of a turbogenerator (t.g.) on open circuit, and flux waveforms for it and for an 18-pole salient-pole (s.p.) generator. The t.g. flux waveform is inherently trapezoidal because of the distributed field winding and uniform length of airgap. Prominent ripples caused by the stator slots and teeth have been omitted (see Section 28.2.4).

The salient pole generator flux waveform is inherently rectangular, with short sections of low density between the poles. In this example, however, the shape is closer to sinusoidal, as the fundamental component shows, because the machine was designed with shaped pole shoes. These ensure that the length of the airgap at the extremities of the pole shoe is greater than on the pole centre line. The slot ripples which arise as a result of the stator slot opening have been retained: they are not quite symmetrical because the stator did not have a whole number of slots per pole (actually $10\frac{1}{2}$). Where integral slots per pole are used, great care is needed in the design and spacing of the pole face damper damper bars in order to avoid further additional ripples appearing in the voltage waveform.

These waveforms were produced using finite element analysis: the harmonic contents of the flux waves are shown in *Table 28.1*. The reference value of 100% is the peak of the fundamental component of the open-circuit wave of each machine.

Even harmonics do not normally occur on open circuit, because the waveshape is the same under all the poles, and is symmetrical about each pole centre line. All the harmonic flux density curves go through zero at the interpolar axes. The total flux per pole is, therefore, the fundamental component plus or minus the flux of one pole-pitch of each harmonic: 'plus' if the harmonic has a negative peak at the positive peak of the fundamental, and vice versa.

Wieseman[44] gave curves for a salient pole machine relating the peak flux densities of the fundamental and of the third harmonic to the peak of the actual flux wave. The multipliers

$$\frac{\text{Peak of fundamental}}{\text{Peak of actual flux wave}} = \text{€}1, \quad \text{say}$$

and

$$\frac{\text{Peak of third harmonic}}{\text{Peak of fundamental}} = \text{€}3, \quad \text{say}$$

were deduced as functions of the pole arc to pole pitch ratio, the minimum airgap to pole pitch ratio and maximum edge airgap to minimum airgap.

Ginsberg *et al.*[45] gave curves from which the flux wave fundamental and harmonics up to the 11th can be estimated, again in terms of the airgap lengths and pole profile.

With usual pole shoe profiles, Cl is between 1.0 and 1.1, i.e. the actual flux wave is a bit flatter than a sine wave, and the total flux per pole is rather more than the fundamental flux Φ, alone. Wieseman's factor $k\Phi$ is the total flux divided by the fundamental component. Narrow poles with an edge airgap of about twice the minimum tend to give a peaky waveform with Cl just less than 1.0 and $k\Phi$ down to about 0.93. Wider poles and a uniform airgap give $k\Phi$ *up* to 1.06. The third harmonic flux can be made small by using a pole arc: pole pitch ratio close to 2/3; this together with a ratio maximum: minimum airgap ratio of 1.5 gives a $k\Phi$ *value close to* unity. If Φ is calculated from the e.m.f. equation, total flux and harmonic fluxes can be calculated.

Ginsberg *et al.*[46,47] provide curves from which the harmonic e.m.f.s on load could be calculated, as a per unit value of the fundamental. For detailed analysis, harmonic fluxes now are calculated by computer, e.g. by finite-element solution of the magnetic-field equations.

28.2.2 Open-circuit e.m.f.: integral slots per pole

28.2.2.1 Fundamental frequency e.m.f.

The e.m.f. induced in a single conductor lying parallel to the shaft axis has the same waveform as the flux wave. It contains all the space harmonics in the same proportions as in the flux density. It may also contain ripples caused by local

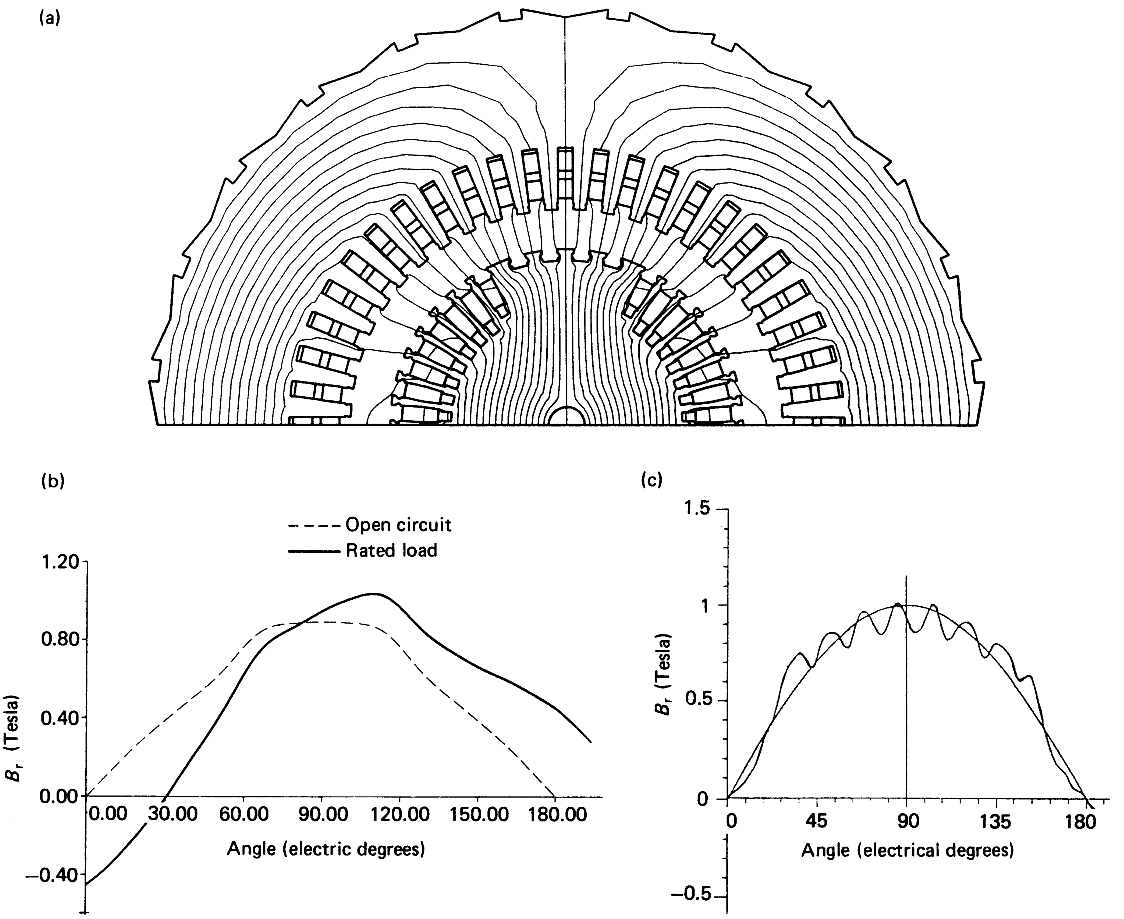


Figure 28.1 (a) Flux distribution of a turbogenerator on open circuit; (b) Flux waveforms of a turbogenerator; (c) Flux waveform of a salient pole generator on open circuit

variations in density caused by the winding slots (see Section 28.2.4). The root-mean-square (r.m.s.) value of the fundamental frequency component of the e.m.f. in a coil of one turn, spanning exactly one fundamental pole pitch, is

$$e_1 = 4.44 f\Phi_1 \text{ (volts)} \tag{28.1}$$

where Φ_1 is the fundamental flux per pole (in webers) and f is the frequency (in hertz).

28.2.2.2 Pitch factor, k_{p1}

If there are N_p slots per pole, and the coil spans S teeth, it is short-pitched by an electrical angle

$$\beta = \left[\left(1 - \frac{S}{N_p}\right)\right]180°$$

and the e.m.f.s generated in each of the two coil sides will be out of phase with respect to each other by β degrees.

Table 28.1 % harmonic contents of flux waves

	Harmonic order							
	Fundamental	3	5	7	9	11	13	15
Turbogenerator								
Open circuit	100	5.5	0.65	2.3	1.4	0.07	0.37	0.4
Rated load	112	13.0	2.34	2.4	1.5	0.2	0.42	0.46
Salient pole generator								
Open circuit	100	6.5	2.4	4.7	3.9	1.2	0.5	1.2
Rated load	108	18.5	3.2	4.8	4.3	2.0	0.4	1.2

The total e.m.f. is, therefore that of a full-pitch coil reduced by the factor

$$k_{p1} = \cos \tfrac{1}{2}\beta \tag{28.2a}$$

or, in another form,

$$k_{p1} = \sin \frac{S}{N_p} 90^\circ \tag{28.2b}$$

28.2.2.3 *Distribution factor* k_{d1}

If there are q slots per pole per phase ($q = N_p/3$ for a three-phase machine), where q is an integer, the phase difference between successive slots causes the sum of their e.m.f.s to be less than q times the e.m.f. per slot by the distribution factor, or spread factor k_{d1}. This is because vectorally each successive coil voltage is at an angle to its predecessor. This angle is determined by the total number of stator slots distributed around the bore of the stator.

$$k_{d1} = \frac{\sin q(\alpha/2)}{q \sin(\alpha/2)} \tag{28.3}$$

where α is the electrical angle between adjacent slots. α is numerically equal to $180^\circ/N_p$ electrical degrees or $60^\circ/q$ for three phases, with the usual 60° coil phase spread. Then, for a three-phase winding,

$$k_{d1} = \frac{\sin 30}{q \sin(30/q)} \tag{28.4}$$

28.2.2.4 *Skew factor*, k_{s1}

If the winding slot, or the longitudinal axis of the pole, is skewed at an angle δ to the shaft axis, there is a progressive change in phase of the e.m.f. along each conductor. The conductor e.m.f., and hence the phase e.m.f., is reduced by the skew factor

$$k_{s1} = \frac{2 \sin(\delta/2)}{\delta} \tag{28.5}$$

Then, for a phase comprising of T_{ph} turns, divided into g parallel circuits, the fundamental frequency e.m.f. is

$$\begin{aligned} E_1 &= 4.44 f_1 \Phi_1 \frac{T_{ph}}{g} k_{d1} k_{p1} k_{s1} \text{ (volts)} \\ &= 4.44 f_1 \Phi_1 \frac{T_{ph}}{g} k_{w_1} \text{ (volts)} \end{aligned} \tag{28.6}$$

28.2.2.5 *Harmonic e.m.f.s*

For the following reasons a lot of attention is now being paid to the harmonic content of the generator's voltage waveform especially with the advent of increasing amounts of private power generation equipment connected to networks.

(1) Limits of telephone interference factor (t.h.f.) are specified in many standards, and the weighting factors emphasise harmonics that can be caused by slot ripples if the design does not suppress them.
(2) Legislation is impending in the form of the Electromagnetic Compatibility Directive (EMC) 89/336/EEC. This will specify limits for the maximum levels of emitted electromagnetic radiation, and of the levels of voltage or current that can be impressed on the power system or on the load connected to the generator.
(3) Generators are being more frequently required to provide a backup supply to uninterruptible power supplies if the mains supply fails. These schemes are used to feed loads that are sensitive to voltage distortion, such as computers, television monitors and control systems.
(4) There is increased concern that harmonic currents may be large enough to cause unacceptable extra heating and losses, both in the generator and in equipment supplied by it.
(5) Inadvertent malfunction of protection and switching equipment can occur particulaly if the equipment is set to trigger at the zero crossing point of the voltage waveform.

It is difficult to estimate how large the harmonic currents arising from the harmonic voltages will be. Their values depend on the impedances of the system and of the generator at the harmonic frequency, and on the nature of the connection between their neutral points.[42]

The lower-frequency harmonics, up to say the 13th order, are produced by harmonics in the main flux wave. These result from the non-sinusoidal distribution of magnetomotive force (m.m.f.) and the non-uniform radial permeance around the circumference of the airgap. Hence they are often called *rotor permeance harmonic e.m.f.s*. The magnitude of each compared with the fundamental e.m.f. depends on the harmonic flux density relative to the fundamental density, and on the harmonic wavelength relative to the spacing and span of the stator coils, i.e. on the winding factors k_d, k_p and k_s at the relevant frequency.

E.m.f.s at frequencies associated with the number of stator slots are caused by: (a) variations in gap permeance as the poles pass the stator slots and teeth; and (b) flux waves produced by currents induced at slot frequency in the damper winding, field winding and solid steel of the rotor. These e.m.f.s are generally called *slot ripple e.m.f.s* (see Section 28.2.4).

Considering only the field form harmonics, and taking q as an integer, the pole span of the nth harmonic flux is $1/n$ of the fundamental span, and the harmonic flux per pole

$$\Phi_n = \frac{\Phi_1}{n} \cdot \frac{B_n}{B_1} \text{ (webers)}$$

where B_1 and B_n are the peak or average flux densities.

The harmonic e.m.f. generated as a result of this flux is

$$E_n = 4.44 f_n \Phi \frac{T_{ph}}{g} k_{wn} \tag{28.7}$$

where

$$k_{wn} = k_{dn} k_{pn} k_{sn}$$

or

$$\frac{E_n}{E_1} = \frac{B_n}{B_1} \frac{k_{wn}}{k_{w1}}$$

Because any angle α on the fundamental scale embraces *na* of the *n*th harmonic, the winding factors become (where *q* is an integer)

$$k_{pn} = \cos\frac{1}{2}n\beta; \text{ or } \sin n\frac{S}{N_p}90^\circ \qquad (28.8)$$

$$k_{dn} = \frac{\sin n\,30}{q\,\sin(n \cdot 30/q)} \qquad (28.9)$$

$$k_{sn} = \frac{2\,\sin[(n\delta)/2]}{n\delta} \qquad (28.10)$$

The r.m.s. value of the total e.m.f. per phase is then

$$E_{pn} = \sqrt{\left(E_1^2 + E_3^2 + \cdots E_n^2\right)} \qquad (28.11)$$

The lower harmonic e.m.f.s are usually a few per cent of E_n and the higher orders smaller still. The total e.m.f. E_{ph} is rarely significantly greater than the fundamental, E_1 although the harmonic e.m.f.s themselves may be troublesome.

For $q = 1$, k_{d1} is 0.966, decreasing steadily to 0.955 for $q = 9$ or more. k_{dn} decreases rapidly with increasing *q*, *leading to a reduction in the* e.m.f. harmonics caused by space flux density harmonics in the field form. Unfortunately, k_{dn} rises periodically to equal k_{d1} for $n = m6q \pm 1$ (where *m* is any integer), i.e. for the slot frequency harmonics. If flux components exist at these frequencies they are usually small, but they appear in undiminished proportion as harmonics in the e.m.f. For example, for $q = 3$

n	1	3	5	7	9	11	13	15	17	19	21
k_{dn}	0.966	0.667	0.217	0.177	0.333	0.177	0.217	0.667	0.960	0.960	0.667

k_{pn} is less than k_{p1} for most slot numbers and coil pitches, but again $k_{pn} = k_{p1}$ at the slot frequencies. For comprehensive tables of k_{dn} and k_{pn}, see references 1 and 22.

k_{sn} can be made nearly zero for the slot frequencies by skewing slots or poles by one slot pitch in the length of the core. Then k_{s1} is still very nearly unity.

In principle, a particular harmonic e.m.f. can be eliminated by choosing a number of slots that allows the coil to be short-pitched by exactly half the harmonic wavelength, i.e. by $1/n$ of the pole pitch, so making k_{pn} zero. In practice, it is very rarely necessary or acceptable to impose this constraint on other design considerations. Most windings have a coil pitch close to 5/6 of the pole pitch, as this usefully reduces both the fifth and seventh harmonic e.m.f.s. This pitch is easily obtainable with any even number of slots per phase per pole.

For certain slot numbers, and harmonic orders, k_{dn} becomes negative. k_{pn} becomes negative for certain pitches and harmonic orders. Thus the harmonic e.m.f. and the harmonic flux that produces it have opposite signs, each relative to its fundamental.

In a star-connected winding, third-order and other triplen e.m.f.s do not appear between the line terminals because they are in phase with each other in all three phases and cancel out. Triplen currents cannot flow unless they have a path via a connection made to the star point. Triplen e.m.f.s act in series round a closed delta winding, and would cause circulating currents, extra losses and, perhaps, overheating. For this reason, and because a star point is usually needed for earthing, generators rarely use delta windings.

Sometimes especially for generators operating at voltages below 1000 V, 2/3 pitch coils are used to suppress the triplens in the phase e.m.f. This allows the star point to be connected directly to earth without giving rise to circulating third harmonic currents. This is especially desirable if multiple generators are to be run in parallel with the neutral point on each machine tied to earth. On higher voltage machines, large circulating triplen currents are avoided by using a high resistance or impedance in the connection of the star point to earth and the more economical and advantageous 5/6 pitch stator winding can be used.

28.2.3 Open-circuit e.m.f.: fractional slot windings

Equation (28.7) applies, but k_{wn} and the harmonic e.m.f.s are less than when *q* is an integer. In a winding with $q = a + (b/c)$ slots/pole/phase (where *a*, *b* and *c* are integers and *b* and *c* have no common factor) each phase has exactly $(ac + b)$ slots arranged in *c* groups in *c* pole pitches. The position of each phase group of *a* or $(a + 1)$ slots relative to its own pole is different for each of the *c* groups. Taking the *c* groups together, the $(ac + b)$ slots are uniformly distributed within an angle of 60° electrical (the coil-to-coil connections allow for the 180° phase difference between adjacent poles). Therefore, the phase angle a_c between the e.m.f.s that are electrically adjacent is $60/(ac + b)$, although the electrical angle between slots that are physically adjacent is

$$\alpha_s = \frac{60}{q} = \frac{60c}{ac + b}$$

i.e. $\alpha_e = \alpha_s/c$ as if the winding had *cq* slots/pole/phase. The $ac + b$ coils must be joined in series to form one repeatable section of the phase.

The winding factors that apply to the fundamental and harmonic m.m.f.s produced by a winding carrying balanced sinusoidal currents apply also to the e.m.f.s generated in it by the fundamental flux wave and any harmonic waves (including those produced by the damper winding, see Section 28.2.4). Walker and Kerruish[38] consider the e.m.f. and m.m.f. Liwschitz-Garik[36,39] considers the m.m.f. of fractional-slot windings. Their formulae for K_{dn} look different, but give the same values for a given winding. Equation (28.9), modified by putting *cq* in place of *q*, gives the same values too, except at the slot frequencies of a fractional-slot winding (see Section 28.2.4 and reference 43). In general. k_{dn}, and k_{pn} are much smaller if *q* is not an integer than if it is. References 34, 35, 36 and 40 describe other methods of analysis and other types of winding.

28.2.4 Slot ripple e.m.f.s[33,43]

The causes of harmonics in the open-circuit e.m.f. wave at frequencies associated with the number of stator slots are:

(1) The presence of harmonics at slot frequencies in the main flux wave produced by the rotor.
(2) Cyclic variations in the flux distribution in the airgap, as the poles pass the stator teeth and slots. There is very little change in the total flux per pole, because change is resisted by currents induced in the damper cage and field winding.
(3) Flux waves produced by currents induced in the rotor (damper cage, solid pole shoes, and field winding). These currents are caused by the ripple produced in the gap flux density by the stator slotting.

The stator e.m.f. ripples are greater if:

(1) The machine has a whole number of stator slots per pole. Ripple e.m.f.s can be greatly reduced by choosing an appropriate fractional-slot winding.

(2) The stator slot opening is large compared with the airgap length. A slot opening to airgap length ratio greater than 2 emphasises the ripple (and the eddy current losses in the damper or pole surface).
(3) The rotor has a low impedance damper cage rather than solid pole shoes, which have higher impedance.
(4) The effective pole shoe arc covers other than a whole number of stator slot pitches. However, flux fringing makes this a rather imprecise relationship.

Walker[43] investigated mathematically the occurrence of slot ripple e.m.f.s, and illustrated his findings with oscillograms in which ripples, or their absence, were related to constructional features of the machine. He concluded that the main cause of such ripples is the presence of slot frequency currents in the damper cage of a salient pole machine, unless it is designed specifically to avoid them. They arise because the radial permeance of the airgap is less opposite each slot than it is opposite adjacent teeth. This imposes on the main flux waveform a ripple of wavelength equal to one stator slot pitch. The peak-to-peak depth of the ripple (its double amplitude) is approximately proportional to the *local* mean density of the flux wave (neglecting iron saturation). Hence the ripple is greatest over the pole shoe arc and dwindles to zero at the quadrature axis. At each slot or tooth position the mean flux density, and the ripple, rise and fall as the poles sweep past. This may make the ripple appear to rotate, but because it is caused by the stator slots, it is fixed in position relative to the stator. Therefore it does not induce e.m.f. directly into the stator winding.

It does, however, induce e.m.f.s at slot frequency into any available rotor circuits, most importantly into the rather low impedance damper cage. The frequency is $2N_\mathrm{p}f$ hertz, where $2N_\mathrm{p}$ is the number of slots per pole pair and f is the synchronous frequency ($2N_\mathrm{p}$ may or may not be a whole number). The number of damper bars is chosen primarily to provide effective damping, and the angular pitch of the bars is usually within 15% of the stator slot pitch. If the bars are symmetrically and similarly placed in all the poles, the ripple e.m.f.s in corresponding bars in successive poles circulate current from pole to pole. This produces a m.m.f. with a wavelength equal to two fundamental pole pitches, pulsating relative to the rotor at $2N_\mathrm{p}f$ hertz and rotating with it. This is equivalent to two waves of half amplitude rotating relative to the rotor at $2N_\mathrm{p}$ times synchronous speed, one forwards and one backwards. Relative to the stator the speeds are, therefore, $(2N_\mathrm{p} \pm 1)$ times synchronous speed, and harmonic e.m.f.s of these two orders are induced in the stator. These e.m.f.s are not reduced, relative to the fundamental, by their pitch and spread factors, if N_p is an integer.

As Walker points out, with a whole number of stator slots per pole pair, the pole-to-pole damper cage currents can be avoided, or at least greatly reduced, by offsetting the bars 1/4 of a stator slot pitch to the left in say the north poles, and 1/4 slot pitch to the right in the south poles. This can dramatically improve the e.m.f. waveform, and is standard practice when N_p is an integer.

A practical point of mechanical design arises. The bars usually occupy 75–80% of the circumferential width of the pole shoe. The slots for the outermost bars must be placed so that they do not cause the pole shoe to be overstressed by centrifugal force. Alternatively, the damper bar slots in this position may be closed or omitted.

Walker also shows that if N_p is odd, or fractional, the backward-rotating field does not occur, and the e.m.f. ripple frequency is $(2N_\mathrm{p} + 1/k)f$ hertz, where k is the denominator of $2N_\mathrm{p}$ when it is expressed in its lowest terms, thus:

$$2N_\mathrm{p} = \frac{N}{p} = \frac{A}{k}$$

where N is the total number of stator slots and p is the number of pole pairs. Second-order slot ripple at $(4N_\mathrm{p} \pm 1)f$ (for integral N_p) or $(4N_\mathrm{p} \pm 1/k)f$ may sometimes contribute significantly to the telephone harmonic factor, because the weighting factor is near its maximum at the corresponding frequencies.

Tables showing combinations of k and $2N_\mathrm{p}$ that may, and those that will not, cause pole-to-pole currents and hence slot ripple e.m.f.s are given in reference 43. These are reproduced, for k up to 6, in *Table 28.2*. If b/c in $q = a + b/c$, is less than 1/4, ripple e.m.f.s are usually tolerably small.

In a turbogenerator, the rotor tooth-tops, slot wedges and the damper winding will carry stator-slot-frequency currents. However, the gap-flux ripple at the rotor surface is not great, because the stator slot opening is less than the gap length, except in smaller machines (say below 30 MW). Although q is usually an integer, it usually is more than 6, so the ripple e.m.f. is rarely objectionable.

Walker has also shown that the uniformly spaced rotor bars, in conjunction with integral slots per pole per phase, may cause additional e.m.f. ripples at frequencies of $[(N'/p) \pm 1]f$ Hz, where N' is a whole number given by the rotor circumference divided by the rotor slot pitch. Again this e.m.f. ripple is rarely troublesome.

In summary, design features used to reduce harmonics in the open-circuit e.m.f. are as follows.

(1) Use a fractional slot winding, choosing the number of slots per pole pair from Table 2 of reference 43 (or *Table 28.2*).
(2) Choose an optimum pole shoe arc, and increase the gap length towards the edges of the pole, to reduce the harmonics in the main flux wave. An arc of 65–70%

Table 28.2 Slots per pole pair that do and do not produce slot ripple e.m.f.s*

k	*Fundamental harmonic*	*Second harmonic*
Slots per pole-pair that produce slot ripple e.m.f.s		
1	$2K$	
2	$K \pm \frac{1}{2}$	K
3	$2K$, $2K \pm \frac{2}{3}$	$K \pm \frac{1}{4}$
4		K, $K \pm \frac{1}{3}$
5	$K \pm \frac{1}{4}$	$K \pm \frac{1}{8}$, $K \pm \frac{3}{8}$
6	$2K$, $2K \pm \frac{2}{5}$, $2K \pm \frac{4}{5}$	K, $K \pm \frac{1}{5}$, $K \pm \frac{2}{5}$
	$K \pm \frac{1}{6}$, $K \pm \frac{1}{2}$	$K \pm \frac{1}{12}$, $K \pm \frac{1}{4}$, $K \pm \frac{5}{12}$
Slots per pole-pair that do not produce slot ripple e.m.f.s		
1	$2\ K \pm 1$	$K \pm \frac{1}{2}$
2	K	$K \pm \frac{1}{2}$, K
3	$2K \pm 1$, $2K \pm \frac{1}{3}$	$K \pm \frac{1}{2}$, $K \pm \frac{1}{6}$
4	K, $K \pm \frac{1}{2}$	$K \pm \frac{1}{2}$, $K \pm \frac{1}{4}$, K
5	$2K \pm 1$, $2\mathrm{K} \pm \frac{1}{5}$, $2K \pm \frac{3}{5}$	$K \pm \frac{1}{2}$, $K \pm \frac{1}{10}$, $K \pm \frac{3}{10}$
6	K, $K \pm \frac{1}{3}$	$K \pm \frac{1}{2}$, $K \pm \frac{1}{6}$, $K \pm \frac{1}{3}$, K

*K is any integer; k is the denominator of the number of slots per pole pair.

of the pole pitch is usual, with a gap length at the pole shoes edges 1.3–1.7 times the minimum gap at the pole centre. With fractional slotting, or integral *q* of 5 or more, a parallel gap will often suffice. This is convenient for solid pole shoes, as it allows the shoes to be skimmed to give the correct rotor diameter after they have been bolted on. With a pole shoe arc of 2/3 of the pole pitch, the third harmonic flux is theoretically zero with a parallel gap, but a graded gap may be chosen to reduce the fifth and seventh harmonics.

(3) Skew the stator slots, or the poles, by one stator slot pitch to reduce the fundamental slot ripple harmonic, or by half a slot pitch for the second order harmonic. Skewing is often used on small machines, but it becomes rather awkward and expensive on large ones.

(4) For a salient pole machine with a whole number of slots per pole pair, offset the damper bars to the left and right alternately on successive poles. An offset of 1/4 of the stator slot pitch each way will largely eliminate the fundamental slot harmonic and will somewhat reduce the second-order harmonic. If necessary, the latter can be reduced further by skewing the stator slots or the damper bars by half a stator slot pitch.

(5) With a whole number of stator slots per pole pair, it may sometimes be worthwhile to offset the poles in pairs to the left and right alternately, in order to avoid ripple caused by slot frequency currents in the field winding. Again, an offset of 1/4 stator slot pitch each way will reduce the fundamental slot ripple.

(6) Make the ratio of the gap length to the width of slot opening as large as other requirements will permit. However, the gap length is governed chiefly by the shortcircuit ratio (approximately $1/X_d$) that is specified, and to make it longer merely to reduce slot ripple would usually be unacceptably expensive.

The most effective and economically acceptable methods are to choose a suitable fractional slot winding and shape the pole shoes (for the lower harmonics): or, if an integral slot winding is selected, it is important to choose the most suitable damper bar pitch and to offset the bars as in (4) above.

Stromberg[40] explains how the fractional-slot winding reduces the harmonic e.m.f.s, gives reduction factors for various winding arrangements and indicates certain special constructions that can be used if a particularly good e.m.f. waveform is required.

28.3 Alternating current windings

In synchronous generators, the a.c. output winding is on the stator, except in some very small machines and some for particular purposes. For example, the a.c. winding and the diodes of a brushless exciter are necessarily on the rotor.

Low voltage machines with voltages up to 4160 V and with outputs of up to 2500 kVA, may have single- or two-layer windings with mush-type coils wound with round enamelled wire secured in semiclosed slots. Larger machines have two-layer windings using diamond-shaped short-pitched coils wound with insulated rectangular copper strip, secured in open slots. Each coil has one of its sides in the top layer and the other in the bottom. The coils and the connections between them are most often of the lap type but can be of the wave type. See *Figure 28.2* for a diagrammatic explanation of these terms.Where such coils would be physically too difficult to manufacture, either because of their size or weight, single half coils, known as bars, are used. Bars may also be of either the lap or wave type. Almost always, in a three-phase winding, each phase-group of coils or bars occupies an arc of exactly or nearly 60° electrical

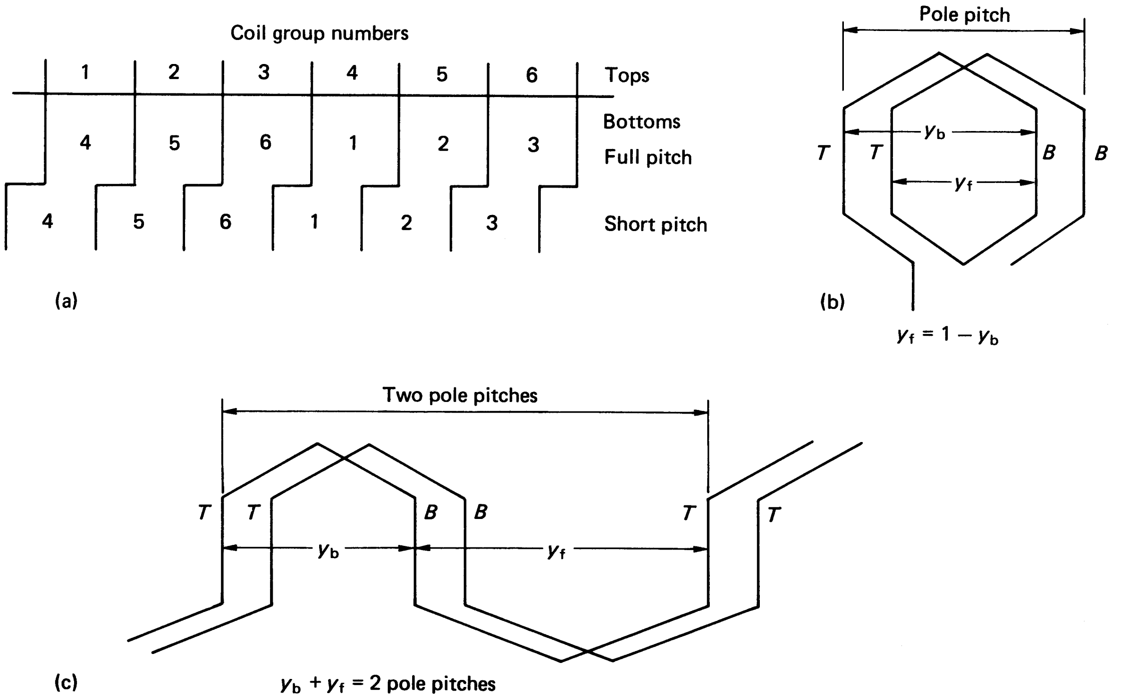


Figure 28.2 Stator coil arrangements: (a) coil groups, *q* an integer; (b) short-pitched lap connection; (c) short-pitched wave connection

under each pole. As explained previously, the number of slots per phase per pole for the stator winding may either be integral or non integral.

Most two-pole and four-pole turbogenerators use a bar type winding and a whole number of slots per pole per phase. Some large low-speed machines use wave-type windings; smaller machines quite often have skewed stator slots.

The two-layer short-pitched coil winding has the advantages that:

1. the coils all have the same shape, which reduces the tooling needed for forming and insulating;
2. a neat endwinding is obtained which is not difficult to support;
3. the coil span can be chosen to reduce harmonic e.m.f.s produced by flux wave harmonics, and to reduce harmonic m.m.f.s produced by the load current;
4. fractional slotting can be used, for the same purpose as in (3); and
5. in multipolar machines, several identical parallel circuits per phase can be formed, giving greater freedom to optimise the design.

Single-layer windings are now rarely used for generators, and are not considered here (see references 11 and 22).

28.3.1 Choice of slot number

Liwschitz-Garik[22] gives a very thorough treatment of this subject, to which acknowledgement is given. Considering only three-phase windings, let:

N = total number of slots = total number of coils
$2p$ = number of poles

$$N_p = \frac{N}{2p}, \text{ slots per pole}$$

$$N_{ph} = \frac{N}{3}, \text{ slots per phase}$$

$$q = \frac{N_p}{3} = \frac{N}{6p}, \text{ slots/pole/phase}$$

g = number of parallel circuits per phase
y = full coil pitch
y_f = front coil pitch
y_b = back coil pitch

See *Figure 28.2* for a diagrammatic explanation of these terms Then:

1. $N/3g$ must be an integer, to provide equal numbers of coils in each parallel circuit of each phase; and
2. the coils must be arranged symmetrically around the airgap, to the same pattern in all phases, to give equal phase e.m.f.s, spaced 120° electrical apart.

28.3.2 Integral-slot windings

The quantity q can have any practicable value, usually between 2 in multipole machines (though these are more likely to have fractional-slot windings) and 10 in large two-pole turbogenerators. The $6q$ coils in $6q$ slots per pole pair are almost always arranged in six groups, each of q coils, each group occupying exactly 60° electrical in each layer. *Figure 28.2(a)* shows the arrangement of full-pitch and short-pitch coils. The pitch is usually about 5/6 of the pole pitch which is chosen to minimise both the 5th and 7th harmonics.

If the left-hand top-layer conductor of each group is called the 'start' of the group, the e.m.f.s acting from start to finish of groups 1 and 4 are 180° out of phase; similarly, for groups 2 and 5, and 3 and 6. Phase A is formed by connecting F1 to F4 to put groups 1 and 4 in series. S1 is the start of phase A, and S4 the finish. Connecting S1 to F4 and F1 to S4 puts the two groups in parallel. Similarly, for phase B (groups 3 and 6) and phase C (groups 5 and 2).

If Phase A starts at slot 1, phase B can start at any slot numbered $1 + 2q + 6nq$, and phase C at $1 + 4q + 6nq$ ($2q = 120°$ and $6nq = 360°$, n being any integer from 0 to p).

In a short-pitched wave winding, $y_b < y_f$, but $y_b + y_f$ is equivalent to two pole pitches. The step of $2p \times (y_r)$ must be one slot more or less than y_f so that a connection can be made to the coil side next to the one at which that tour round the winding started.

28.3.3 Fractional-slot windings

These windings, in which q is not an integer, have the advantage that they can generate as good a waveform with few slots per pole per phase as an integral-q winding with many more slots. Fractional slot windings are frequently used in multipolar machines where there is not room for q to exceed 3 or 4. For example with $q = 3\frac{6}{7}$ the effect of harmonics in the gap flux is reduced as effectively as with an integral value of q of 27, an impracticable number. In this example, it would be necessary to have 14, or a multiple of 14 poles to allow the full pattern of groups of coils to be achieved that are required to give an arithmetic average figure of $3\frac{6}{7}$ slots per pole per phase.

Fractional q also gives the designer a wider choice of slot numbers (for most numbers of poles), enabling flux densities to be adjusted more easily on a given frame size.

28.3.3.1 Arrangement of the coil groups

1. Writing $q = a + (b/c)$ where a, b and c are integers and b and c have no common factor, then c is the least number of pole pitches that will contain a whole number of slots per phase, viz. $(ac + b)$ slots. In the c pole pitches, there are $3(ac + b)$ coils that form a complete three-phase unit of the winding. $2p/c$ ($= F$, say) of these units make up the whole winding, which has $N = (2p)/c\ 3(ac + b)$ slots and coils (of course $N = 6pq$).

   F is the highest common factor of N and $2p$, because

   $$\frac{N}{F} = \frac{N}{2p/c} = 3(ac + b)$$

   and is the number of slots in the unit that occupies c pole pitches, and $c = (2p)/F$. Of course N is a multiple of 3 and of F.
2. A balanced three-phase winding can only be obtained if c is neither 3, nor a multiple of 3. Then, in the c pole pitches in each layer of the winding, each phase occupies b groups of $(a + 1)$ slots, and $(c - b)$ groups of a slots.
3. Starting from slot 1 of each unit of c pole pitches, the larger and smaller slot groups occur in different sequences in the three phases, but the phase e.m.f.s are balanced. The pattern of larger and smaller groups is the same in all F units.
4. Taking all three phases together, the pattern of larger and smaller slot-groups is complete in $c/3$ pole pitches. It occurs $3F$ times in the whole winding, but does not itself form a balanced three-phase unit.

(5) The sequence of a-slot and $(a+1)$-slot groups can be found as follows which is the same for lap as for wave windings.

60° electrical contains q slot pitches. Starting from 0° at slot number 1, the centre lines of the first and last slots of successive slot groups (often called phase belts) must lie within the electrical angles 0–60°, 60–120°, ..., 420–480°, etc., from slot 1. Therefore, the centre line of the last slot of the nth group lies just within nq slot pitches of slot 1, where n is an integer. In general, nq is an integer plus a fraction, say $nq = n(ac+b)/c = X + (Y/c)$. Then, the slot number $(1+X)$ is the last slot of the nth group. But when $n = c$, $2c$, etc., $X = (ac+b)$, $2(ac+b)$, etc., and $Y = 0$. Then slot number $(1+X)$ lies exactly 60n° from slot 1, and so is the first slot of the $(n+1)$th group. *Table 28.3* illustrates the method for a 12-pole machine with 135 slots, $q = 3\frac{3}{4}$, $c = 4$, and $F = 3$. The slot sequence 4, 4, 4, 3 can be determined by considering only c steps of q slots but the complete table allocates all slot numbers to phases.

The fraction b/c determines the numbers of larger and smaller coil groups and their sequence in the winding. References 8 and 22 contain tables of coil group sequences for a wide range of values of b/c.

(6) Within the c pole pitches of the repeatable unit all the conductors of a phase have different positions relative to the poles. Thus their e.m.f.s are out of phase, and all the cq coils must be connected in series. But the e.m.f.s of the F units of a phase are in phase, and can be put in parallel if necessary (see Section 28.3.4).

(7) $2q$ slot pitches span 120°, but the smallest integral number of pitches covering 120° is $2cq$. Slots spaced at multiples of $2cq$ pitches are the obvious starting slots for phases A, B and C. In the example, slot 31 is 120°, and slot 16 is 240° from slot 1. But as with q an integer, the starts can be at other phase groups that are 2 or $(2+6n)$ phase groups apart; e.g. phase B could start at slot 9, although that is 128° away from slot 1. The sum of the coil e.m.f.s is independent of the order in which they are connected (so long as the polarities are kept correct). Starts and finishes can therefore be arranged in the most convenient positions. This can be used to advantage to ensure that the leads from each phase group are distributed around the machine, thus avoiding two adjacent leads having full line-to-line potential between them.

(8) The bottom coil sides have the same pattern as the tops but their groups are displaced to the left or right by the amount of the coil pitch.

(9) Sometimes multipolar windings are used that do not obey all these rules; for example, an existing core design may be usable with a few empty slots or unconnected coils.[8,11,22]

28.3.4 Parallel circuits

28.3.4.1 *Integral slot windings*

Each phase has two groups of coils per pole pair, i.e. there are $2p$ groups in the machine. Whether the coils are lap or wave connected, the groups can all be put in series, all in parallel, or in series–parallel connection. For the latter, the number of parallel circuits in each phase must be a factor of $2p$, say t, so that there can be $2p/t$ groups in series in each parallel path. For example, for $2p = 10$, only two or five parallel circuits are possible.

28.3.4.2 *Fractional slot windings*

In each repeatable section, occupying c pole pitches, the $ac+b$ coils of each phase must be put in series. Hence the maximum number of parallels per phase is the number of sections $F = (2p)/c$. If less are needed, their number must be a factor of $(2p)/c$, so that there will be $2p/ct$ repeatable sections in series in each parallel path. For example, for $2p = 40$ and $q = 2\frac{1}{5}$, there are $40/5 = 8$ repeatable sections and 2, 4 or 8 parallels are possible. With $2p = 40$, but $q = 2\frac{1}{4}$, $(2p)/c = 10$, only two or five parallels are possible.

28.3.4.3 *Concentrated or distributed parallel circuits*

Whether q is integral or fractional, in each parallel circuit the coils that are in series may be arranged under adjacent poles, or may be distributed around the machine under alternate poles. *Figure 28.3* shows two possible arrangements for two parallel circuits in one phase of an eight-pole winding. The concentrated arrangement shown in *Figure 28.3*(*a*) has the advantage that it reduces the unbalanced magnetic pull caused if the rotor is offset radially from the true centre of the stator bore. If, for example, in this 8 pole machine, the airgap becomes narrower opposite pole 2, the gap flux density tends to increase there, and to decrease opposite pole 6. The change in induced e.m.f.s circulates current round the parallel circuits, which tends to reduce the difference in flux densities and so reduce the unbalanced magnetic pull (u.m.p.).

28.4 Coils and insulation

28.4.1 Service conditions

The effects that may eventually cause a breakdown of the insulation of a winding are described below.

Table 28.3 Slot, or coil, grouping for a fractional slot winding*

Pole No.	*1*			*2*			*3*			*4*			*5, etc.*
Slot group No., n	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*	*11*	*12*	*13*
nq	$3\frac{3}{4}$	$7\frac{1}{2}$	$11\frac{1}{4}$	15	$18\frac{3}{4}$	$22\frac{1}{2}$	$26\frac{1}{4}$	30	$33\frac{3}{4}$	$37\frac{1}{2}$	$41\frac{1}{4}$	45	$48\frac{1}{4}$
Slot Nos	1–4	5–8	9–12	13–15	16–19	20–23	24–27	28–30	31–34	35–38	39–42	43–45	46–49
Phase	A	C′	B	A′	C	B′	A	C′	B	A′	C	B′	A
No. of slots	4	4	4	3	4	4	4	3	4	4	4	3	4

*12 poles. 135 slots, $q = 3\frac{3}{4}$ slots/phase/pole, $c = 4$ poles per section, $F = 3$ sections.
Top coilsides 1–4, 13–15, 24–27, 35–38, with their associated bottoms, form one parallel circuit of phase A.

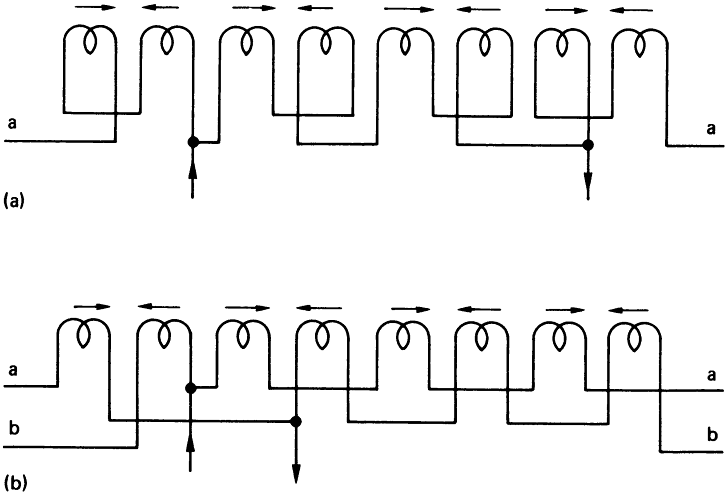


Figure 28.3 Parallel connection of coils: (a) concentrated; (b) distributed

Thermal ageing This affects all windings, and is the main reason for setting limits for winding temperature (see Section 28.5). In smaller machines running at full class H temperatures it is the most usual cause of deterioration.

Electrical stress The trend in stress levels is continually upwards as manufacturers seek to minimise costs and reduce the effect of the thermal barrier of the insulation which is a factor in determining a machine's output. This requires considerable care in the design of the insulation system, in the manufacture of the coils and their insertion in the stator slots. Values of 3 kV/mm are now common with even higher figures being achieved by some manufacturers. The voltages encountered in field windings increase with machine size up to ceiling voltages (see Section 28.14) of around 1500 V on the largest generators. Thyristor excitation systems also superimpose spike voltages sometimes requiring extra insulation.

Electromagnetic forces Stator and field coils are both subjected to forces due to the magnetic fields operating in the machine. Any looseness of an a.c. winding in its slot will quickly lead to mechanical abrasion of the insulation on the coils. On higher voltage machines, equipped with a corona shield, this will lead to slot discharge with eventual failure of the main wall of the insulation causing an earth fault in the slot portion. Field coils able to move on their poles or in their slots will also lead to failure through abrasion.

Mechanical forces Rotor windings carry centrifugal forces that become very onerous in high speed and large machines. Interturn insulation and coil-to-earth insulation placed under the pole shoe and adjacent to supporting 'V' blocks used between poles must withstand the stress caused by centrifugal force and temperature. Non-supported parts of the winding, between 'V' blocks and in the endwinding overhangs are particulaly vulnerable especially in the event of an overspeed. In long rotors, especially two-pole turbine-type rotors, expansion of the winding relative to the rotor body may damage the insulation. Stator windings of long machines may also suffer from differential expansion between the copper and the core. In large generators double frequency vibration caused by electromagnetic forces may abrade insulation or even cause conductor strands to break due to fatigue.

Vibration External vibration, imparted by the prime mover can lead to premature failure of the stator or rotor. In the case of the stator this results from lateral vibrations setting up resonance or fatigue in mechanical components or in the endwinding bracing system. Rotor failures are associated with the effect of running at or near the critical speed of the shaft system leading to high levels of vibration or high tortional perturbations leading to fatigue failures of components.

External electrical fault conditions Mal-synchronising, line-to-line faults or short duration interruptions of the supply can all impart high mechanical and electrical forces on both the stator and rotor windings. Under these conditions, mechanical failure of the stator endwinding bracing, looseness of the winding in the slot or separation of the phase groups can all occur. Rotor field coils suffer from mechanical forces under these faults which can lead to coil connection failures, failure of the insulation or damge to components in the rectifier assembly if a brushlesss machine. Lightning strikes and switching transients arising from the operation of circuit breakers can also lead to failure of the inter-turn insulation in the end groups of the stator windings if the machine is not suitably protected by lightning arresters and/or surge diverters. Similarly, opening a field circuit breaker without suitable discharge resistors fitted can give rise to very high voltages of sufficent magnitude to break down the field coil insulation.

Environmental Contamination of the stator coil insulation by dirt, moisture or oil encourages surface tracking which may eventually become severe enough to penetrate the insulation.

28.4.2 Stator coils[5,11,22,53]

The various types of stator coils most used for generators are described below. The type used depends primarily on the size, output and voltage of the generator, but the choice is, of course, influenced by the maker's facilities and practice.

28.4.2.1 Mush coils

These are used for smaller machines, up to 2500 kVA, at voltages up to 4160 V, in single layer or two-layer windings. Coils are wound with round wire insulated with enamel or enamel and glass tape. The number of turns per coil depends on the size, voltage and frequency of the machine, and the chosen number of slots. The turns are laid at random in slots that may be open, but are often semi-closed, especially if the airgap is short. The main insulation between the coil and the core is a slot liner of a tough but sufficiently flexible sheet, such as polyamide or aramid paper (Nomex). Two-layer windings may have a separate liner for the top and/or bottom coil side (a top or bottom 'box') or a separator between the coils.

28.4.2.2 Diamond coils

These are arranged in two layers in open slots and used for a wide range of outputs and voltages. Coils are wound with rectangular copper strip that is small enough to have acceptably low eddy current losses in service. Sizes are usually within the range 5 mm × 2 mm to 10 mm × 4 mm. The strip is insulated with enamel or enamel plus a thin covering (say 0.2 mm) of lapped or braided glass or polyester-glass plus resin. The number of turns needed are wound into a flat loop on a looping machine and the straight sides are then pulled sideways to form a coil with the required axial length and transverse span. Depending on the equipment used, strip up to about 10 mm wide by 4 mm deep can be pulled. The width-to-thickness ratio should not much exceed 3, to avoid buckling at the bends. If necessary, several strips are wound in parallel to provide the conductor area needed to carry the current, or to make the coil flexible enough to be pulled without damaging the insulation.

Normally a coil has either one or two strips across the width of the slot. Additional inter-turn insulation, enclosing the number of strips that form a complete conductor may be applied after the coil is pulled. However, it is often practicable to apply this insulation as the copper strips are wound on the looping machine, provided that the insulation contains enough resin (see later).

Coils too large to be pulled are shaped on formers. Those for turbogenerators and large hydrogenerators are made as half coils (also known as 'bars') because full coils would be too difficult to handle. In addition, coils made for large 2-pole machines are exceedingly stiff and are difficult to insert through the bore of the machine. *Figure 28.4* shows typical coil cross-sections, all to the same scale. *Figure 28.4*(*a*) shows a typical pulled diamond coil for a diesel-engine-driven 14-pole generator of 9.5 MVA where the conductor strands are not transposed. The coils in (*b*) and (*c*) are formed as bars and incorporate a Roebel transposition which is used to reduce the eddy currents circulating within the full coil (see para 28.4.2.3. below). Coil (*b*) is for a hydrogenerator of 110 MVA at 600 rev/min and (*c*) is for a 590 MVA two-pole turbogenerator. In (*c*) all the strands are squarish tubes to allow for direct water cooling of the conductors: a mixture of tubes and solid strips is often used, especially for the bottom coil side, where eddy current losses are lower than in the top coil side. An appropriate choice of the number and dimensions of tubes and strips can lead to a reduction in the sum of d.c. and a.c. copper losses.

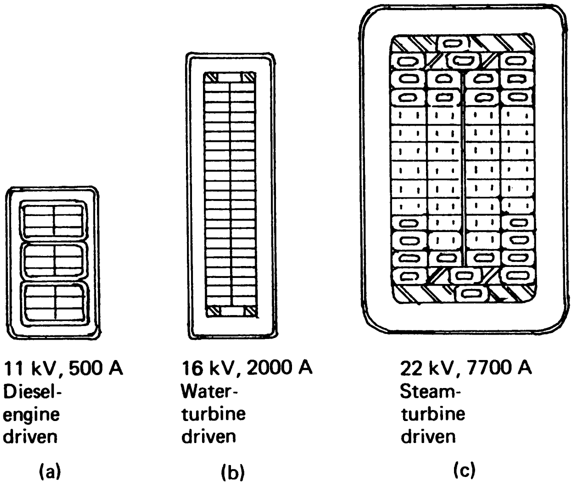


Figure 28.4 Stator coil sections: (a) three turns of six strips; (b) single-bar, two-stack Roebel; (c) single-bar four-stack Roebel

28.4.2.3 Transposition[51,52]

If each turn consists of more than six or eight strips it is desirable, and in larger conductors necessary, to reduce the eddy current losses set up by the leakage flux set up across the slots by the load current. In principle this requires the e.m.f.s induced by the leakage flux to be equal in all strips in the length between points where they are all joined together. In a multistrip coil side for a turbogenerator or hydrogenerator this is achieved within each slot length by using a 360° Roebel transposition. In this construction, each strip is given two edgewise bends separated by half the slot length and differently positioned in the various strips so that when assembled the strips mesh together. When the strips are assembled to form the coil side, each strip occupies each position in the height in turn and they all do so for equal distances. Hence the voltages induced by the cross-slot flux is equal in each and only a small current circulates between strips due to endwinding leakage fields.

In large generators (say 500 MVA or more) the endwinding leakage fields can cause significant circulating currents. The effect can be reduced by using a 540° Roebel transposition. This reverses the positions of the strips in one endwinding relative to the other and approximately cancels the e.m.f.s. arising from the endwinding leakage fields.

In all Roebel coils, the radius of the edgewise bends must be sufficiently generous to avoid damage to the strip insulation. Usually a slip of, for example, Nomex is placed between strips at the cross-overs. The undulating top and bottom surfaces are made level by applying a suitable filler, such as a filled resin dough, which is cured when the conductor stacks are consolidated, before the main insulation wall is applied.

Large turbogenerator slots are often wide enough to require four, not two, stacks of tubes, or strips and tubes. The eddy current losses are rather less if stack 2 is transposed with stack 3, and stack 1 with stack 4, rather than 1 with 2 and 3 with 4. For smaller bars with relatively few strips, and for full coils, the Roebel transposition in the slot is not justified: simpler schemes are possible which are sufficently effective in reducing circulating losses. For example, in a bar winding individual strips can be joined at the coil-to-coil connections in insulated, segregated groups, not all together as is done with a Roebel bar winding. The strips are joined so that within a phase group of coils the groups of strips occupy a succession of positions in the slots that gives a sufficiently good balance of leakage flux e.m.f.s.

With full coils the natural roll-over of the conductors at the noses inverts the strips, i.e. a strip nearest the wedge in the top coil side is nearest the slot bottom in the bottom coil side. Summers[52] gives formulae for calculating the reduced circulating current loss that can be obtained by additionally inverting one or more turns in the overhang portion. He concludes that the most generally useful arrangement is to invert the strips of one turn only at the connection end before forming the last turn; or to invert them at the back end after making only the first half-turn. These sorts of transposition are only effective on 2 or 3 turn coils—above this number it is not necessary to incorporate a transposition in the coil overhangs to reduce the circulating losses to an acceptable value.

28.4.3 High voltage insulation systems[53,58–61]

Two systems[60] are in general use: resin rich (r.r.) and vacuum pressure impregnation (v.p.i.). Both use mica paper (mica flake, which was universally used until the early 1960s, is very rarely used today). The mica paper is bonded with a thermosetting resin to a thin backing material,[54] frequently woven glass or polyester fabric or polyester film. Polyester, epoxy and epoxy-novolac resins, or mixtures of them, are most widely used for r.r. and v.p.i. systems, with appropriate hardeners and catalysts to control the curing process. Resin rich tapes contain about 30% resin, dried to the B stage, at which it can be stored in cold conditions (5–10°C) long enough to be convenient in manufacture. It can be applied at room temperature by hand or by a taping machine; the machine gives better control of lapping and tension. Resin rich tape can be applied as additional turn insulation[56] to a group of strips that is then wound and pulled to form a coil of several turns. Tape for the v.p.i. process contains 4–10% of a different resin system to make it handable and is more fragile than the resin rich tape.

Normally, the strips in the straight sides of the coil, or in the bar, are consolidated in a heated press before any turn insulation or main wall insulation is applied. If the turn insulation (r.r.) has been applied as the pulled-coil loop was wound, this and the strips are cured and bonded at the same time. If turn taping is to applied after this consolidation stage, a release film is put between turns to allow the additional taping to be applied. After any turn taping and the main wall insulation has been applied, the slot portion of the coil is again consolidated under heat and pressure to cure the resin contained in the mica tapes. For turbogenerators and large-hydrogenerators the endwinding portions are consolidated too, both before and after the insulation is applied to improve their thermal conductivity.

In the v.p.i.[59] process, the low-resin content mica paper is applied after the conductor strips have been consolidated. If individual coils or bars are being manufactured, these are placed in an autoclave. Air is drawn out of the tape under vacuum in the autoclave, which is then flooded with low-viscosity resin under pressure. After impregnation, excess resin is drained off, and the coil is pressed to size in a hot press.

Alternatively, for machines up to 5.3 m in diameter and 5 m in length (dependent on the manufacturer's facilities) the global v.p.i. impregnation process can be used.[57] Here, the coils are wound at the so-called 'white stage' into the stator core and the whole wound stator is then placed in a suitable pressure vessel. Absorbent packings and lashings are used in the endwinding bracing structure (e.g. glass or polyester tapes and polyester fleece). Again, a vacuum is drawn to remove any air trapped in the insulation materials followed by a pressure cycle where resin is introduced. Following impregnation, the wound stator is immediately transferred to an oven where it is baked to cure the resin. If treated with an epoxy resin, the stator would be rotated during the baking cycle to ensure both an even distribution and retention of the resin. Stators impregnated with a polyester resin do not require rotation during the curing cycle as the resin gels before becoming sufficently fluid to flow out of the stator. The resin fills the gap between the coils and the core, improving thermal conductivity and permitting some increase in current for a given temperature rise in service. Complete impregnation of the endwinding structure increases its strength and its resistance to moisture, dirt and contaminants.

The r.r. and v.p.i. processes can both produce good quality stator insulation systems. Essentials to achieve a high level of quality are:

(1) uniform lapping and tension of all tapes;
(2) correct choice of resin, tapes etc., to suit the process;
(3) r.r. tapes must be allowed to warm up to room temperature before being applied;
(4) v.p.i. resin must have a low enough viscosity at the impregnating temperature, and an economic life over many cycles of storage and impregnation;
(5) careful control of each cycle of the processes to ensure sound consistent results. In the consolidation stages, the soaking temperature (about 150°C), the heating and cooling rates, and the application of pressure all need careful control. During the v.p.i. process, times, temperatures, the degree of vacuum and the amount of pressure are all critical to achieving a satisfactory result.

28.4.3.1 Electric stress control

At line voltages up to about 5 kV the insulation material and thickness depend very much on mechanical and manufacturing factors. The nominal stress (phase voltage divided by insulation thickness) may be up to about 1.5 kV/mm. At higher voltages electrical stress determines the design, and nominal stress on resin-mica systems is usually 2.5 to 3.0 kV/mm.

At or above 6 kV, the outer surface of the slot part of the coil (whether r.r. or v.p.i.) must be adequately earthed to the stator core to avoid corona discharge in the gap between coil and core. This gap occurs because of the need to have a clearance to allow the coils (or bars) to be wound. The surface may be painted with a conducting paint, but preferably should be taped with a low resistance tape, e.g. a graphite loaded glass or polyester, before the final consolidating press. At the ends of the core the longitudinal stress gradient along the coil surface must be kept low enough to avoid breakdown of the air or surface tracking and eventual failure of the insulation. The surface may be painted with a higher resistance paint than is used on the slot part, but preferably is taped with, for example, a tape loaded with silicon carbide. The length of this stress grading treatment depends on the machine voltage, the insulation thickness and the voltage–current characteristic of the material used. (The resistance decreases with increasing stress.) The treatment must be effective for the short-time high-voltage tests on the coils (greater than twice line voltage to earth) as well as for the long-term operating voltage.

28.4.4 Insulation testing

28.4.4.1 Acceptance tests

Acceptance tests[66] are made on the finished generator to prove that it meets contractual requirements. Tests include:

(1) d.c. insulation resistance (IR), recorded after applying the test voltage continuously for 1 and 10 min;
(2) polarisation index (PI)

$$\text{PI} = \frac{\text{IR after 10 min}}{\text{IR after 1 min}};\ \text{and}$$

(3) high-voltage test with, usually, power frequency voltage, applied after satisfactory IR and PI results have been obtained. The test voltage is maintained for 1 min between each phase and earth in turn, the other phases being earthed.

Procedures and test voltages are specified in BS EN 60034-1:1998 which is equivalent to IEC 60034-1:1994 and replaces BS 4999: Part 101:1987, in ANSI C50.10 1990, NEMA MG1: Part 32-1998, and other national specifications. (see Section 28.21.) The latest issue of the appropriate specification should be consulted for complete information; broadly, the test voltages are, except for low-voltage machines rated less than about 1 kW:

A.c. windings: 2 × rated line voltage + 1000 V.

Field windings:
Rated voltage $V_f \leq 500$; 10 V_f, minimum 1500 V.
Rated voltage $V_f > 500$; 2 V_f + 4000 V.

Document BS EN 50209:1999 specifies tests on conductor bars and coils for machines with rated voltage U_N from 5 to 24 kV. Tests include: (1) measurement of loss tangent (tan δ) and loss tangent tip-up (Δtan δ) on all or some coils of a machine set; and (2) voltage tests on the strand insulation, turn insulation and main insulation. The specification applies to generators of 5 MVA and above, and to 1–5 MVA ratings by agreement. It may or may not be specified as a contractual requirement. The tan δ limits are, for rated voltages U_N of 5–11 kV inclusive:

(1) at 0.2 U_N, tan δ not greater than 30×10^{-3} for any sample coil or bar;
(2) (tan δ at 0.6 U_N–tan δ at 0.2 U_N), $\frac{1}{2} \not> 2.5 \times 10^{-3}$; and
(3) Δ tan δ over any step of 0.2 U_N, $\not> 5 \times 10^{-3}$.

Limits (2) and (3) apply to 95% of the test samples; 5% are acceptable at 3×10^{-3} for item (2) and 6×10^{-3} for item (3).

The measurements must be made at room temperature before the samples are heated to at least 90°C, and again after they have cooled to room temperature. Guard electrodes at the ends of the slot length of the bar exclude the loss in the stress grading from the measurement. R.r. and v.p.i. systems can comfortably meet these limits. Typical values are:

Voltage	$0.2U_N$	U_N
tan δ	10–15	15–20 $\times 10^{-3}$

Most of the increase occurs between 0.8 U_N and U_N. Of course tan δ of a globally impregnated winding can only be measured on complete phases, and the limits stated above cannot apply; comparison with individual coils is necessary.

28.4.4.2 Quality assurance tests

Quality assurance tests are made at the manufacturer's discretion at suitable stages of manufacture. They include:

(1) Dimensional, mechanical and dielectric tests on incoming materials.
(2) Voltage tests such as:
    (a) between insulated strands in a conductor, at 110–250 V r.m.s.;
    (b) between turns of multi-turn coils;
    (c) on individual coils or bars;
    (d) on groups of coils after they have been wedged in the slots, but not connected; and
    (e) on each phase of the completed winding before the acceptance tests.

Supply frequency voltages in (c) and (d) and (e) must be rather higher than the final test voltage; (b) must be an impulse test unless the turns are cut through at the coil nose: it is obviously undesirable to cut a full-wound coil, so a Biddle or other surge tester is used.

BS EN 60034–15:1996 specifies rated phase-to-earth impulse withstand voltages for machines rated 3–15 kV inclusive, with form-wound coils. For the standard lightning impulse, a 1.2/50 μs wave, the rated impulse voltage has a peak value $U_p = 4U_N + 5$ kV, where U_N is the rated voltage. The standard recommends that impulse test voltages should not be applied to a complete machine. It describes test procedures on sample coils, and specifies test levels of ($4U_N + 5$) kilovolts between the conductor and a dummy slot, and half that value between turns, i.e. applied across the ends of the coil.

International discussions are being held to agree upon a withstand level for impulse voltages with steeper wavefronts probably down to rise time of 0.2 μs. In service many generators are protected to some extent from impulse voltages by the impedances of transformers or cables. These reduce the peak voltage and steepness of the wavefront, so the generator does not suffer the full impulse generated by some forms of switchgear. Alternatively, lightning arresters and/or surge diverters may be fitted at the machine terminals to protect the stator windings from fast fronted surges.

(3) IR and PI measurements before (2)(d) and (2)(e).
(4) (a) tan δ test on all or some coils or bars.
    (b) tan δ test on each phase of the complete winding and between phases before impregnation.
(5) Measurement of integrated discharge magnitude using a dielectric loss analyser,[69] on individual coils, phases and the complete winding.
(6) Measurement of the partial discharge value on individual phases or the complete winding.
(7) Measurement of the resistance per square of the corona shield on the slot part of the coil: 2–30 kΩ/square is acceptable.

28.4.4.3 Diagnostic tests in service[67–70,79–82]

The electrical, mechanical and thermal stresses of normal service cause gradual degradation of the insulation. If this general deterioration can be adequately monitored, preventive maintenance of the winding can be co-ordinated with other planned maintenance in an attempt to avoid the cost of failures and unplanned outages. It is true that some generators work in less harsh environments than some motors do, or have closed air circuit cooling arrangements so suffer less contamination, but the cost of an unplanned outage or the damage caused through inadequate maintenance can be appreciably higher. Consequently, more and more attention is now being placed on the need for continuous on-line monitoring and preventative maintenace based on regular inspections for generating plant.

No single test can indicate the extent of deterioration at a particular time. It is necessary to review the results of several non-destructive tests made at reasonably regular intervals, preferably starting with a 'foot-print' from when the machine was first manufactured.

Simons[69] recommends measurements of:

(1) IR and PI with direct voltage at say 1, 2.5, 5.0 kV, appropriate to the machine's rated voltage U_N (r.m.s.):
(2) IR, capacitance C to earth, and integrated discharge energy at rated frequency (50 or 60 Hz) with voltages up to U_N to earth; and
(3) C and tan δ at steps of 0.2 U_N up to about U_N to earth. Tests are made, as usual, on each phase to earth with the other phases earthed, and on the whole winding to earth.

A local defect, unless it is a severe one, will not greatly affect the results. It may, therefore, be desirable to apply a proof h.v. test at say 1.2–1.5 U_N (r.m.s.) to earth at supply frequency, or a 0.1 Hz voltage with a peak of 1.7 U_N (r.m.s.).

28.4.5 Rotor coils

See Sections 28.15.3 and 28.17.2.3.

28.5 Temperature rise

Limits of temperature or of temperature rise are specified in national and international standards (see Section 28.21: BS EN 60034-1:1998, BS EN 60034-3:1996, ANSI C50: Parts 10–15 and NEMA MG1: Part 32).

The limiting values apply at rated load under specified ambient conditions, of which the most important figure is the temperature of the primary coolant (air, hydrogen or water) entering the machine or the winding. Temperatures of stator windings are measured either by resistance for machines rated below 5 MVA, or by embedded temperature detectors (e.t.d.), either of the thermocouple or resistance element type, above this rating or when supplied to NEMA MG1. Temperatures of rotating windings, usually field windings, are measured by resistance. In small machines without e.t.d.s fitted, winding resistance or surface temperatures would be measured on test, and monitoring in service may be of air temperature only. Temperature rises are calculated above the temperature of the primary coolant. Alternatively, but less often, it may be calculated above the temperature of the cooling water entering a heat exchanger.

Machines are designed to comply with the temperature rise limits specified as long as the primary coolant temperature does not exceed 40°C. If the ambient air temperature is high, or there is a water-cooled heat exchanger, the primary coolant temperature may exceed 40°C. Then the design rises are correspondingly reduced, so that the limiting total temperatures are not exceeded. This is especially likely to occur with turbogenerators using turbine condensate, which may enter the heat exchanger at a temperature of up to about 35°C. Conversely, if the temperature of the primary coolant entering the machine is less than 40°C, then the standards allow the temperature rises to be adjusted upwards to maintain the same total temperature. Below a primary coolant temperature of 30°C, any further adjustment is by agreement between the customer and the supplier.

The aim of all standards is to keep the temperature of the winding insulation down to a value at which the insulation, and therefore the generator, will have an acceptably long life. Standards do not specify or imply a lifetime. Accepted norms for life are 20 years for large capital equipment installed in prime power applications. A life expectancy of 10 years is more appropriate for industrial and smaller units. Some machines last much longer than these figures partly because they do not operate for long times near their temperature limits. Large machines or those subject to frequent load changes, or whose reliability is especially important, are commonly specified to meet class B temperature limits, although they have class F insulation in order to achieve a protracted life.

In all windings, the design must allow for the difference between the observable temperature and the hotspot temperature that ages the insulation most rapidly. In an indirectly cooled high voltage stator coil this difference may be 10–20 K, and the innermost insulation may run close to the classification temperature for the insulation system (130°C for class B and 155°C for class F). In an indirectly cooled turbogenerator rotor too, the temperature difference across the main slot insulation is a major component, but the temperature rise of the gas in the 'air' gap, and the surface-to-gas temperature difference, are important. In a salient-pole rotor, the end parts of bare strip-on-edge field coils are well cooled, but down the length of the coils turns can overheat, especially on four- to eight-pole rotors that have 'V' block coil supports between the poles.

Direct cooling avoids the temperature drops through the insulation, and permits higher current densities with acceptable temperature rises, e.g. up to about 10 A/mm^2 in a turbogenerator rotor hydrogen cooled at 5 bar absolute pressure, compared with about 3 A/mm^2 with indirect air cooling. Axially cooled rotors, with ends-to-middle gas flow, must have a modest temperature rise by resistance to avoid excessive temperatures at the midlength. Radial flow, or a combination of radial and axial, gives a mean temperature closer to the permissible hotspot temperature.

In a directly cooled stator coil the best indication of copper temperature is given by the temperature of the coolant, hydrogen or water, where it leaves the coil. The copper temperature is up to about 10 K higher than the hydrogen temperature, and only approximately 1 K above the water. The traditional e.t.d. between coil sides is still used as the routine temperature indicator in service, though some turbo generators are instrumented to measure the outlet temperature of the water from every coil. (For stator windings in new turbogenerators, water has superseded hydrogen, and it is the only choice for hydroelectric generators.) If the water flow fails the copper temperature rises very quickly, say 20 K/min, so the turbine output should be automatically reduced to avoid gross overheating and failure of insulation.

Gas turbine driven generators (see Section 28.21: BS 5000: Part 2 and ANSI C50: Part 14) are specified differently from others because the maximum output available from the turbine changes quite widely as the inlet air temperature changes. The generator must deliver the range of outputs that is the base capability of the turbine over the whole range of air temperature. If the generator is cooled by ambient air, and is allowed to operate up to limits of total temperature, the changing temperature rise allowed as the air temperature changes gives the generator a capability that matches that of the turbine more closely than if the rise were fixed at the rated-load value at all loads and ambient temperatures. Hence a smaller and cheaper generator can be used to cover the turbine rating across the operating temperature range.

The range of still higher outputs that is the peak capability of a gas turbine or internal combustion engine is handled by allowing higher total temperatures in the generator. At these, the insulation ages much more quickly than at standard temperatures. This is accepted because: (1) peak operation demands more frequent engine or turbine maintenance, so will not be used frequently or for long periods, and (2) many internal combustion engine or gas-turbine-driven sets are not expected to have a life of 10 years or more.

If a gas turbine driven generator has water-cooled heat exchangers, its range of permissible temperature rise will probably be less than the ambient air range due to the long thermal time constant of the source of the cooling water. The capability of the generator no longer matches the turbine output so closely and its frame size and design details are determined by the maximum turbine output but using the primary coolant temperature corresponding to the minimum cooling water temperature.

28.6 Output equation

The generator output in apparent power (S volt-amperes) is a function of the stator bore diameter D, the active axial length L, the speed n, the specific magnetic loading B, the specific electric loading A, and the stator winding factor K_w. The output is given by

$$\begin{aligned} S &= 11 K_w B A D^2 L n \text{ (volt-amperes)} \\ &= G D^2 L n \end{aligned} \tag{28.12}$$

where G is the output coefficient ((V-A s)/m^3), B is the average fundamental frequency airgap flux density $((2p\Phi)/(\pi D L)$ tesla), $2p$ is the number of poles, A is the ampere-conductor density $((I_c N)/(\pi D)$ ampere/metre), where N is the total number of stator conductors $= 2T_{ph} \times$ number of phases, or number of slots $\times$ conductors per slot, I_c is the current in each conductor (= phase current/g amps), n is the speed in (rev/s), g is the number of parallel circuits per phase and 11 is a numerical multiplier when all dimensions are in metres. K_w is the product of the individual winding factors. $(K_d K_p K_s)$ and Φ, is the fundamental component of the flux per pole (in webers).

The stator winding is designed to develop the specified voltage with B and A close to chosen values. For different types and sizes of generator, B does not vary widely: it is limited primarily by the degree of saturation that is acceptable in the various parts of the magnetic circuit. Typical values of B are in the range 0.5–0.65 T on no load, corresponding to peak densities in the airgap of say 0.7 to 0.95 T. On load the increase of total flux and the distortion caused by armature reaction will cause the peak density to rise by about 15–20%.

The electrical loading A varies very widely with size and with the intensity of cooling, short-circuit ratio, reactances, etc.

G is numerically more convenient if expressed as

$$\frac{\text{kV-A}}{D^2 L \text{ rev/min}}$$

i.e. in (kV-A min)/m^3. *Figure 28.5* gives typical mean values of G against rated MVA; the values range from 5 to 35, but variations can easily be ±15% for a given output. With G in (kV-A min)/m^3, A is approximately $(6G)/B$ kA/m. Assuming $B = 0.6$ T, A ranges from 50 to 350 kA/m.[27]

The steady increase in G for salient pole generators is a result of the scale effect and progressive improvements in design, mainly in more effective cooling of the stator and field windings. For turbogenerators, a sharp increase in G occurred in the 1950s with the introduction of direct cooled rotor windings using hydrogen at 3 bar absolute pressure. Rapid increase in unit ratings and in G soon occurred, using hydrogen at pressures up to 5 bar absolute and improved methods of circulating it through the windings. It was necessary also to improve stator cooling in order to handle

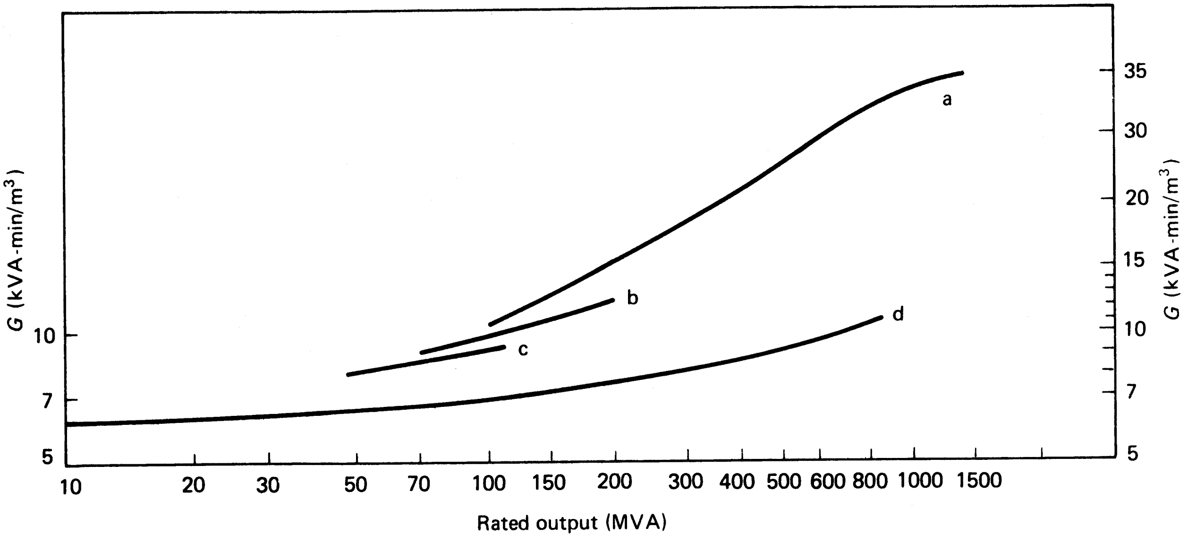


Figure 28.5 Output coefficients, G. Turbine generators: (a) directly hydrogen cooled + stator winding water; (b) directly air cooled; (c) indirectly hydrogen cooled; (d) salient-pole generator

the higher capability of the rotor. Water cooled stator windings have been universally adopted for machines of say 200 MW or more, although direct hydrogen cooling had been used for many stator windings up to 800 MVA output.

In recent years the demand for electric power has grown steadily at the rate of 5% per annum though there has been much public outcry against nuclear power and against acid rain from large fossil-fired stations. Hence there has been almost no call for turbogenerators larger than 500 MW recently, though the recent power shortages on the west coast of America and the insatiable demand for power both in the developing and the developed nations means that there is now renewed interest in nuclear power to meet those needs. In order to maximise efficiency and to satisfy the economics of building nuclear power stations, large generator units possibly up to 1500 MW may well be required in the near future. In recent years, there has been a large number of combined heat and power (CHP) systems have been installed. Here the waste heat from the prime mover, which may be a gas or diesel internal combustion engine or a gas turbine is used for a supplementary purpose such as district heating. This has the effect of increasing the efficiency of the power station. Another possibility to achieve enhanced efficiency has been the use of combined cycle systems where the exhaust gases from a gas turbine are used to raise the steam to be used in a steam turbine located alongside or on the same shaft as the gas turbine. These schemes require generators with ratings up to 150–200 MW at 3000 or 3600 rev/min. For low first cost, and for cheap installation and operation, air-cooled machines have been developed using directly cooled rotors, reaching output coefficients exceeding those of the early low hydrogen pressure machines.

28.6.1 Some design parameters

28.6.1.1 Specifed requirements

Appendix B of BS EN 60034-1:1998 lists information that should be supplied to the manufacturer.

A generator is required to meet specified values of output (kilovolt-amperes, voltage, and power factor) at a specified speed under stated operating conditions, of which the cooling conditions are most important. Specified limits of temperature rise or of total temperature must not be exceeded. The machine must maintain the performance during a life that is specified (or understood by custom and practice), and should do so with reasonable maintenance but without needing major repair. Normally a maximum value of X_d (or a minimum short-circuit ratio) is specified. A minimum $X''_d\Leftarrow$ and or a maximum $X'_d\Leftarrow$may also be specified if currents under fault conditions are to be limited to a certain value or if voltage dips when starting specified loads are to be limited. National specifications recommend preferred voltages (e.g. ANSI C50: Parts 12, 13 and 14), but for large units connected by transformers to the system the generator voltage can be chosen to give the most suitable design.

28.6.1.2 Length to diameter (L/D) ratio

For two- and four-pole cylindrical rotor machines the cheapest design is usually one using the smallest diameter that does not lead to excessive length. The ratio usually lies in the range 3.5–6, perhaps up to 7 for the largest outputs (over 1000 MW) where the strength of retaining ring material limits the diameter to a surface speed of about 220 m/s (1.35 m diameter at 3000 rev/min). With such ratios, two-pole rotors for more than about 10 MW will have first and second critical speeds below the running speed. Higher ratios are avoided because they would lower the critical speed still more and make the rotor very sensitive to small (e.g. thermal) changes in balance.[93]

Cooling also becomes more difficult in very long machines. Salient pole machines with four poles and outputs of, say, 3–30 MVA may have a L/D close to unity but, because there are so many combinations of output, speed, overspeed, reactance, inertia, etc., no simple rule can cover the whole range of outputs. The ratio of L to pole pitch varies rather less than L/D, but still runs from 2.5 to 6 for outputs in the range 150–800 MVA, associated with speeds in the range 500 to 83 rev/min.

28.6.1.3 Stator winding

The most suitable type of winding is usually obvious from the specified current and voltage and the manufacturer's established practice (see Sections 28.3 and 28.4). The cost of making and installing the winding will be smaller the fewer the coils, but too few will cause difficulties. The current per slot, and the copper cross-sectional area, must be such that specified temperature rises are not exceeded. With too few slots, they are necessarily wide, and the total loss per slot may cause excessive temperature rise. With too many slots (and coils) they become narrow and deep, and much slot space is used for insulation; the coils are awkward to form and to support effectively in the endwinding. The ratio of slot depth to width should preferably not exceed 7 (4 – 6 is usual).

The ratio of slot width to airgap length should be about 1½ to 2 with solid pole shoes; with laminated shoes or poles 2½ to 3½ is permissible. Higher ratios increase the tooth ripple e.m.f., and the eddy current losses in the pole surface.[32,33] This is particularly important in salient pole machines with short airgaps especially machines with solid pole shoes. In turbogenerators the airgap is so long that this effect is rarely troublesome.

The tooth width must be enough to avoid excessive magnetic saturation on load, allowing for the radial vent ducts in the core. Roughly, the width of the tooth tip is about half the slot pitch.

28.6.1.4 Short-circuit ratio, synchronous reactance, and rotor temperature rise

The short-circuit ratio (SCR) and X_d are defined with reference to *Figure 28.10*.

$$\mathrm{SCR} =\Leftarrow \frac{0\mathrm{b}}{0\mathrm{d}} \text{ and } X_d = \frac{0\mathrm{d}}{0\mathrm{a}} =\Leftarrow \frac{\mathrm{SF}}{\mathrm{SCR}}$$

where SF is a saturation factor, $\frac{0\mathrm{b}}{0\mathrm{a}}$

which usually has a value between 1.1 and 1.2. X_d is almost inversely proportional to the length of the airgap, and the SCR is approximately directly proportional to the length, l_g. Changing l_g adjusts SCR and X_d, but has only a small effect on other reactances.

For a given output and frame size, increased SCR requires increased excitation, especially if the length of the airgap is increased to raise the SCR, which may cause excessive field temperature rise. Then, unless the cooling can be improved, a lower electrical loading or a larger frame size becomes necessary.

Evidently, a lower SCR permits more output from a given frame size which will be determined by the thermal limit set by the rotor: of course the stator must also be adequately cooled.

As turbogenerator ratings increased, lower values of SCR came to be accepted, partly as a matter of necessity to restrain the increase in physical size, but largely because faster control of excitation and of turbine valves, and shorter fault-clearing times, made it possible to operate safely closer to the stability limit. SCRs fell from the usual range of 0.8–0.6 to 0.6–0.45 for turbogenerators.

Hydrogenerators too have benefited from improved excitation control, power system stabilisers and faster clearing times, but are still required to supply a higher line-charging capability than most turbogenerators. Their SCRs are therefore usually in the range of 0.8–1.5. Other salient pole generators up to say 40 MW (see Section 28.18) have an SCR usually between 0.5 and 0.8. with some up to 1.0 for particular locations.

28.6.1.5 Other reactances (see Section 28.8 for the identification of these reactances)

The stator leakage reactance is a large component of the subtransient and transient reactances. The slot leakage flux contributes most to the leakage reactance, as the endwinding leakage flux is comparatively small, especially in long machines.

Kilgore[84] expresses each reactance as the product of a reactance factor

$$X = \frac{AK_{\mathrm{w}}}{\sqrt{2}\cdot \text{€}_1 B_{\mathrm{g}}}\text{(per unit, p.u.)}$$

and a permeance λ calculated from the geometry of the path of the flux associated with the reactance concerned. For the stator slot, λ is approximately proportional to the slot depth divided by the total width of the $3q$ slots in each pole pitch. Hence for a given output and frame size, a winding with fewer, wider and, therefore, shallower slots has a smaller slot leakage reactance than one with more, narrower and deeper ones. It should cost less too, but the copper will be less well cooled, unless it is cooled directly.

$$X'_{\mathrm{d}} = X_1 + X_{\mathrm{F}}$$

and

$$X''_{\mathrm{d}} = X_1 + \left(\frac{X_{\mathrm{F}}X_{\mathrm{D}}}{X_{\mathrm{F}} + X_{\mathrm{D}}}\right)\Bigg]$$

where X_{F} and X_{D} depend on the dimensions of the pole and the damper cage, and these are chosen primarily to suit the electromagnetic and mechanical design requirements. For most salient-pole machines the ratio of $X'_{\mathrm{d}}/X''_{\mathrm{d}} = 1.4$ to 1.6; for most turbogenerators the ratio is 1.3–1.5.

The electrical design can be manipulated to change X''_{d} and X'_{d} by changing X_1 and there is a little scope for changing X_{D} to alter the ratio $X'_{\mathrm{d}}/X''_{\mathrm{d}}$. Reducing the electrical loading A will reduce X'_1 but will require a larger DL product to increase the main flux with unchanged B. Leaving empty slot above the stator winding increases X'_1, but requires a corresponding increase in the outer diameter of the core to avoid increased back of core flux density. So specifying reactances different from those naturally occurring with the otherwise optimum design will raise the cost.

Some compromise between the values of X'_{d} and X''_{d} may have to be made. A high X''_{d} is desirable to limit the initial short-circuit current and hence reduce the duty of the associated switchgear. A low X'_{d} may be preferred because it would reduce the voltage dip when load is suddenly applied. For example to start a large motor on a fairly small generator or generator group. Or it may help to retain stable operation after a disturbance on the power system. In a generator with laminated salient poles, the damper cage design can be adjusted to bring $X'_{\mathrm{d}}/X''_{\mathrm{d}}$ down towards 1.4, but solid poles and shoes permit almost no adjustment from the usual 1.5–1.6. In turboalternators the rotor tooth tops and wedges, and any damper winding beneath them, are in parallel. Again, the dimensions and wedge materials must first satisfy mechanical and magnetic requirements, and only a very small adjustment of the $X'_{\mathrm{d}}/X''_{\mathrm{d}}$ ratio is possible.

Kostenko[11] and Ames[2] use methods similar in principle to Kilgore's for calculating reactances. Kostenko gives formulae for λ for different slot shapes; Ames presents a detailed treatment of gap leakage reactance that is considered to give better results for short gap designs. Ames also has a useful section on the currents and torques resulting from three-phase and single-phase short circuits, and on voltage dip caused by a suddenly applied load.

Today, analysis using finite element techniques is used to derive reactance figures with increased accuracy.

28.6.1.6 Inertia constant, H

A minimum inertia of the whole generator set may be needed: to help to maintain transient stability during a system fault; to limit the overspeed of a hydrogenerator set; or to limit the cyclic speed irregularity of a set driven by an internal combustion engine.

When H *is* defined as the stored energy at rated speed in watt-seconds, divided by the rated volt-amperes, then in SI units stored energy $= 1/2\,J\omega^2$ metre-newtons. i.e. watt-seconds, and

$$H = \frac{1}{2}\frac{J\omega^2}{\mathrm{VA}}\text{ seconds} \tag{28.13}$$

where J is the polar moment of inertia of the rotor ($= mk^2$ kgm^2, where k is the radius of gyration); ω is the speed (in rad/s); and m is the mass (in kg). The term WR^2 is also frequently used in this context where GD$^2 = 4\times$ WR^2. If this is the case, the term 1.37 in equation 28.14 is increased to 5.48 for an inertia figure quoted as the WR^2.

Alternatively, in usual engineering units,

$$H = \frac{1.37\ GD^2(\text{rev/min})^2}{\text{Rated kVA}}\text{ seconds} \tag{28.14}$$

where GD^2 is the polar moment of inertia (in kg-m^2)

$$GD^2 = m(2k)^2 = 4J$$

Typical values of H for the generator alone are

Salient pole	
Medium speed:	1–2 s
Low speed (100–150 rev/min):	2–4 s
Hydrogenerator	Usually 3–5 s, can be up to 8 s
Cylindrical rotor	
2 and 4 pole	2–4 s

A directly coupled steam turbine may have 2–4 times as much inertia as its generator, but a hydraulic turbine adds relatively little. A hydrogenerator may therefore have to be larger than the electrical performance would require, whereas the turbogenerator designer can usually accept the inertia of the design that meets the other criteria.

28.7 Armature reaction

Balanced three-phase sinusoidal currents in the stator winding produce an approximately sinusoidal m.m.f. wave rotating round the airgap at synchronous speed $n_s = f/p$ rev/sec. The wave continuously changes shape between two extremes, which occur in turn every 1/12 of a cycle of the current. The wave can be represented as a fundamental plus harmonics thus:

$$\text{m.m.f.} = 1.35\frac{IT_{\text{ph}}}{g}\left[k_{\text{w1}}\sin(\omega t-\theta)+\frac{1}{5}k_{\text{w5}}\sin(\omega t+5\theta)+\frac{1}{7}k_{\text{w7}}\sin(\omega t-7\theta)+\cdots\frac{1}{n}k_{\text{wn}}\sin(\omega t\pm n\theta)\right] \quad (28.15)$$

where $1.35((IT_{\text{ph}})/g)k_{\text{W1}}$ ampere-turns per pole is the amplitude of the fundamental wave and θ is a space angle in electrical degrees (one fundamental pole pitch = 180°). Harmonic orders n are $n = 6m \pm 1$ where m is any integer. The sign of n is opposite to that of 1 in $n = 6m \pm 1$.

Each harmonic rotates at a speed inversely proportional to its order; the higher order of each pair, the $(6m + 1)$, goes forwards (i.e. the same direction as the fundamental) and the lower one $(6m - 1)$ goes backwards.

The harmonic fluxes induce small e.m.f.s at supply frequency in the stator winding, not large enough to affect the r.m.s. value of the phase voltage. However, they induce currents at $6m$ times supply frequency in the rotor damper winding, pole faces, etc. These can cause significant extra losses and local temperature rises.[32,33] The fifth and seventh harmonic m.m.f.s are the largest, but with a 5/6 pitch winding, for which $K_{\text{p5}} = K_{\text{p7}} = 0.259$, and spread factors of about 0.2, they are reduced to about 1% of the fundamental m.m.f. The m.m.f. winding factors are the same as those applied to the calculation of open-circuit e.m.f.

Figure 28.6 shows how the gap flux wave shown in *Figure 28.1(c)* is distorted by armature reaction, and shows a small increase in the (mainly slot) harmonics. Harmonic currents supplied by the generator are not often troublesome, unless the load has large capacitance, or resonates to give a low impedance at a particular frequency. The harmonic content in the terminal voltage will depend on the impedances of the load and the generator at the harmonic frequencies.

In fractional slot windings the currents produce also even harmonics and subharmonics: the latter have wavelengths that are multiples of the fundamental double pole pitch. In large multipolar machines such as large hydrogenerators, the subharmonic flux may cause unacceptable deflection and vibration of the stator core, especially if it is a bit shallow radially behind the slots, or has a natural frequency of vibration close to a subharmonic frequency. Some such stator frames have been damaged by this effect (see Liwschitz[36,34] and Walker[38])

28.7.1 Cylindrical-rotor machine

The gap flux wave developed by the fundamental stator ('armature') m.m.f. acting alone is nearly sinusoidally distributed because of the uniform airgap. The gap flux developed by the rotor ('field') m.m.f. acting alone has a trapezoidal distribution. On load the gap flux results from the stator and rotor m.m.f.s in combination. The fundamental components of the distributed m.m.f.s can be represented by phasors of peak values F_a and F_f ampere-turns per pole, respectively, each directed along the corresponding axis of maximum m.m.f. The fundamental component of gap flux on load is proportional to the phasor sum F of F_a and F_f (neglecting magnetic saturation).

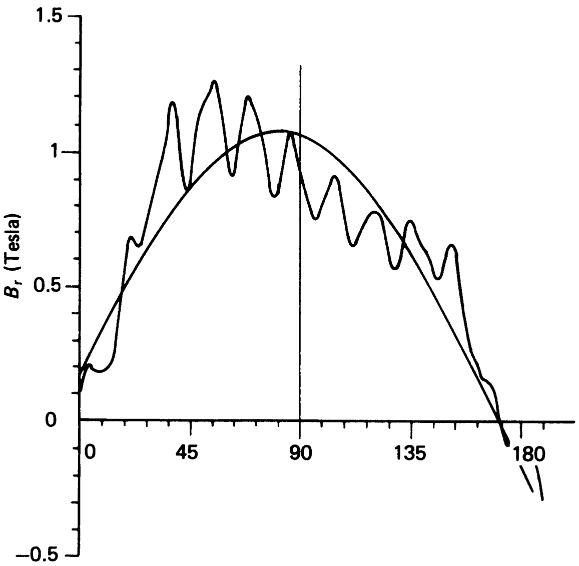


Figure 28.6 Salient-pole flux waveform on load

F_f is centred on the pole axis (the direct or d axis) to which the interpolar axis (the quadrature or q axis) is in electrical space quadrature. In general, the axis of F_a is displaced from the d axis by an angle β depending on the load and the power factor. F_a can be resolved into components F_{ad} and F_{aq}, respectively, on the d and q axes. *Figure 28.7(a)* shows, for a balanced three-phase cylindrical-rotor machine, the stator and rotor current-sheet patterns for an instant of zero current in stator phase C. (The black areas represent outward, and the cross-hatched areas inward, current direction.) The m.m.f. phasors F_a and F_f are displaced by angle, β. The resultant m.m.f. acting on the airgap is the phasor sum $F_f + P_a = F$. Assuming each m.m.f. to develop an individual flux, the e.m.f.s E_f, E_a and E_{ph} in *Figure 28.7(b)* are respectively induced by F_f, F_a and F. Then E_{ph} is the terminal e.m.f. for a representative stator phase.

28.7.2 Salient-pole rotor machine

The gap reluctance is far from uniform, and a given stator m.m.f. acting on the q axis produces less flux than it would if acting on the lower reluctance of the d axis. This is indicated in *Figure 28.7(c)*. The d axis flux is distributed almost sinusoidally, but the q axis flux contains significant space-harmonics, chiefly the third. To deduce the d and q axis fluxes, hence e.m.f.s and reactances, it is necessary to resolve F_a into the axis components F_{ad} and F_{aq}, then to evaluate separately the fluxes they produce. This is done in detail using, for example, finite element analysis of the field. Hand calculation can be done using coefficients presented

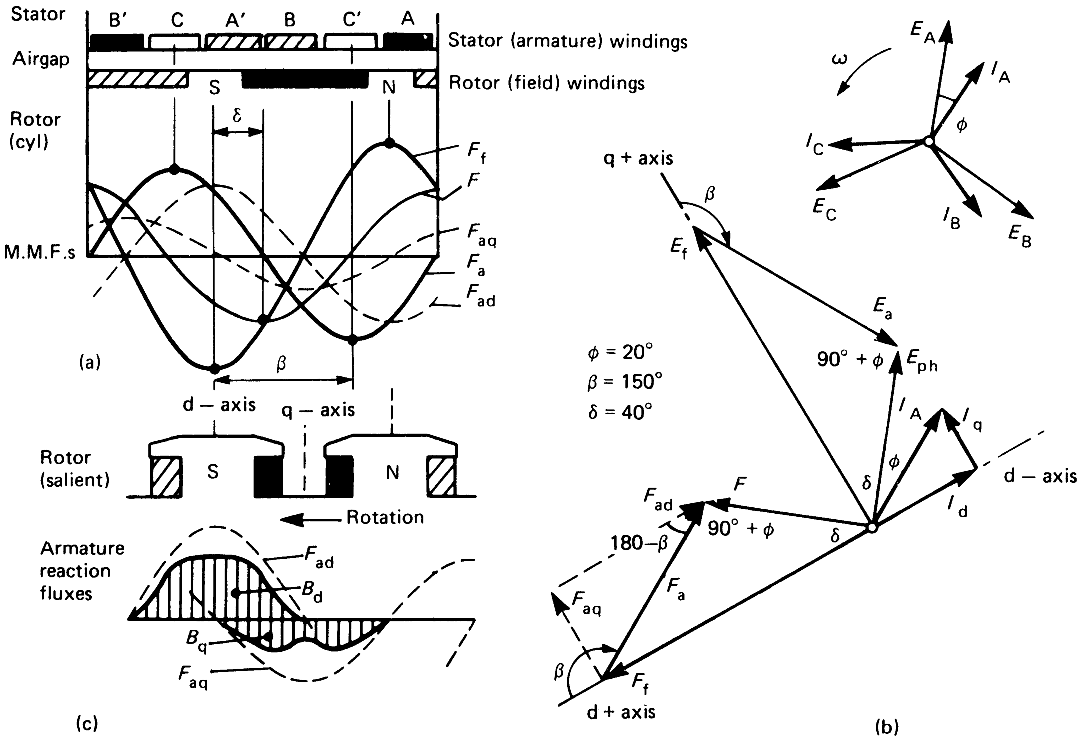

Stator
B′ C A′ B C′ A
Stator (armature) windings
Airgap
S N
Rotor (field) windings
Rotor (cyl)
δ
F_f
F
M.M.F.s
F_{aq}
F_a
F_{ad}
(a)
β
d – axis
q – axis
Rotor (salient)
S
N
Rotation
Armature reaction fluxes
F_{ad}
B_d
B_q
F_{aq}
(c)
ω
E_A
I_A
φ
I_C
E_C
I_B
E_B
q + axis
β
E_f
E_a
φ = 20°
β = 150°
δ = 40°
90° + φ
E_{ph}
I_A
I_q
φ
δ
d – axis
I_d
F
F_{ad}
90° + φ
180–β
δ
F_a
F_{aq}
β
F_f
d + axis
(b)

Figure 28.7 Airgap m.m.f. distribution

by Wieseman[44] and by Ginsberg *et al.*[45] in terms of the pole and airgap profile, and the minimum gap reluctance at the pole centre line.

28.7.3 Magnitudes and equivalence of stator and rotor m.m.f.

For a balanced three-phase winding with 60° phase spread, the peak of the fundamental component of the m.m.f. wave is

$$F_a = 1.35 K_w \frac{I T_{ph}}{gp} \text{ (A-t/pole)} \qquad (28.16)$$

where I is the r.m.s. phase current, g is the number of parallel circuits per phase, T_{ph} is the total number of turns per phase, and $2p$ is the number of poles. With a cylindrical rotor the proportions of the trapezoidal m.m.f. waveshape and the peak of its fundamental component depend on the field form factor C_f which is a function of the ratio

$$\lambda = \frac{\text{Total slotted arc}}{\text{Total circumference}}$$

i.e.

$$\lambda = \frac{\text{Number of slots}}{\text{Number in the circumference if all were cut}}$$

$$\text{e.g. } \lambda = \frac{28}{37} = 0.757$$

$$C_f = \frac{8\pi^2}{\lambda\varsigma} \sin(\lambda \cdot 90°) \qquad (28.17)$$

λ is usually 0.65 to 0.75, and C_f therefore from 1.06 to 1.0. Then the peak fundamental m.m.f. of the rotor is given by $C_f I_f N_f$ A-t/pole, where N_f turns per pole carry the field current I_f.

On load the rotor m.m.f. F_f must be such that when combined with the armature reaction m.m.f. F_a the net m.m.f. F must be sufficient to provide the flux needed to generate the e.m.f. E. Hence F_a must be put in terms of the rotor m.m.f. that would develop the same flux as F_a, i.e. equating the peak fundamental m.m.f.s of stator and rotor:

$$C_f I_f N_f = 1.35 k_w \frac{I T_{ph}}{gp} \text{ (A-t/pole)} \qquad (28.18)$$

Hence

$$I_f N_f = \frac{1.35}{C_f} k_w \frac{I T_{ph}}{gp} \text{ (A-t/pole)} \qquad (28.19)$$

Putting C_f at a typical value of 1.03, the rotor equivalent of the reaction m.m.f. F_a is

$$F_{af} = I_f N_f = 1.31 k_w \frac{I T_{ph}}{gp} \text{ (A-t/pole)} \qquad (28.20)$$

With a salient pole rotor, with a distributed stator winding, a concentrated rotor winding, and a non-uniform gap reluctance, it is necessary in effect to calculate and equate the fundamental fluxes, not the m.m.f.s, along the direct axis.

Wieseman gives curves for estimating the factor C_{dl}, which is the ratio of the fundamental airgap flux produced by armature reaction m.m.f. directed along the direct axis to that which would be produced if the airgap were uniform and equal to the effective gap at the pole centre line. Then the peak fundamental flux density produced on the d axis by F_{a} is proportional to $C_{\mathrm{dl}}F_{\mathrm{a}}$. For typical profiles C_{dl} lies between 0.85 and 0.95.

The peak fundamental d axis flux produced by a rotor m.m.f. $I_{\mathrm{f}}N_{\mathrm{f}}$ is proportional to $C_1 I_{\mathrm{f}}N_{\mathrm{f}}$. The value of $I_{\mathrm{f}}N_{\mathrm{f}}$ that makes $C_1 I_{\mathrm{f}}N_{\mathrm{f}} = C_{\mathrm{dl}}F_{\mathrm{a}}$ is F_{at}, the rotor m.m.f. equivalent of F_{a}. Then

$$F_{\mathrm{af}} = I_{\mathrm{f}}N_{\mathrm{f}} = \frac{C_{\mathrm{dl}}F_{\mathrm{a}}}{C_1} \qquad (28.21)\Leftarrow$$

$$C_1 = \Leftarrow\frac{\text{Peak of fundamental flux wave}}{\text{Peak of actual flux wave}}$$

Putting in typical values of C_{dl}, and C_1

$$F_{\mathrm{af}} = (1.08 \text{ to } 1.2)k_{\mathrm{w}}\frac{IT_{\mathrm{ph}}}{gp}(\text{A-t/pole}) \qquad (28.22)\Leftarrow$$

Figure 28.20 shows how F_{at} is combined with F_{e} to calculate F_{f}, the on-load ampere-turns per pole.

28.8 Reactances and time constants[11,83,84,86–91]

In order to evaluate the steady-state behaviour of a synchronous generator or its response to changes of load, excitation and system disturbances, a mathematical model of the machine is required. Reactances have been defined which, with winding resistances where significant, can form an appropriate equivalent circuit for which behaviour equations can he written; these equations can then be solved to determine the performance of the machine.

The reactances commonly employed are described below. They are associated with a two-axis model, represented on the d axis by an equivalent stator winding, the field winding and a damper winding, and on the q axis by a second equivalent stator winding and one damper winding. Circuit impedances can normally be taken as reactances because the resistances are comparatively small; however, the resistance values directly influence the time constants.

Voltages, currents and reactances are usually expressed in per-unit (p.u.) terms of rated voltage and current; other values are put in p.u. by dividing actual voltage and current by the rated values. The unit reactance is the ratio rated phase voltage/rated phase current. Thus a reactance having a voltage drop of 0.2 p.u. when carrying 0.5 p.u. current has a value of 0.4 p.u. Stator voltages, currents and reactances are all per-phase values.

Each reactance is associated with a particular component of flux produced by either the d or the q axis component of current in the stator winding. A d axis current produces a d axis flux: as shown in *Figure 28.7*, the conductors in which it flows are near the q axis but they form coils magnetising on the d axis. Similarly a q axis current produces a q axis flux. The numerical value of each reactance is then the fundamental frequency e.m.f. per phase generated by the associated flux, divided by the corresponding component of current. Usually reactances are defined with rated current in the d or q axis and are termed ‘rated’ or ‘unsaturated’ values. With the heavy currents occurring under short-circuit conditions, saturation reduces the flux per ampere, and the saturated reactance values are lower.

The reactance values can be derived from the machine geometry by calculating first the permeance of the associated flux path and then the inductance L, i.e. the flux linkage with the stator winding per ampere of stator current in the required axis. The reactance (in ohms) is $2\pi fL$. Difficulties arise in defining exactly the flux paths and permeances and in allowing for saturation at high flux densities, even though a large part of the leakage flux paths is in air.

28.8.1 Armature leakage reactance

The armature leakage reactance X_1 results from stator leakage flux that crosses the stator slots, flux that passes circumferentially from tooth to tooth round the airgap without entering the rotor and flux linking the stator endwinding. X_1 is a component of all the positive- and negative-sequence reactances. Since the flux paths are independent of the rotor, the reactance has the same value for both axes; and since the flux paths are the same for positive- and negative-sequence currents, X_1 has the same value for both.

28.8.2 Magnetisation (armature reaction) reactances

The magnetisation (armature-reaction) reactances are associated with the synchronously rotating flux set up by positive-sequence current in the stator winding, i.e. by balanced phase currents. Suppose the field winding is rotating synchronously but is open-circuited, and a three-phase e.m.f. (V) of correct phase sequence is applied to the stator winding. After the initial damper circuit currents have delayed, a steady stator current I flows. If the e.m.f. is phased so that I produces flux along the pole axis, i.e. $I = I_{\mathrm{d}}$, then I_{d} has a value sufficient to establish an airgap flux that induces a stator e.m.f. $= V$ (neglecting the stator resistance and leakage reactance drops). V/I_{d} is the direct axis magnetising reactance X_{ad}. If the phasing is such that I produces only q axis flux, i.e. $I = I_{\mathrm{q}}$, then V/l_{q} is the corresponding reactance X_{aq}. In most generators, X_{ad} lies between 1 and 2.2 p.u. In salient-pole machines, X_{aq} is typically 0.6 X_{ad}. In cylindrical-rotor machines the airgap is of uniform length, but the rotor slots slightly increase the q axis reluctance, so X_{aq} is usually approximately 0.9 X_{ad}.

28.8.3 Synchronous reactances

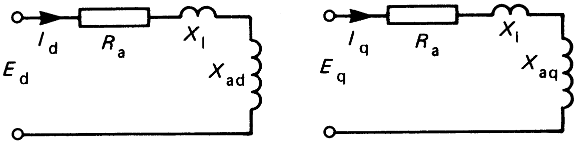
The synchronous reactances are the total reactances presented to the applied stator voltage when the rotor is running synchronously but unexcited:

$$X_{\mathrm{d}} = X_{\mathrm{ad}} + X_1 \text{ and } X_{\mathrm{q}} = X_{\mathrm{aq}} + X_1 \qquad (28.23)\Leftarrow$$

The magnetising and synchronous reactances are steady-state values applicable with balanced phase currents of constant r.m.s. value. They are defined by considering the machine to be excited by stator current only. The two axis reactances can be represented by the equivalent circuits in *Figure 28.8*.

28.8.4 Transient and subtransient reactances

The transient and subtransient reactances relate to conditions that arise when the m.m.f. on the magnetic circuit of the machine is suddenly changed. Consider the conditions



(a) Direct axis **(b) Quadrature axis**

Figure 28.8 Equivalent circuits for d- and q-axis armature reaction reactances

following the sudden application of a three-phase supply of voltage V to the stator winding, with the rotor running synchronously and its field winding closed but unexcited. The three-phase stator currents develop an m.m.f. that rotates synchronously with the rotor. (There are direct-current (d.c.) components as well, but they are not relevant here.) Suppose that the instant of switching is such that the stator m.m.f. is impressed on the pole axis. The flux linkages of the stator winding must induce an e.m.f. that balances V (neglecting resistance). If there were no current paths on the rotor, the stator would immediately carry currents necessary to magnetise the machine to the required flux level, i.e. $I = V/X_d$. However, if there are closed rotor circuits available (i.e. the field winding, damper windings and solid iron poleshoes), currents are induced in them, inhibiting the rise of flux through the rotor poles and so forcing the flux into rotor leakage paths of high reluctance. Hence the initial stator current must be larger than V/X_d. The leakage paths are largely circumferential in the pole faces, from pole to pole in the gap between adjacent salient poles, and across the wedges and tooth tips in a turbogenerator rotor. This path adds only a small permeance to that of the stator leakage paths alone, so the effective reactance is not much more than X_l.

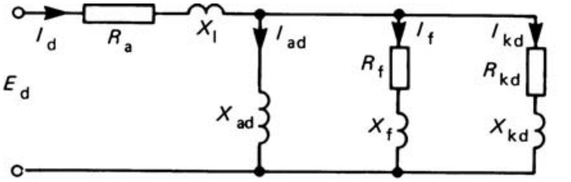
I^2R losses in the damper circuits cause these currents to decay rapidly, enabling the flux to penetrate past the dampers into the pole and field winding region. The permeance of the available flux path therefore increases, decreasing the stator current needed to maintain the required stator flux linkage. Thus the induced rotor currents, and the stator current, decrease in unison, at first rapidly as the damper currents decay, then for a time more slowly, until eventually the induced current in the field winding has disappeared, the flux is fully established along the main flux paths and the stator current is settled at its magnetising value $I = V/X_d$. The main decay time is called the transient period, and the brief initial decay period is the *subtransient*.

The machine as seen from the supply system can be represented by the equivalent circuit shown in *Figure 28.9*(*a*), where I_{kd} and I_f are the components of I_d needed to balance the induced damper and field currents respectively. The effective impedance increases progressively from its initial value of X_r in series with the other three circuits in parallel, through X_l in series with X_{ad} and the field circuit in parallel when I_{kd} has reached zero, to $X_l + X_{ad}$ in the steady state. Thus the transient reactance X'_d and subtransient reactance X''_d are

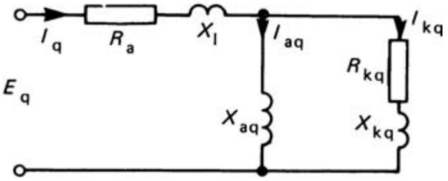
$$X'_d = X_l + \frac{1}{1/X_{ad} + 1/X_f} \tag{28.24a}$$

and

$$X''_d = X_l + \frac{1}{1/X_{ad} + 1/X_f + 1/X_{kd}} \tag{28.24b}$$



(a) Direct axis



(b) Quadrature axis

Figure 28.9 Equivalent circuits for d- and q-axis reactances

If the moment of switching is such that flux is established along the quadrature axis, similar arguments apply except that q axis flux has no net linkage with the field winding. There is, therefore, no q axis reactance corresponding to X_f on the d axis, and in this simple model the q axis circuit is as shown in *Figure 28.9*(*b*).

The q axis subtransient reactance is

$$X''_q = X_l + \frac{X_{aq}X_{kq}}{X_{aq} + X_{kq}} \tag{28.25}$$

In practice, the induced current paths in the iron on the d and q axes change as the flux distributions change, so neither axis can be accurately represented by a single damper circuit, with fixed X and R and therefore one fixed time constant. This is especially true of solid cylindrical rotors, in which the tooth tops and slot wedges form a surface damper cage of relatively high resistance with a time constant typically less than 50 ms. As surface currents decay, lower resistance current paths in the poles and beneath the slots become effective, introducing higher reactances with time constants up to a few seconds. Hence the machine can be represented more closely by having two damper windings on each axis. Traditionally the direct axis reactances X'_d and X''_d and associated time constants have been deduced from oscillograms of a sudden symmetrical three-phase short-circuit test, assuming only one damper circuit. Values appropriate to different levels of magnetic saturation can be found by testing at a number of voltages (tests at or near full voltage are rarely done because they cause very heavy forces on the windings). The short-circuit test measures only d axis values; other tests for these, and for q axis quantities, are given in IEEE Publication 115: 1983 and IEC Publication 34–4 1985; IEEE 115A 1987, describes frequency response tests (see Section 28.8.1).

In a salient pole rotor with laminated poles and specific damper cages, the damper circuits are more clearly defined, and a model with one damper on each axis (as well as the d axis field) is accurate enough for many purposes. Such a model has in the past been used for turbogenerators too, but advances in design and test procedures, and in computer analysis, have encouraged the use of the more elaborate models.

28.8.5 Negative-sequence reactance

Unbalanced load or fault conditions are usually analysed by the method of symmetrical positive, negative and zero phase-sequence components (p.p.s, n.p.s. and z.p.s. components). Currents of n.p.s. in the stator produce an m.m.f. rotating at synchronous speed in a direction opposite to that of the rotor. This m.m f. acts upon the d and q axes in turn, inducing double-frequency currents in any available rotor circuit. The stator presents a low reactance to n.p.s. current, taken to be the mean of the subtransient d and q axis values, i.e.

$$X_2 = \frac{1}{2}(X''_{\mathrm{d}} + X''_{\mathrm{q}}) \qquad (28.26)$$

28.8.6 Zero-sequence reactance

Zero-sequence currents in the three phases are equal and in time phase, and their combined effect is to produce a stationary field alternating at supply frequency, therefore inducing a stator e.m.f. of that frequency. The m.m.f. and therefore the flux are small compared with the p.p.s. and n.p.s. components, and they depend heavily on the coil-span, falling from a maximum value to zero as the span decreases from full pitch to 2/3 pitch. Accordingly X_0 is usually quite small.

28.8.7 Reactance values

Ranges of typical values of reactances and time-constants are given in *Table 28.4*. Since the ranges all depend on the details of the machine design, there will be exceptions to the following generalisations, which do however indicate normal trends.

For a given output and speed, the physically smaller machine will have a higher current loading, so all reactances will be higher than those of a larger, and therefore 'slacker', design. At a given speed, reactances tend to rise with increasing rated output, since higher electrical loading and more intensive cooling are needed to attain more output per unit volume of active material. For a given output, low-speed machines are physically larger, and tend to have higher reactances, than high-speed designs.

28.8.8 Reactances and time constants

The reactances and time constants are based on equivalent circuits, such as those in *Figure 28.9*. Time constants are given by inductance/resistance ratios, i.e. the ratio of $X/2\pi f$ to R. In the formulae below, $2\pi f$ is written as ω, the angular frequency. All time constants are in seconds if all X and R values are expressed consistently in per-unit or ohmic values.

28.8.8.1 Open-circuit

With the stator winding open-circuited, the leakage impedance (X_1 and R_{a}) has no influence and the transient behaviour is determined by the inductance and resistance of the field winding. The open-circuit transient time constant is

$$T'_{\mathrm{do}} = (X_{\mathrm{ad}} + X_{\mathrm{f}})/\omega R_{\mathrm{f}} \qquad (28.27)$$

The subtransient duration depends primarily on damper-circuit currents: if we neglect the effect of R_{f}, the time constant is

$$T''_{\mathrm{do}} = \left[\left(X_{\mathrm{kd}} + \frac{X_{\mathrm{ad}} X_{\mathrm{f}}}{X_{\mathrm{ad}} + X_{\mathrm{f}}}\right)\right] \frac{1}{\omega R_{\mathrm{kd}}} \qquad (28.28)$$

28.8.8.2 Short circuit

With the stator short circuited, current and flux changes are influenced by X_1, but R_{a} is usually negligible. The transient time constant is

Table 28.4 Synchronous generators: typical reactances (p.u.) and time constants (s)

Parameter	*Symbol*	*Turbogenerator*	*Salient-pole generator*		*Compensator*
			With dampers	*Without dampers*	
Synchronous reactance					
d axis	X_{d}	1.0–2.5	1.0–2.0		0.8–2.0
q axis	X_{q}	1.0–2.5	0.6–1.2		0.5–1.5
Armature leakage reactance	X_1	0.1–0.2	0.1–0.2		0.1–0.2
Transient reactance					
d axis	X'_{d}	0.2–0.35	0.2–0.45	0.2–0.45	0.2–0.35
q axis	X'_{q}	0.5–1.0	0.25–0.8	0.5–1.0	0.25–0.8
Subtransient reactance					
d axis	X''_{d}	0.1–0.25	0.15–0.25	0.15–0.3	0.1–0.25
q axis	X''_{q}	0.1–0.25	0.2–0.8	0.5–1.0	0.2–0.8
Negative-sequence reactance	X_2	0.1–0.25	0.15–0.6	0.25–0.65	0.15–0.5
Zero-sequence reactance	X_0	0.01–0.15	0.04–0.2		0.03–0.2
Time-constants					
Time-constants					
D.c.	T_{a}	0.1–0.2	0.1–0.2		0.1–0.2
Transient	T'_{d}	1.0–1.5	1.5–2.0		1.5–2.5
Subtransient	T''_{d}	0.03–0.1	0.03–0.1		0.03–0.1
Open-circuit transient	T'_{do}	4.5–13	3–8		5–8

$$T'_d = \left(X_f + \frac{X_1 X_{ad}}{X_1 + X_{ad}}\right)\frac{1}{\omega R_f} \qquad (28.29)$$

The subtransient short circuit time constant depends primarily on damper-circuit parameters:

$$T''_d = \left(X_{kd} + \frac{X_{ad} X_f X_1}{X_{ad} X_f + X_f X_1 + X_1 X_{ad}}\right)\frac{1}{\omega R_{kd}} \qquad (28.30)$$

The armature short circuit time constant relates to the rate of decay of d.c. components of stator current that occur from the beginning of a sudden short circuit:

$$T_a = \frac{X_2}{\omega R_a} = \frac{X''_d + X''_q}{2\omega R_a} \qquad (28.31)$$

28.8.8.3 q Axis

If the machine is represented by the simple model with only one q axis damper circuit, only subtransient time constants occur. These are

$$T''_{qo} = (X_{aq} + X_{kq})/\omega R_{kq} \qquad \text{open circuit} \qquad (28.32)$$

$$T''_q = \left(X_{kq} + \frac{X_{aq} X_1}{X_{aq} + X_1}\right)\Big/\omega R_{kq} \qquad \text{short circuit} \qquad (28.33)$$

28.8.8.4 Other relations

From the foregoing, it is found that

$$T'_d/T'_{do} = X'_d/X_d \qquad (28.34)$$

$$T''_q/T''_{qo} = X''_q/X_q \qquad (28.35)$$

$$T''_d/T''_{do} = X''_d/X'_d \qquad (28.36)$$

28.8.9 Potier reactance

The Potier reactance X_p is an estimate of armature leakage reactance deduced from the open-circuit and zero-power factor (z.p.f.) curves; it is used in one method of calculating the on-load field current allowing for saturation. X_p is slightly higher than the true X_1, especially for salient-pole machines where pole saturation is greater than on normal load because of the greater pole-to-pole leakage flux for the z.p.f. conditions. Hence using X_p somewhat overestimates the load excitation.

28.8.10 Frequency-response tests[87–89]

The d and q axis parameters can be measured by injecting one-phase current into the stator winding over a frequency range (typically 1 mHz to 1 kHz), the rotor being stationary with its d and q axes in turn aligned with the stator field. By fitting an expression in the Laplace form

$$X_d(s) = \frac{(1 + sT'_d)(1 + sT''_d)}{(1 + sT'_{do})(1 + sT''_{do})} \qquad (28.37)$$

to the curve of d axis reactance against frequency, X_d and the time constants can be found, and the corresponding X'_d and X''_d values derived, appropriate to a machine model with one damper circuit on the d axis.

To fit the q axis reactance-frequency curve reasonably closely, it is necessary to assume two damper circuits, which give an expression similar to the d axis expression quoted, but with q axis quantities. Thus X'_q and T'_q values are deduced, as well as X''_q and T''_q, despite the absence of a field winding on the q axis.

More accurate fits to the measured frequency–response curves may be obtained by using an equivalent with more damper circuits and corresponding time-constants. Techniques have been developed for taking frequency–response curves on the machine in service in order to obtain values more appropriate to the load condition. There is an extensive literature in the *IEEE Journal* (*Power Apparatus and Systems*). A review of the subject, and some specific papers are contained in IEEE Publication 83TH0101-6-PWR, Symposium on Synchronous Machine Modelling for Power System Studies, February 1983.

ANSI/IEEE 115A—1995 describes 'Standard procedures for obtaining synchronous machine parameters by standstill frequency response testing'. It contains eight references, and an appendix showing how operational impedances transfer functions and the *R* and *L* values of the equivalent circuit can be deduced from the measurements.

Measurements of the behaviour of machines and systems when they are subjected to test disturbances have shown that for some stability calculations the simpler circuits, even with the subtransient effects neglected, are adequate. However, for calculating short-circuit torques, the subtransients must be included; and for detailed analysis of the effects of excitation control, more elaborate models are needed.

28.9 Steady-state operation

28.9.1 Open- and short-circuit characteristics

Prediction of the operation of a generator in the steady state is based on the open- and short-circuit characteristics shown in *Figure 28.10*.

With the stator winding on *open circuit*, the field current I_f produces a mutual flux linking the stator and rotor windings, plus a relatively small rotor leakage flux linking the rotor winding only. The corresponding stator winding flux

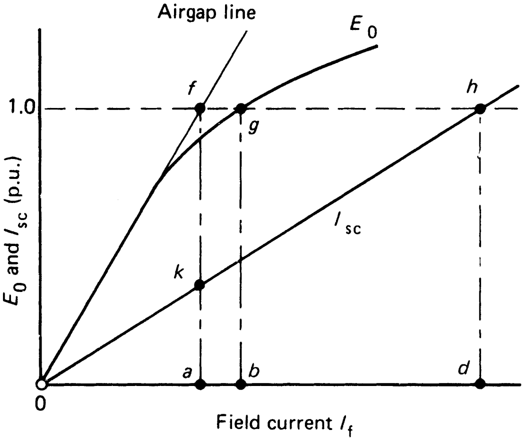


Figure 28.10 Open- and short-circuit characteristics

linkage per phase Ψ_0 generates the stator e.m.f. At rated speed, the value of I_f represented by 0b generates rated e.m.f. E_o represented by bg; 0a is the current needed to overcome the reluctance of the airgap, and ab is required for the iron parts of the magnetic circuit.

With the stator winding *short circuited*, field current 0d circulates rated stator current I_{sc}, represented by hd. Armature-reaction m.m.f. produced by I_{sc} is wholly demagnetising, and the difference between it and the field m.m.f. (0d) develops a mutual flux Φ_{sc}, sufficient to induce an e.m.f. E_{sc} equal to the stator leakage reactance drop $I_{sc}X_1$. This neglects the stator winding resistance R_a and any harmonic fluxes developed by the stator and rotor m.m.f.s.

The flux Φ_{sc} is too small to cause magnetic saturation; hence I_{sc} is proportional to I_f. As both E_{sc} and X_1 are proportional to frequency, the short-circuit characteristic is almost independent of speed; nevertheless, it is usual to obtain it at rated speed.

Still neglecting saturation and armature resistance, a field current $I_f = 0a$ gives $E_o = af$ on open circuit and $I_{sc} = ak$ on short circuit. Thus on short circuit the stator appears to present a reactance $X_{du} = E_a/I_{sc} = af/ak$, a constant representing the *unsaturated* d axis synchronous reactance $X_{ad} + X_1 \cdot X_{du}$ is usually defined in terms of I_f as the ratio 0d/0a in *Figure 28.10*. The short-circuit ratio is defined as 0b/0d which from the geometry is (0b/0a)(l/X_{du}). Here 0b/0a is a saturation factor, usually in the range 1.1–1.2.

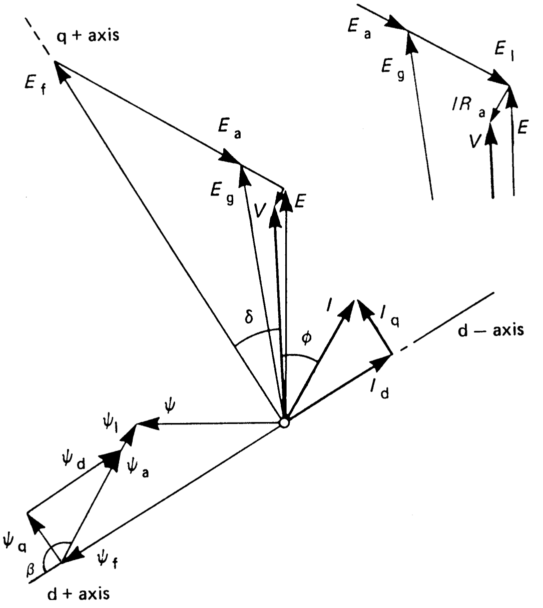


Figure 28.11 Phasor diagram for a cylindrical-rotor generator on load

28.9.2 Phasor diagram and power output

The resultant m.m.f. of the stator current I and the held current I_f develops an airgap flux Φ which induces the stator phase e.m.f. E_{ph}. The stator leakage flux Φ_l induces the e.m.f. E_l.

Detailed analysis allowing for local flux distribution and the variation of magnetic permeability in ferromagnetic parts of the magnetic circuit is necessary in design. However, the performance of a generator on load and under fault conditions can be examined conveniently (and, for many purposes, adequately) by combining the e.m.f.s considered to be produced by I_f alone and I alone, taken separately. Neglecting saturation, the machine can be represented by an equivalent circuit of constant reactances, and with e.m.f.s proportional to their respective currents. Usually the effect of stator-winding resistance can be neglected. Further, by assuming the airgap flux distribution to be sinusoidal, phasor diagrams can be employed. Finally, for a cylindrical-rotor machine the uniform gap length makes it permissible to assume that the d and q axes have equal reluctances. In practice, X_q is usually $0.85X_d$ to $0.95X_d$.

Adopting the conventions of IEC Publication 34–10, the basis is generator action, with power positive when it flows from generator to load. An induced e.m.f. is $e = -d\Psi/dt$, where Ψ is the linkage between flux and stator winding. This means that the e.m.f. phasor lags the flux (or linkage) phasor by 90°⇐electrical. Then, since the flux linking a stator phase winding is in time phase with the current, the phasor diagram in *Figure 28.11* applies for a generator on inductive load, and *Figure 28.12* shows the corresponding equivalent circuit. Here Ψ_f is the stator linkage produced by the field current I_f alone; it would induce on open circuit the e.m.f. E_f. The load current I produces the linkage Ψ_a (the armature-reaction effect) reducing E_f to the airgap e.m.f. E_g. Stator leakage flux produces the linkage Ψ_l and the e.m.f. E_l, and the total induced e.m.f. is E. Subtraction of the volt drop IR_a gives the terminal voltage V. With the convention that the inductive voltage drop leads the current by 90°⇐, the two equivalent equations in phasor terms are

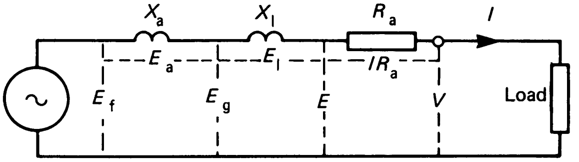


Figure 28.12 Equivalent circuit for a cylindrical-rotor generator

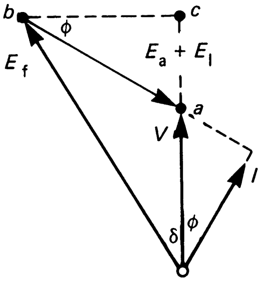


Figure 28.13 Simplified phasor diagram

$$V = E_r + E_a - E_l - IR_a \quad \text{or} \quad V = E_f - I(X_a + X_l) - IR_a$$

If, as is usually permissible, R_a is neglected, then V and E coincide, simplifying the diagram to that in *Figure 28.13*. Here ab on a voltage scale represents IX_d. If the scale is divided by X_d then ab represents I. The angle abc is ϕ, the power-factor angle; bc represents the active-power component of I; but bc is also $E_f \sin \delta$, and hence the power per phase is $(VE_f/X_d) \sin \delta$.

At no load, V and E_f coincide. Hence δ is the *power* (or *load*) angle, i.e. the angle by which the rotor must be driven forward relative to the resultant flux (i.e. forward from its no-load position) to deliver active power. If V and E_f are fixed, then the active-power output P is proportional to sin δ, reaching a maximum for $\delta = 90°$.

For a *salient-pole* machine, account must be taken of the differing d and q axis reluctances. The q axis component current I_q has a lower flux-producing effect than in a cylindrical-rotor machine, i.e. X_q is smaller than X_d. *Figure 28.14* shows the phasor diagram for an output at lagging power factor, with Ψ_l regarded as part of Ψ_a and E_l as part of IX_d. The position of the q axis and the length $0e = E_f$ are found by dividing ab at c such that $ab/ac = X_d/X_q$. The voltage triangle abd is similar to the current triangle I, I_d, I_q, each side being the appropriate current multiplied by X_d. As $ac/ab = de/db = X_q/X_d$, it follows that $de = I_q X_q$, whence

$$0e = V + I_d X_d + I_q X_q = E_f$$

The triangle ade is similar to the flux linkage triangle $\Psi_{ad}\ \Psi_{aq}\ \Psi_q$.

Again δ is the load angle. It is less than that for a cylindrical-rotor machine of the same X_d delivering the same active power at the same voltage and excitation. From the geometry of the phasor diagram it can be shown that the active power output is

$$P = VE_f \frac{1}{X_d} \sin\delta + V^2 \frac{X_d - X_q}{2X_d X_q} \sin 2\delta \qquad (28.38)$$

The second term is the power that is available with zero field excitation ($E_f = 0$) and which is developed as a reluctance torque and power that depend on the different axis reluctances X_d and X_q. *Figure 28.15* shows power–angle curves for a salient-pole machine with typical values of X_d and X_q and for different excitation levels. In a cylindrical-rotor machine X_q is approximately 0.9 X_d, so the reluctance torque ($E_f = 0$) is small, and the power–angle curve is not far from sinusoidal.

28.10 Synchronising

Almost all a.c. generators operate in parallel with others. This raises the problem of switching a machine safely into service ('synchronising') and ensuring that it subsequently remains in synchronism. It is here assumed that a generator is to be connected to a system large enough to fix its voltage and frequency regardless of changes in load and excitation on an individual generator.

28.10.1 Synchronising procedure

The following conditions must be satisfied by the incoming machine with respect to the network bus-bars: (I) the speed must be such that its frequency is close to that of the bus-bars (preferably about 0.2% high); (2) its r.m.s. voltage should equal the bus-bar voltage within ±5% and

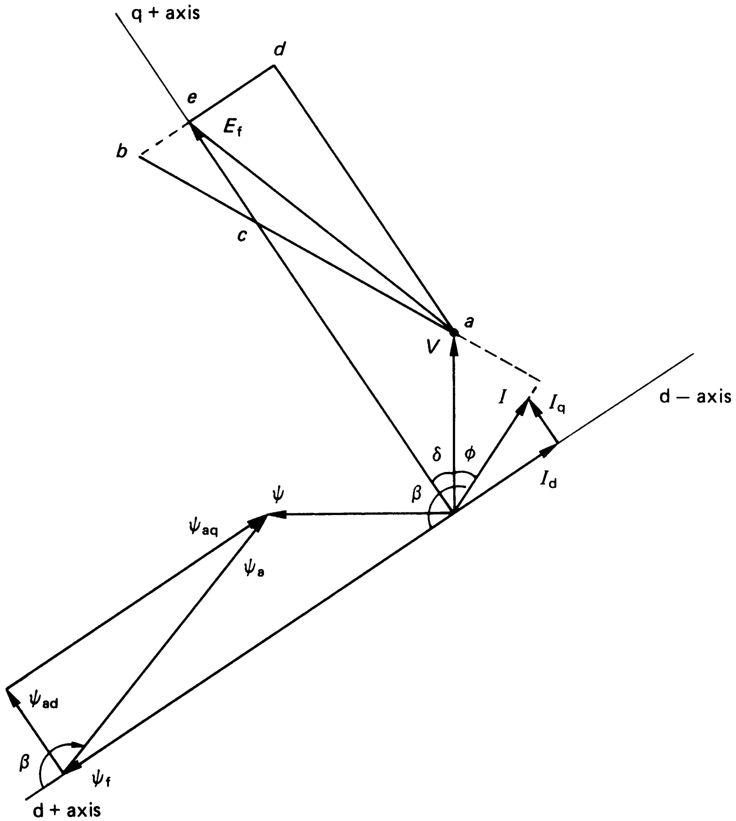


Figure 28.14 Phasor diagram for a salient-pole generator on load

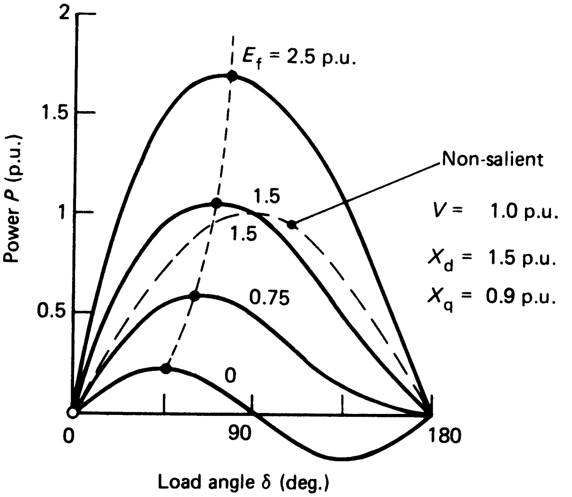


Figure 28.15 Power–angle relationships for a salient-pole generator on load

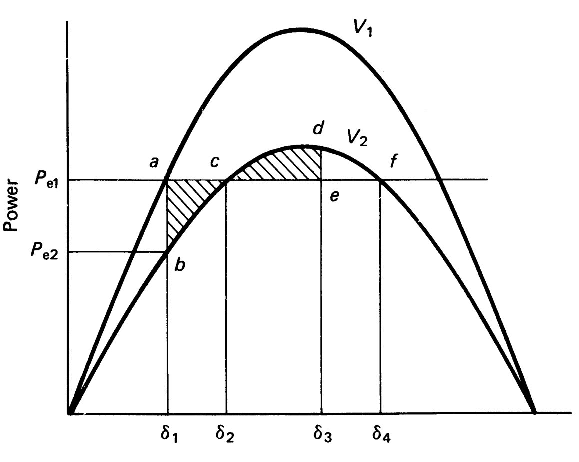


Figure 28.16 Equal-area criterion

(3) the machine and bus-bar voltages must be momentarily in phase, or within ±5° of phase coincidence.

In manual synchronising, condition (3) involves the use of a synchroscope to indicate to the operator the relative phase positions. Allowance must be made by the operator for the 0.2–0.4 s time-lag between initiation of switch closure and the actual closure of the switch contacts.

Alternatively, the process may be carried out by automatic means which monitor the conditions and initiate switch closure at the proper instant.

28.10.2 Synchronising power and torque

If a generator running in parallel with others is disturbed from its steady state, for example by a small change in system voltage, the electromagnetic torque no longer balances the driving torque (presumed constant), and the rotor swings from load angle δ_1 to a new angle δ_2 at which the balance is restored. The equal-area graphical criterion for stability is useful to give a picture of the oscillation process with a single machine,[172,173] but is inadequate for accurate calculation. In *Figure 28.16* the initial operating point a moves to b as the terminal voltage drops from V_1 to V_2, reducing the electrical output. The rotor oscillates between δ_1 and δ_3 before settling at δ_2, again delivering output P_{e1} equal to the mechanical power input. Areas abc and cde represent the energy interchanged between rotor kinetic energy and electrical energy.

If P_{e1} is too close to the peak of the V_2 power curve, area cdf is less than abc, so the rotor can never be slowed to synchronous speed on its forward swing. It rises to a mean speed a little above synchronous, say up to 1%, delivering rather less than the initial load as an induction generator. As the rotor poles slip past the stator m.m.f. wave the speed fluctuates, and there are large swings in current, power, reactive and voltage. An e.m.f. at slip frequency is induced into the field winding, and may reach four or five times the excitation voltage needed at rated load. This forms one of the design criteria for the diodes or thyristors of the excitation system. In a turbine-generator, increased leakage flux in the end regions causes rapid heating of the core ends. E.m.f.s induced between the laminations have in a few machines initiated local damage to the core-plate insulation, and serious core faults have developed in subsequent service. Many machines have pole-slip protection to trip the machine if synchronous running is not restored very quickly.

If $\Delta\delta$ is a small deviation from δ_2, the accelerating, or retarding, torque ΔT is approximately $\Delta\delta$(slope of the T–δ curve at δ_2) $= \Delta\delta T_s$ where T_s is the synchronising torque coefficient. The corresponding synchronising power $\Delta P = \omega_m \Delta T = \Delta\delta P_s$ where ω_m is the synchronous speed (in mechanical rad/s $= 2\pi$ rev/s) and P_s is the synchronising coefficient.

If the rotor swing were very slow the appropriate P–δ curve would be the steady-state curve such as in *Figure 28.15* and

$$P_s = \frac{\mathrm{d}P}{\mathrm{d}\delta} = \frac{VE_f}{X_d}\cos\delta + \frac{V^2(X_d - X_q)}{X_d X_q}\cos 2\delta \qquad (28.39)$$

where P_s, the total for three phases, is in watts/electrical radian if V and E_f are line voltages (not phase) and reactances are in ohms/phase. T_s is then in newton-metres/electrical radian. Usually per unit (p.u.) values are more convenient, using rated apparent power and phase voltage and current as bases. Then P_s is in p.u. power/electrical radian, and T_s p.u. is numerically the same.

28.10.3 Rotor oscillation

The synchronising torque acting on the inertia of the coupled rotors constitutes an oscillatory system with a natural frequency

$$f_n = \frac{1}{2\pi}\sqrt{\frac{T_{\mathrm{spu}}\omega_s}{2H}}\ \text{(hertz)} \qquad (28.40)$$

where $\omega_s = 2\pi f_s = 314$ for $f_s = 50$ Hz and 377 for 60 Hz.

This formula ignores the effect on the frequency of the damping torques and, more significantly, the additional synchronising torques developed by currents induced in damper cages and solid iron.

In practice too, the natural frequency for almost any machine and system lies in the range 1–3 Hz, within which the time of a half-cycle of oscillation is comparable with,

Table 28.5 Comparison of steady and transient states

Calculation Basis	E_f	$E'^{\Leftarrow}$	$\delta\varsigma$ (°)	$\delta'^{\Leftarrow}$ (°)	P_m	$P'^{\Leftarrow}_m$	P_s	$P'^{\Leftarrow}_s$	f_n	$f'^{\Leftarrow}_n$
Steady state	1.58	—	66	—	1.05	—	0.43	—	0.46	—
Transient	—	0.97	—	14.3	—	3.38	—	3.76	—	1.3

or less than, the transient time constant. The field flux linkages do not have time to change significantly, and the effective reactances are the transient, not the synchronous, ones.

With this assumption a round-rotor machine can be represented simply as a constant e.m.f. E' behind the transient reactance $X'^{\Leftarrow}_d$: $E'^{\Leftarrow}$is determined by conditions before oscillation starts, and is assumed not to change as δ', the angle between E' and the terminal voltage V, changes.

Then the peak of the transient power curve is

$$P'^{\Leftarrow}_{\max} = \frac{VE'^{\Leftarrow}}{X'_d} \qquad (28.41)\Leftarrow$$

and at angle $\delta'^{\Leftarrow}_{;}$

$$P' = \Leftarrow\frac{VE'^{\Leftarrow}}{X'^{\Leftarrow}_d}\sin\delta\varsigma^{\Leftarrow} \qquad (28.42)$$

The synchronising power coefficient

$$P'_s = \Leftarrow\frac{VE'}{X'_d}\cos\delta\varsigma' \qquad (28.43)$$

in p.u./electrical radian.

Calculations from the steady-state and transient representations are compared in *Table 28.5* for a generator set with $X_d = 1.5$, $X'_d = 0.25$ and $H = 8$ s, with an initial load of 1.0 p.u. at 0.95 p.f. leading.

In many investigations this approximate model may be good enough, at least as a starting point. For large swings, and to get more accurate results, an appropriate equivalent circuit representation is needed. The damping and induction torques can be included, as also can the influence of fast-acting excitation systems. Computer solution of the resulting equations is necessary. The effect of series reactance X_e between the machine and the fixed voltage bus is to reduce the synchronising torque and lower f_n. X_e is allowed for by adding it to the machine reactances.

28.11 Operating charts

Operating charts based on *Figure 28.11* or *28.14* define the operating limits imposed by the prime mover, excitation, load and stability. Saturation is ignored and the reactances are deemed constant.

28.11.1 Cylindrical-rotor generator

In *Figure 28.17*, 0ab is the synchronous reactance triangle. The fixed phase voltage V is represented by 0a to a scale of say v V/cm. Point b represents rated load, where ab is the rated stator current to a scale of v/X_d A/cm, at the power factor cos dab. ad and ag are the active and reactive components of ab.

0b represents E_f, the phase e.m.f. behind X_d. The corresponding value of I_f can be read off the airgap line of the open-circuit characteristic shown in *Figure 28.10*; the value of I_f is not seriously inaccurate at leading power factors where magnetic saturation is low.

To a scale of $(3Vv)/X_d$ V-A/cm, ab represents rated apparent power, ad the active power, and ag the reactive power.

a0 represents, at v/X_d A/cm, the magnetising current V/X_d drawn from the system at no load if I_f is reduced to zero, and also the corresponding vars V^2/X_d at the scale $(3Vv)/X_d$ V-A/cm.

The P and Q axes are usually marked in megawatts and reactive megavolt-amperes, respectively, or in p.u. terms where, if $V = 1$ p.u., ab =⇐rated apparent power = 1 p.u., ad =⇐active power, cos $\phi\varsigma$p.u. ag = reactive volt-amperes, sin $\phi\varsigma$p.u. and a0 = rated MVA/X_d = $(1/X_d)$ p.u.

Operation must be so controlled that the operating point is within the boundary set by (i) an arc of centre a and radius ab representing rated stator current, (ii) an arc bh of centre 0 representing rated field current, and (iii) the line bdb$_1$ representing the rated active power output of the prime mover.

The line 0m, corresponding to a load angle $\delta\varsigma$= 90°, shows the theoretical maximum power (since the output falls for $\delta\varsigma$= 90°), while 0b$_1$ is the lowest field current for which rated power can be delivered, the corresponding stator current then being ab$_1$.

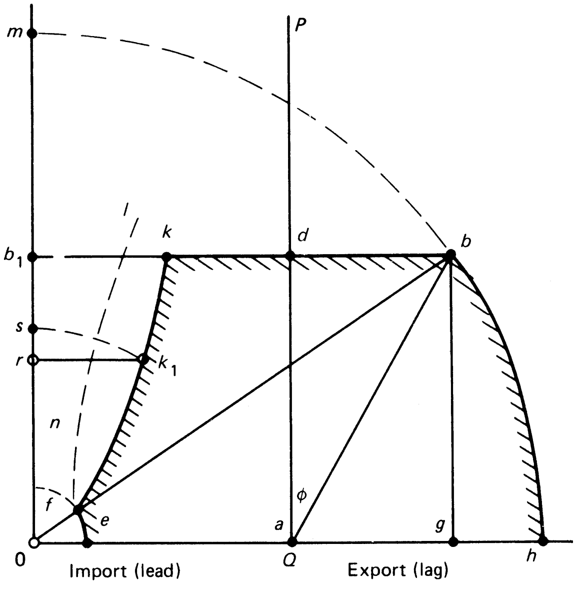


Figure 28.17 Operating chart for a cylindrical-rotor generator

Instability at all loads occurs when the generator (being underexcited) absorbs the reactive power 0a of value $1/X_d$ p.u. The line 0a to the current scale is the zero-power-factor component of stator current; considered as delivered to the power system, 0a leads the terminal voltage V.

Stable operation in practice is not possible up to the theoretical steady-state limit line 0m. It is usual to construct a practical stability line such as 0fk on which, from each point such as k_1, operation for a given excitation $0s = 0k_1$ is not permissible in the region rk_1. The load increment rs may be a fixed fraction of rated load or may change progressively between no load and full load. Often, a minimum acceptable field current is defined to avoid pole-slip at low load when a system voltage drop occurs (for the synchronising torque depends on VI_f). The minimum excitation is typically 20% of that needed on no load giving the limiting arc ef.

With a constant active-power output, the stator leakage flux in end-winding spaces increases as the power factor becomes more leading. This increases the loss, and the temperature in the end core packets and clamping plates will rise. Hence an end-heating limit line ln may be specified. Either the practical stability or the end-of-core temperature may set the limit on the reactive power that can be absorbed.

The operating chart can form the dial of a P–Q meter having a pointer moving parallel to each axis. Where the pointers cross indicates the load point, and the margins between the output and the several limits are readily observed.

The effect of reactance X_e (of, say, a transformer or power line) between the generator and the fixed-voltage bus-bar is allowed for by adding X_e to X_d and X_q. Then for a cylindrical rotor machine the steady-state stability limit is reached where $\delta_B = 90°$: the reactive power Q_L is $V_B^2/(X_d + X_e)$ and the maximum real power P_L is $V_B E_f/(X_d + X_e)$. At any stable load condition

$$P = \frac{V_B E_f \sin \delta_B}{X_d + X_e} \qquad (28.44)\Leftarrow$$

and

$$Q = \frac{V_B(E_f \cos \delta_B - V_B)}{X_d + X_e} \qquad (28.45)\Leftarrow$$

The phasor diagram in *Figure 28.18*(*b*) shows that the presence of X_e requires the generator to operate over a range of terminal voltage. Generators are usually designed to deliver rated megavolt-amperes at rated power factor over a voltage range of ±5%, and a frequency range of ±2% or so. As V_T and f move away from the rated (1 p.u.) values, the rotor or stator temperature rises will increase. If V_T or f goes outside the design range, load may have to be reduced to avoid unacceptably high temperatures. See, for example, IEC 34–3 (Section 28.21).

To provide a voltage range of, say, 10% rather than 5% adds significantly to the size and cost of a large machine. In service the p.u. reactances increase as the voltage decreases, and this reduces the stability margin. Hence many generator transformers have on-load tap-changers, to reduce the range of generator voltage needed.

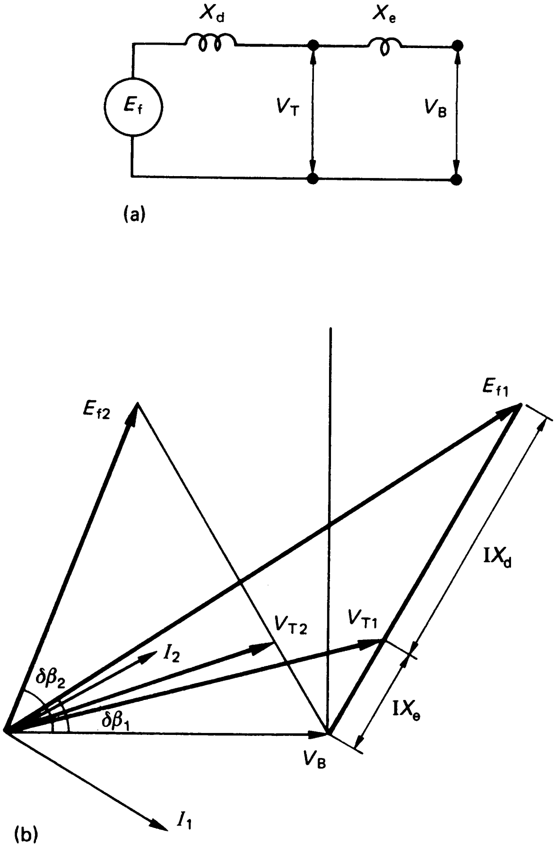


Figure 28.18 Effect of system reactance

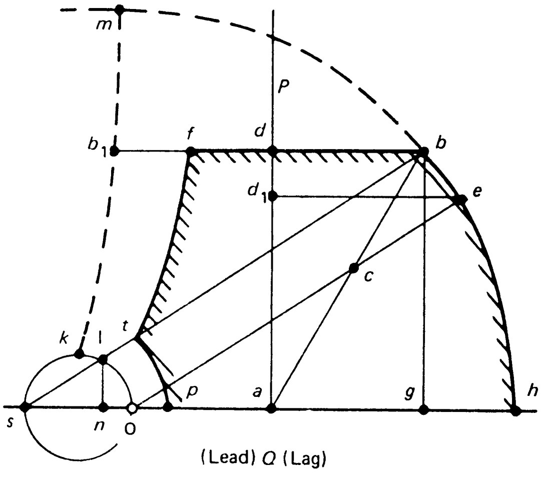


Figure 28.19 Operating chart for a salient-pole generator

28.11.2 Salient-pole generator

From the phasor diagram the P–Q chart for operation at fixed V_T of 1 p.u. can be drawn as in *Figure 28.19*; ab, ad and ag are the stator current or volt-ampere quantities, as before; $ac/ab = X_q/X_d$; $a0 = 1/X_d$; $as = 1/X_q$; $be \perp oe$; $slb \parallel oe$; and $lb = oe$. Hence $ols = 90°$; and the semi-circle constructed on diameter 0s represents a zero-excitation boundary. Angle bs0 = e0a = the load angle δ. At any load, E_f and I_f are represented by a length such as lb, on the line through s. (For any constant E_f and I_f, the locus of b is not quite a circular arc, but this becomes evident only with low excitations near the stability limit.)

ad$_1$ is the power contributed by E_f, and $d_Id = ln$ is the reluctance power derived from the difference between X_d and X_q.

The theoretical steady-state stability limit is the line skb$_1$m. One method of finding the line is to use the expression $dP/d\delta = VE_f \cos\delta/X_d + V^2(X_d - X_q)\cos 2\delta/X_dX_q$ to find consistent values of E_f and δ that make $dP/d\delta = 0$ (or see reference 8, article 6.5). A practical limit such as ptf can be constructed as for the round rotor generator, but with modern excitation controllers operation even beyond the theoretical limit, or with negative excitation, may be accepted, at least as a temporary condition.

With external reactance X_e, the intercepts of the zero-excitation circle lie at a0 $= 1/(X_d + X_e)$, and as $= 1/(X_q + X_c)$. For a given E_f the peak power at the stability limit is also reduced.

28.12 On-load excitation

Use of the constant unsaturated value of X_d leads to values of E_f higher than those that occur for a practical machine. Thus a cylindrical-rotor generator with $X_{du} = 2$ p.u. and carrying rated load at p.f. 0.85 lagging would have $E_f = 2.67$ p.u. while a practical machine would saturate at around 1.5 p.u. If I_t were read from the airgap line for $E_f = 2.67$ p.u. it would be about 15% low. At leading power factor, for which saturation levels are low, the error would be small. However, a more accurate estimate of the field current I_f is needed in the design of the excitation system and its cooling, and to determine the open-circuit e.m.f. that would be reached if the automatic voltage regulator failed to limit the excitation on sudden load rejection.

Flux distributions can now be calculated in considerable detail using computer programs that contain information on the geometry of the magnetic circuit, m.m.f.s of stator and rotor currents and values of iron permeabilities appropriate to the local flux densities. Hence, open-circuit curves and excitation on load can be calculated without much labour from design data once a program has been proved. However, methods based on phasor diagrams and adjusted reactances, making separate allowance for saturation, are still of value if a suitable program or the detailed design information, is not available. They are also needed when calculating excitation from test results on built machines.

Such methods add the rotor m.m.f. phasor needed to generate the no-load voltage to that needed to balance armature reaction. They differ in the choice of the voltage and in the way allowance is made for the effects of saturation.

The open-circuit short-circuit and zero-power-factor characteristics (o.c.c., s.c.c. and z.p.f.c.) are required, either by test or from design calculations. All the methods should give the full-load excitation to within ±5%, or closer at leading power factor. We consider three methods here.

28.12.1 M.m.f. phasor diagram (*Figure 28.20*)

The e.m.f. E behind the leakage reactance X_1 requires a field m.m.f. F_c, read from the calculated or tested o.c.c. F_{ar} is the armature-reaction m.m.f. in rotor terms, obtained by using the equations in Section 28.7 or (if X_1 is known) by calculation from the tested s.c.c. as follows. To circulate stator current I on short circuit, excitation I_n is needed; to generate an e.m.f. to balance IX_1 drop, I_{f2} is needed; hence the armature-reaction m.m.f. is $F_{af} = I_{f1} - I_{f2}$. E_f is the excitation m.m.f. required for the load current I at terminal voltage V and p.f. angle ϕ. *Figure 28.20*(*c*) shows the diagram for a salient-pole machine with a leading p.f. load. As in *Figure 28.14*, ab is divided at c such that ac/ab $= X_q/X_d$ to find point *f*.

28.12.2 The ANSI Potier reactance method (*Figure 28.21*)

ANSI Potier reactance method (ANSI/IEEE Publication 115, Section 4) requires the tested o.c. and z.p.f. characteristics, and is limited to machines that can be loaded for a z.p.f. test. In *Figure 28.21*(*a*), A is the rated-current short-circuit point and D is the rated-current rated-voltage point: to reach D the field current exceeds the rated-load excitation level. Draw DC = A0, and draw CF parallel to the airgap line 0BH. Drop FL perpendicular to DC, and draw triangle 0BAM similar to CFDL. Then FL is the Potier voltage drop IX_p, and DL is the armature-reaction m.m.f.

The argument is that for a given stator current the armature-reaction and leakage-reactance voltage drops are constant, but the latter requires more excitation at the higher levels of saturation.

0G, 0A and FH in *Figure 28.20*(*b*) are excitation currents as in (*a*), with 0G for rated voltage on the *airgap* line. GH is the total excitation required.

The z.p.f. test may have to be performed at less than rated stator current; the z.p.f.c. is then closer to the o.c.c., and DL and FL are smaller. Nevertheless, the Potier reactance is still considered to be $X_p = FL/I$ up to rated value.

Given only the o.c.c. and s.c.c., and no facility for adequately loading the machine, one procedure is to measure the d axis subtransient reactance from a sudden-short-circuit test (or by the IEEE method below) and to use it as X_p.

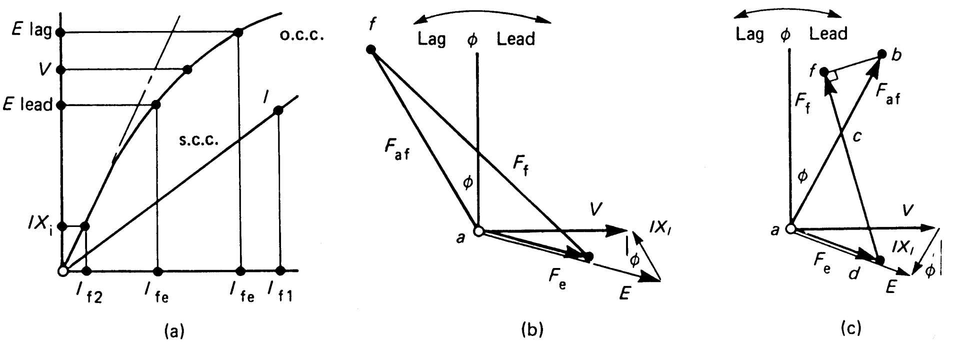


Figure 28.20 M.m.f excitation diagrams: (a) open- and short-circuit characteristics; (b) cylindrical-rotor, power factor lagging; (c) salient pole, power factor leading

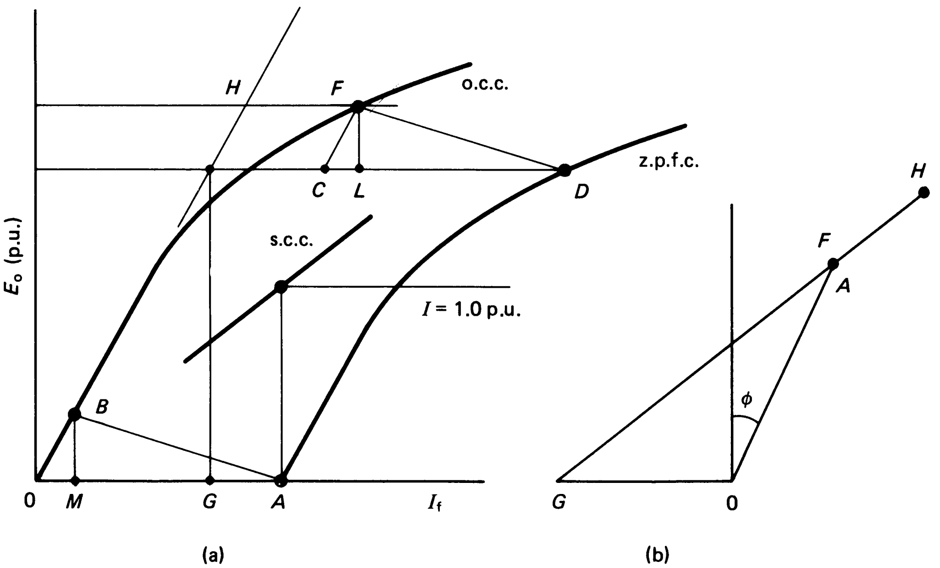


Figure 28.21 Potier reactance excitation diagram

The test method for X''_d⇐in ANSI/IEEE 115, Clause 7.30.25, is to apply a voltage E of normal frequency to each pair of stator terminals in turn and to observe the current I with the rotor stationary. Let the three quotients of E/I be A, B and C. Then with E and I in per-unit values of rated phase voltage and current, $X''_d = (A + B + C)/6$ to an approximation. To avoid rotor overheating, the duration of the test should not exceed the maker's recommendations (e.g. 0.2 p.u. for a time sufficient to read the meters).

28.12.3 Use of design calculation (*Figure 28.22*)

Methods can be refined by allowing for the rotor pole-to-pole leakage. From a knowledge of the airgap flux, the m.m.f.s required for the gap, stator teeth and stator core are calculated. The m.m.f. for the rotor pole and body are calculated from the gap flux and pole leakage. The component phasors are added as in the diagram, where the total rotor m.m.f. per pole bc is obtained from 0a (armature-reaction), 0b (gap, teeth and core) and ac (pole and body). Saliency can be allowed for by dividing 0a at d such that $0d/0a = X_q/X_d$. Then the total rotor m.m.f. is bc′. This differs very little from bc, but the load angle δ'⇐is more accurate.

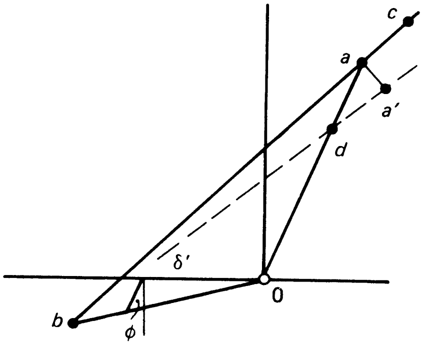


Figure 28.22 M.m.f. excitation diagram from design calculations

28.13 Sudden three-phase short circuit[6,7,11,19]

If a three-phase generator is initially excited to a phase e.m.f. E_o on open circuit, and then the three phases are suddenly short circuited together, the stator winding carries balanced three-phase currents of up to several times full-load value depending on the magnitude of E_o. These currents produce a m.m.f. that rotates synchronously with the rotor, with its axis along the main pole axis, tending to reduce the mutual flux from its initial value. Change of flux linkage in a closed circuit induces therein a current opposing the change. Hence large direct currents are induced in the rotor damper circuits and field winding.

The combined effect of the large and opposing stator and rotor currents is to produce large leakage fluxes around the stator winding, the damper circuits and the field winding while the mutual flux along the main flux axis decreases correspondingly, so that the *total* flux linkage with each winding remains momentarily unchanged. I^2R losses rapidly dissipate stored magnetic energy and the damper currents rapidly decay. Typically, the induced field current reaches its peak a period or two after the short-circuit instant and then decays relatively slowly (*Figure 28.23*).

In the stator, each phase current is asymmetric to an extent depending on how near the phase voltage was to zero at the instant of short circuit; zero instantaneous voltage produces full asymmetry. Thus each phase carries a d.c. component which, at the instant of short circuit, is equal and opposite to the instantaneous a.c. component. These d.c. components produce a stationary m.m.f. sufficient to hold the stator flux linkage momentarily unchanged, i.e. fixed relative to the stator in the position the stator flux linkage occupied at the short-circuit instant. The d.c. rapidly decays, and with it the stationary field; while it persists, however, the rotation of the rotor within it induces rotational-frequency currents in the rotor damper and field circuits. The a.c. in the field is clearly seen in *Figure 28.23*; it is the greater, the less effective the damper circuits are. With no damper at all, the induced d.c. and the zero-to-peak amplitude of the a.c. would initially be equal.

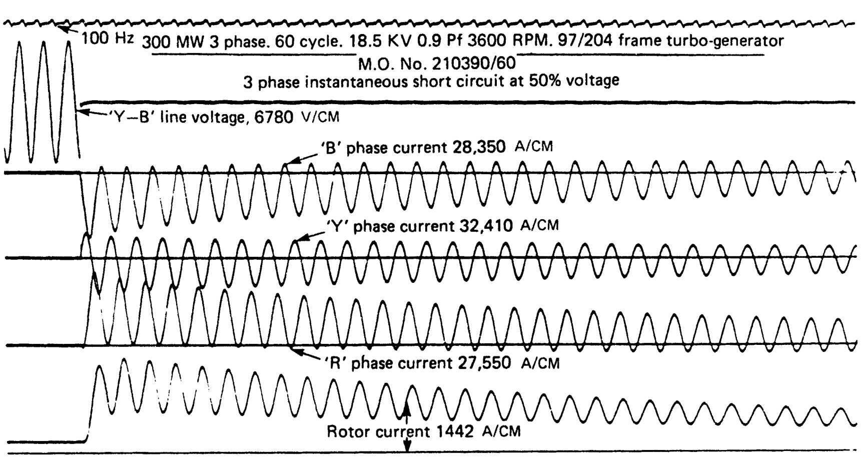


Figure 28.23 Short-circuit oscillogram

The path taken in the rotor by the stationary airgap flux has a greater permeance when the d axis coincides with the stationary flux axis than when the q axis coincides. Hence the flux fluctuates at twice fundamental frequency, and double-frequency current is induced in the stator winding. This current and the a.c. in the rotor decay as the stator d.c. component decays. The magnitude of the second-harmonic stator current depends on the difference between X''_d and X''_q; it is small in turbogenerators and in salient-pole machines with good interconnected damper windings.

In summary, the a.c. components in the stator give rise to the d.c. components in the rotor circuits; after the subtransient period, stator a.c. and field d.c. decay together with the transient short-circuit time constant, T'_d. The stator d.c. components produce the rotor a.c., and these decay together with the armature short-circuit time constant T_a.

Direct-axis reactances and time constants are derived from the oscillograms of a three-phase short circuit as follows (see *Figure 28.24* and Section 28.22: BS EN 60034-4:1995 and ANSI/IEEE Std 115-1995). The oscillograms should record for not less than 0.5 s; I_d can be measured by instruments or by taking a second oscillogram after the steady state has been reached. Speed and field current should be constant throughout. E_o *is* the open circuit phase e.m.f. corresponding to the rotor excitation.

The modern testing technique includes also digitally recording voltages and currents, and using computer programs to analyse the results and to present values of reactances and time constants. The principles are the same as for the analysis of the oscillograms described below.

(1) Draw the envelopes abc and a′b′c′ of one phase-current oscillogram. Then aa′ is the double-amplitude of the prospective current at the instant $t = 0$ of short circuit. (The first current peak is slightly less than ½ aa′ because of the rapid subtransient decrement.) Taking aa′ *as scaled in per-unit terms*, the r.m.s. current is $I''_d = \mathrm{aa}'/(2\sqrt{2})$ and

$$X''_d = E_o/I''_d$$

(2) Project the envelopes in the transient region, cb and c′b′ back respectively to d and d′, ignoring the initial rapid subtransient decrement (this cannot be done with great accuracy). Then $I'_d = \mathrm{dd}'/(2\sqrt{2})$ and $X'_d = E_o/I'_d$.

(3) Repeat steps (l) and (2) for the other two phases and derive the mean values of X''_d and X'_d.

(4) For a closer estimate, the equation relating the r.m.s. value of the a.c. short-circuit current I_t to time may be used:

$$I_t = I_d + (I'_d - I_d)\exp(-t/T'_d) + (I''_d - I'_d)\exp(-t/T''_d) \qquad (28.46)$$

where $I_d = \mathrm{ee}'/(2\sqrt{2})$ is the sustained steady-state short-circuit current, and I'_d and I''_d are the transient and subtransient r.m s. currents respectively, corresponding to dd′ and aa′, respectively, at $t = 0$.

(5) Measure (e.g. in centimetres) the double-amplitude between the envelopes at each current peak and subtract ee′ from each. Plot the results as ordinates on a logarithmic scale, to a linear base of time. This gives the curve abc in *Figure 28.24*(*b*).

(6) Project cb by a straight line to d at $t = 0$. Then

$$[\mathrm{d\ (cm)}/(2\sqrt{2})] \times \text{current scale of oscillogram} + I_d = I'_d$$

whence

$$X'_d = E_d/I'_d$$

[If bc is not linear, the decay is not exponential and I'_d is not a constant. I'_d and T'_d can be estimated from a straight line drawn through chosen points on the curve bc where the transient current components are b and 0.368b (see IEC 34–4: 1985).]

(7) The subtransient r.m.s. current at $t = 0$ is $I''_d = I'_d +$ the rapidly decaying component represented by da. Thus

$$I''_d = I'_d + \frac{\text{Intercept da}}{2\sqrt{2}}\ (\mathrm{cm}) \times \text{Current scale}$$

and

$$X''_d = E_o/I''_d$$

The intercepts between ab and db are drawn to extended current and time scales in the lower part of *Figure 28.24*(*b*). Point f on the ordinate scale corresponds to da.

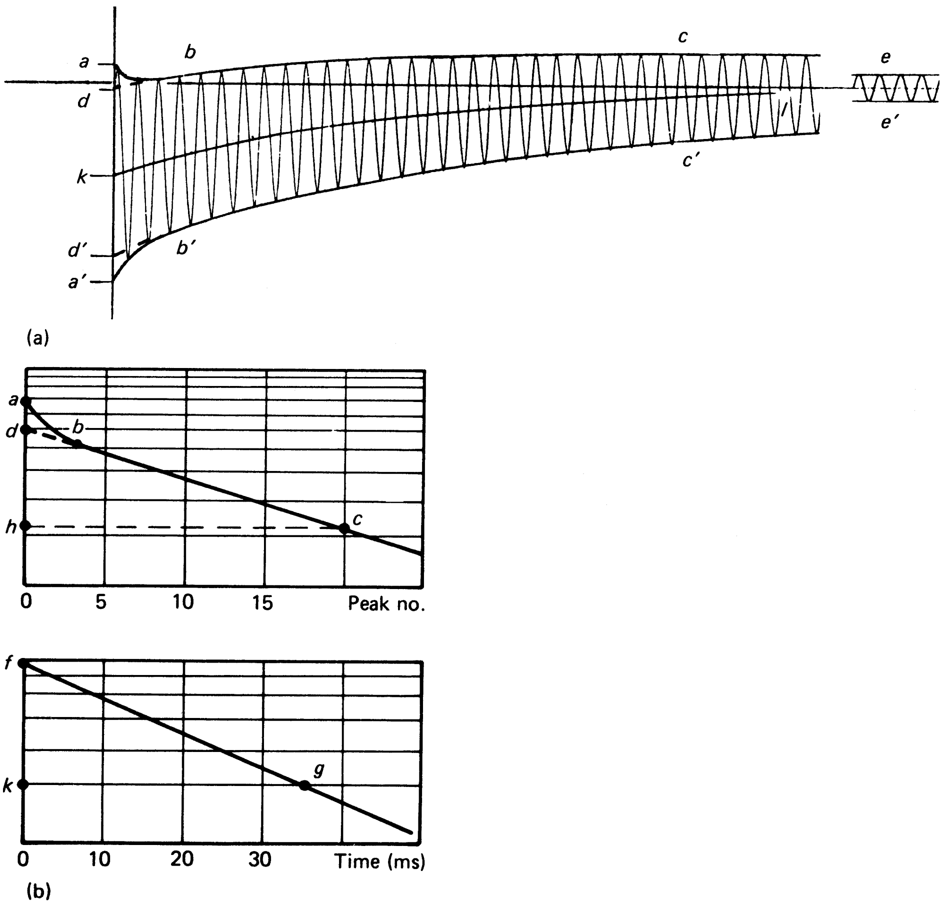


Figure 28.24 Analysis of a short-circuit current oscillogram: (a) current envelope; (b) logarithmic plot

(8) Time constants are obtained from the slopes of the current–time plots. At c on line dc the transient component at time *t* is represented by *h* on the logarithmic ordinate (centimetre) scale and by hc on the time scale. It is related to I'_d at $t = 0$ by

$$I_t - I_d = (I'_d - I_d)\exp(-t/T'_d) \tag{28.47}$$

i.e. $h = d\exp(-t/T'_d)$, consequently

$$T'_d = (hc)/\ln(d/h) \tag{28.48}$$

where *hc* is in seconds and *d* and *h* are in centimetres. T''_d is obtained similarly from the extended-scale plot in *Figure 28.24*(*b*). The $T''_d = kg/(\ln f - \ln k)$.

(9) The procedure (5)–(8) is repeated for the other two phases and the mean values are obtained.

(10) The reactance values decrease with increasing short-circuit current (and therefore with increasing open-circuit voltage E_o). Rated-current values are obtained when at $t = 0$ the transient current I'_d is equal to rated current. A range of short-circuit tests spanning the expected X'_d will give a plot of reactance to a base of transient current I'_d from which the appropriate value can be found. A rated-voltage value, if required, is obtained by testing at 1 p.u. voltage for a small machine without a transformer, but the electromechanical forces on the stator endwindings would be excessive in a large generator. Tests up to 0.7 p.u. voltage simulate a fault on the h.v. side of the transformer in a generator–transformer unit and are more relevant to the service conditions of a large machine.

(11) The armature short-circuit time-constant T_a is determined from the decay of the d.c. components of stator current or from the decay of the a.c. component of induced field current. The latter method is simple: it requires only a log-linear plot of the a.c. component to a base of time, similar to the plot in *Figure 28.24*(*b*). The stator d.c. component is represented by the median line kl of the current envelopes in *Figure 28.24*(*a*). However, if there are significant even-order harmonics then the median is displaced and a waveform analysis is necessary to find the harmonic effect. This will occur in a salient pole machine that has no effective damper winding, and so is not common.

28.14 Excitation systems

A synchronous machine requires an excitation system to provide the field current for magnetising the machine to the desired voltage and, when it is running in parallel with others, determining the lagging reactive power generated or received. It is customary for each generator to have its own self-contained excitation system, which provides the power required to supply the I^2R loss in the field circuit. This varies between about 10 kW/MVA for small machines and 5 kW/MVA for very large units.

Excitation voltage and currents are chosen: (i) to give field winding conductors that are mechanically robust in small machines and not too massive in large ones, (ii) to suit the ratings of available diodes or thyristors, and (iii) to give convenient designs of exciter, and also of slip-rings where these are used. Values range from a few score amperes and volts on very small machines up to say 8 kA at 600 V on the largest turbogenerators. At no load or with leading power factor, control of the exciting current is needed down to about one-third of the value for rated load.

The excitation system must respond to applied signals quickly enough to have the desired effect on the generator flux. Its duties can be broadly classified as:

(1) to control the generator voltage accurately as slow changes of power and reactive loading occur;
(2) to limit the fluctuations of voltage when loads are suddenly imposed or removed;
(3) to maintain steady-state stability; and
(4) to maintain transient stability.

(See references 168, 172 and 173.) These duties require different characteristics of the excitation system: these must be reconciled to provide proper control.

The performance of an excitation system is represented by its response ratio, or its response time and ceiling voltage. BS EN 60034-1:1998 gives definitions, and ANSI/IEEE Standard 421.1 gives definitions and methods of test for these and for other characteristics. In *Figure 28.25*, abdm is the voltage–time curve obtained, starting from the voltage V_e needed on the generator field winding at rated load, when the control is suddenly changed to cause ceiling voltage V_c to be reached as quickly as possible. Definitions are, where area abdf equals area agf:

Characteristic	*ANSI/ IEEE 421.1*	*BS EN 60034-1:1998*
Response time (s)	t_2	—
Response ratio measured over 0.5 s	$\frac{\text{2gf in volts}}{V_e}$	$\frac{\text{2gf in volts}}{V_e}$
Initial response/ second	—	$\frac{1}{V_e}\left(\begin{matrix}\text{Slope of } V-t \text{ curve}\\ \text{at } t=0, V/\text{s}\end{matrix}\right)^*$
High initial response system	A curve such as ak	—

*The initial response is $(V_c - V_e)/(t_1 V_e)$ per second if curve abd is exponential.

The measurements are conveniently made with the exciter on open circuit, but for analysing the generator behaviour values are needed with the exciter supplying the rotor winding. Where such testing is impractical, the on-load values must be calculated using known parameters of the machine(s).

Any system with a response time of 0.1 s or less is called a 'high initial response system'. A thyristor system[166,171] supplied from a transformer (or from an exciter machine that runs continuously at ceiling voltage) is inherently a high initial response system. A brushless exciter system[165,167] can be given high initial response by forcing the exciter field current with a pilot exciter voltage that may be 5–10 times that needed for rated generator output. The exciter field must be designed to have a short time constant; the whole magnetic circuit must be laminated, and damper windings, both deliberate and incidental, must be avoided. For example, clamping bolts and plates should not form closed loops linking flux. The exciter output voltage can be forced to a ceiling value of, say, 2–3 times V_e in less than 0.1 s. The output current rises more slowly, depending on the effective time constant of the generator field circuit. The controller may limit the exciter field current to say 2 ½ times the rated load value. In service such forcing would usually occur only for a few periods of up to about 0.5 s each as the rotor swings and returns to synchronism after a system fault.

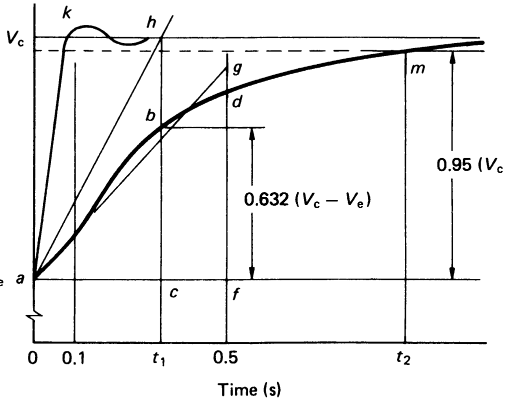


Figure 28.25 Excitation response definitions

A heavily forced main exciter with V_c about $2V_e$ may provide the required *initial* response more cheaply than would a bigger exciter with yet higher V_c but a longer time constant.

Whatever the performance of the excitation system, the change in generator flux is delayed by eddy currents induced in its field winding and any available damper circuits. It is not practicable to constrain the design to avoid these: for mechanical reasons a turbogenerator needs a solid forged rotor (except very small ones), whilst a hydrogenerator with laminated poles needs pole-face dampers. Hence the natural generator time constants have to be accepted, and the excitation system designed to suit them. The extra cost of excitation systems to achieve more and more rapid response may become unjustifiably great in relation to the control actually achieved on the generator.

For the duties (1) to (4) listed above, the characteristics of the excitation system need to be as follows.

(1) To hold the generator voltage within specified limits, which may be between ±3%, and ±0.5% of the set voltage, the excitation system needs a high d.c. gain, but a moderate ceiling voltage is enough to supply the small and slow changes of excitation needed.
(2) To limit the fluctuation of voltage when load is suddenly applied or removed, a large, rapid and well damped response is needed. However a high initial response system, as defined above is often not necessary. For example, starting a large motor demands first a large reactive output from the generator, then increasing real power as the motor runs up. On a small generator or generator group, to avoid the voltage dip being too great and too prolonged, the generator needs a low X'_d, and the excitation must be kept high

for perhaps several seconds. A high ceiling voltage and a response time of say 0.2 s is more useful than a high initial response system with a lower ceiling.

(3) Following a fault and its clearance on a high voltage power system, or a serious sudden loss of generation, a large and rapid response is needed to maintain transient stability. i.e. to restore the voltage and synchronising power flow to hold the several generators in step. Whether or not a high initial response system is essential depends on the particular circumstances, including the inertias of the generator sets and the post-fault reactances of the system. Very low frequency swinging, at 0.5 Hz or less, can occur between generating areas, requiring higher than rated excitation currents to be maintained for several seconds.

(4) Steady-state stability can be improved, in the sense that the generator can run safely close to, or even beyond, the fixed excitation stability limit, if the controller has no dead band, and is designed to have the desired speed of response without introducing a phase shift that causes positive feedback. Such feedback encourages oscillations of rotor angle, etc., which can become intolerable. This is more likely with high-reactance power lines. Long lines with series capacitance inserted to compensate for the line inductance introduce the possible hazard of subsynchronous resonance.[128–133] A minor disturbance, for example normal switching of lines, can cause transient current to flow at the line natural frequency f_n (hertz), often in the range 20–40 Hz. Torque is developed on a generator rotor at system frequency f_s–f_n (hertz). If this is close to a natural frequency of torsional oscillation of some part of a turbogenerator shaft system, the oscillation may increase, sustaining the subsynchronous current. Fatigue damage has occurred on a few turbogenerator sets in this way.

Where stability problems are judged likely, the excitation controller is supplemented by a power system stabiliser. This acts in response to input signals such as voltage, power, rotor angle, or derivatives of these. They cause the controller to adjust the excitation so that torque is developed in the correct sense to damp the oscillations.

Because of its inherently fast response, and because of the mechanical advantages noted in Section 28.17 the self-excited thyristor excitation system is almost always used for hydrogenerators on long lines.

28.14.1 D.c. exciters

These have been superseded by brushless or static thyristor a.c. systems. The exciter was a d.c. generator coupled to the shaft of the synchronous machine, feeding its output to the main field through slip-rings. For high-speed generators of more than about 50 MW, the exciter had to be driven at a lower speed (typically 1000 or 750 rev/min) through gears, or separately by a motor, in order to avoid difficulties of construction and of commutation. On very-low-speed hydrogenerators a directly coupled exciter would be excessively large, so a higher-speed exciter, driven by a motor or perhaps a small water turbine, was used.

For small ratings, the exciter was shunt excited; however, most were separately excited from a directly coupled shunt-excited pilot exciter. Control of the generator excitation was provided by controlling the field current of the main exciter.

28.14.2 A.c. exciters with static rectifiers

Satisfactory service experience with brushless and static thyristor systems has made these early (diode) systems obsolescent. Many remain in service, but many have been replaced by modern equipment. The advent of solid-state rectifiers made it possible to avoid commutators by using an a.c. exciter, directly coupled to the generator and feeding its output via floor-mounted rectifiers to the generator held winding through slip-rings. The exciter can operate at any economically convenient frequency, usually between 50 and 250 Hz, and the system is suitable for generators up to the largest ratings.

The diode cubicles may be cooled by natural convection or forced air flow. Alternatively, especially for large ratings. the diodes may be mounted on water-cooled bus-bars; this greatly reduces the size of the cubicles so that they can, if desired, be mounted on the sides of the main exciter frame, thus avoiding the need for long runs of a.c. and d.c. bus-bars or cables.

The main exciter field is supplied by an a.c. pilot exciter, often a permanent-magnet generator. The excitation of the main generator is controlled by controlling the main exciter field current via the automatic voltage regulator.

The main exciter is usually three-phase, and the diodes are connected in the six-arm bridge circuit, usually with a fuse in series with each diode to interrupt the fault current should a diode break down (which almost always causes it to conduct in both directions, i.e. to act as a short circuit). Diodes are available with current ratings up to 1000 A mean d.c., and peak inverse voltage up to 5 kV (but not both together in one diode). For other than small ratings, each bridge arm has several diodes in parallel; all the diodes are fused and have sufficient current margin to enable the bridge to carry full-load excitation continuously with one or more of the diodes failed and isolated by their fuses. Hence the generator can remain in service until maintenance can be carried out conveniently. Some installations on large generators had a.c. and d.c. isolators to allow parts of the bridge to be worked on without taking the generator out of service.

The diodes must be able to withstand induced transient currents and voltages resulting from system short circuits, asynchronous running, pole-slip and faulty synchronising, as well as from faults in the excitation system itself. Their continuous duty rating must leave some margin for imperfect sharing between parallel paths and for the possible loss of one or more paths, as noted above.

The fuse characteristic is co-ordinated with that of the diode so that fuses should not blow unless a diode fails or there is a short circuit on the d.c. output. The fuses must clear the fault current under the most onerous condition, which is usually that of a failure with the exciter at ceiling voltage. Fuse blowing is easily indicated by a microswitch operated by a striker pin that is ejected from an indicator fuse in parallel with the main fuse.

28.14.3 Brushless excitation

The a.c. exciter has a rotating armature with three or more phases and a stationary field system. It usually is designed for a frequency of 0–5 times the power system frequency. The pilot exciter, when there is one, is usually a permanent magnet generator operating at around 6–8 times system frequency. The diodes and fuses are mounted on the rotor, and the rectified output is led directly to the generator field winding without need of brushes and slip-rings. The diodes are mounted on well-ventilated heat sinks, and special designs of fuse are used to withstand the centrifugal force on the fusible link.

On units small enough to require only one diode and fuse per arm, failure of one diode or fuse leaves the exciter with one phase unloaded; exciters are usually designed to supply full-load excitation in this condition without damage, so that the generator can remain in service until the fault can be repaired conveniently. However, experience shows that the failure rate of diodes is extremely low and that more often fuse links fail mechanically. Hence some makers supply salient-pole generators, up to say 25 MW, with no fuses at all, but use generously rated diodes to provide a large margin. These generators would use up to three diodes in parallel per bridge arm. For turbogenerators up to about 70 MW, some designs use two diodes in series—each of full duty, with one, two or more series pairs in parallel per bridge arm—and no fuses. On large units, redundant parallel paths, individually fused, are provided as in static equipments.

For units that use fuses, the striker-pin indicator type can still be used, the pin being observed by causing it to interrupt a light beam falling on to a photoelectric cell, or it can be observed visually with a stroboscope. Alternatively a neon lamp is connected across the fuse and glows when the fuse blows.

When diodes are in parallel, whether fused or not, if one becomes open circuit the system will continue to function apparently normally unless the remaining diodes are overloaded and eventually fail also. If an unfused diode fails by short circuiting, the short-circuit current in the exciter armature induces fundamental (exciter) frequency current in the exciter field winding. This can be detected and used to trip the set before serious damage is done. Another method of detection is to use stationary pick-up coils to see whether the diode connections are carrying current as they should as they pass the coils.

More elaborate indication, perhaps coupled with measurements of current and voltage and indication of earth fault, can be arranged by telemetry, but the telemetry may be less reliable than the diodes. Frequently instrument slip-rings are used, with solenoid-operated brushes that make contact only when readings are required.

The diodes, and fuses too if they are used, must be rated for the normal duty, including field forcing, and to withstand the abnormal conditions noted in Section 28.14.2.

28.14.4 Thyristor excitation

Direct control of the field current of the synchronous machine by thyristors gives quicker response than can be obtained by controlling the exciter field current, because the time delay in the exciter is eliminated and the machine field current can be forced down by using the thyristors to reverse the machine field voltage. (By contrast, with a diode bridge, the machine field voltage can only be reduced to zero by reversing the exciter field voltage.) This is valuable for generators and synchronous compensators in certain power-system situations: for example, to minimise the voltage dip caused by large and possibly frequent load changes; to maintain transient stability of a generator under short-circuit conditions on the power system; to enable a synchronous compensator to maintain close control of the system voltage by rapid change in its reactive load, to minimise the voltage rise following sudden load rejection; to reduce more quickly the current resulting from a fault between the generator and its nearest protective circuit-breaker when field suppression is the only means available.

The synchronous machine requires slip-rings and brushes, and this is a disadvantage, especially for large machines for which brushgear maintenance may become a significant inconvenience.

The excitation power may be supplied by direct coupled main and pilot exciters, the main exciter working continuously at ceiling voltage. This makes the power supply independent of voltage fluctuations on the power system. Usually though the excitation is supplied from the generator terminals through a step down transformer. This is usually designed to provide the required ceiling voltage when its primary voltage is reduced to about 60% of normal. This ensures that some field forcing can be done even when the power system voltage is depressed by a fault. It does subject the generator field winding to a rather high peak voltage when the system voltage is normal. A lower ceiling is adequate if power-rated current transformers are added in order to derive some excitation from the machine output current, so boosting excitation during the fault. The set is shortened by the absence of the exciter, and this may save costs on foundations and building. For very-low-speed generators the scheme may well be cheaper than a direct-coupled exciter and diodes.

Some excitation systems use diodes and thyristors in combination, e.g. in a full-wave half-controlled bridge circuit. One patented scheme uses a full-wave diode bridge with thyristor ‘trimmer’ control fed from a special excitation winding on the generator stator and from compounding current transformers.

Rotating thyristor systems have not yet been developed commercially, mainly because of technical difficulties in transferring control signals from the stationary equipment and problems concerning the reliability of rotating control circuitry.

28.14.5 Excitation systems circuits

Typical systems are shown in *Figure 28.26*.

(*a*) *Self-excitation* provides a simple and inexpensive scheme for generators up to about 3 MVA, using a one-phase thyristor output stage. With a three-phase thyristor bridge the scheme is applied for the highest ratings. The bridge rectifier may be half-controlled, with thyristors and diodes in combination. Another variant has a diode bridge that provides more exciting current than is demanded and thyristors to divert part of this current from the field winding.

(*b*) *Self-excitation through an exciter* is convenient for brushless sets where the diodes are mounted on the generator-exciter shaft, and where the cost or mechanical complication of a pilot exciter is undesirable. A typical rating limit is 10 MVA.

(*c*) *Separate excitation* provides excitation power independent of the generator output. It is commonly used for generators rated at 10 MVA up to the maximum.

Scheme (*a*) is capable of the most rapid response. In (*b*) and (*c*) some delay is introduced by the exciter time-constant; consequently a high exciter ceiling voltage and a large output from the pilot exciter are needed to obtain a more rapid response.

28.14.6 Excitation control

When a generator operates alone the excitation is controlled to maintain the steady-state voltage within the necessary limits, and to prevent unacceptable variations of voltage when large and sudden changes of load occur. Generators running in parallel may need additional control signals to

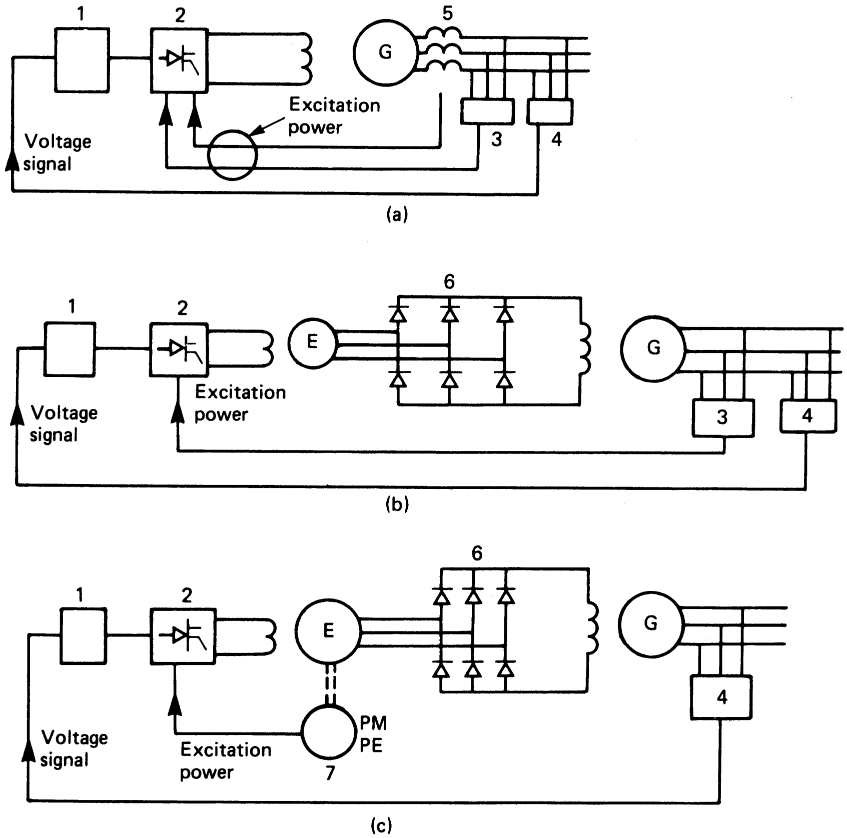

1
2
5
G
Excitation power
Voltage signal
3
4
(a)
6
1
2
E
G
Excitation power
Voltage signal
3
4
(b)
6
1
2
E
G
PM PE
7
Voltage signal
Excitation power
4
(c)

Figure 28.26 Excitation systems: (a) direct self-excitation; (b) self-excitation through an exciter; (c) separate excitation. 1, Control circuits; 2, power-output stage (a.v.r.); 3, excitation power transformer; 4, voltage transformer; 5, current transformers for excitation power; 6, diode rectifier (static or rotating); 7, permanent-magnet pilot exciter

share the total reactive load correctly between them. In an interconnected system, control of steady state and transient stability is a vital duty. Manual control of the excitation is inadequate, and automatic control is provided.

Electromechanical voltage regulators, in use only in old installations, may be of the carbon-pile, vibratory-contact (Tirrill) or rolling-sector (Brown Boveri) type. These have been superseded, initially by magnetic amplifiers, with or without amplidynes, and now by solid-state control systems using transistor amplifiers at low power levels, with thyristors for field-circuit power. The new systems are continuously acting (i.e. have no dead band) and can be arranged to respond to many control signals besides that from the terminal voltage, so the term 'automatic excitation controller' is more logical than 'automatic voltage regulator' (a.v.r.). Power supply for the control circuits is derived from the machine terminals or the pilot exciter.

28.14.7 Basic principles of voltage control

A direct voltage proportional to the generator average terminal voltage is derived via voltage transformers and a diode rectifier circuit. This voltage is compared with a stable reference voltage generated within the regulator. Any difference (the 'error voltage') is amplified and used to control the firing of a thyristor circuit which supplies the excitation, either to the field of the synchronous machine or to its main exciter-field winding. Thus the excitation is raised or lowered to restore the machine voltage to the desired level and the error voltage returns to near zero. The set level is obtained by adjusting the proportion of the machine voltage that is compared with the reference voltage or by adjusting the reference voltage itself. The basic circuit (*Figure 28.27*) is incorporated in *Figures 28.28* and *28.29*.

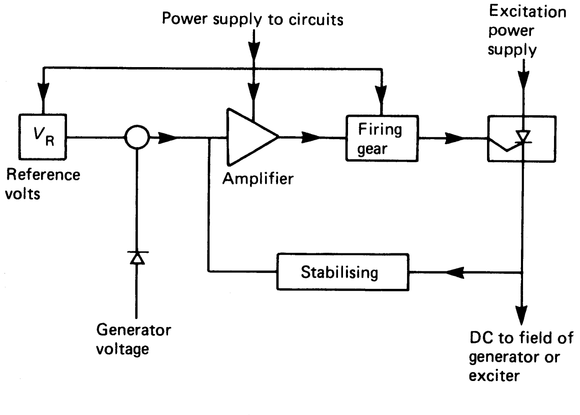

Power supply to circuits
Excitation power supply
V_R
Reference volts
Amplifier
Firing gear
Stabilising
Generator voltage
DC to field of generator or exciter

Figure 28.27 Basic circuit of an a.v.r.

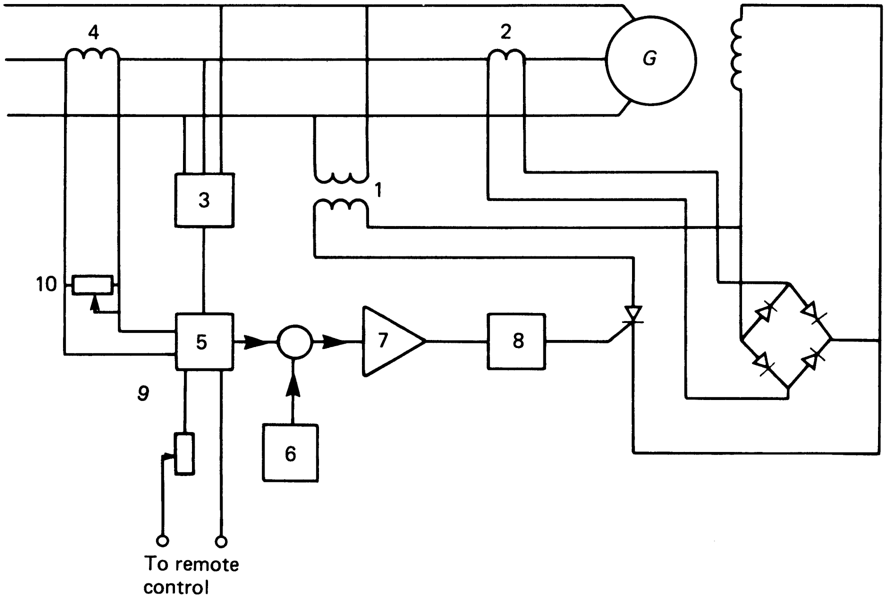

4
2
G
3
1
10
5
7
8
9
6
To remote control

Figure 28.28 Self-excitation with a one-phase thyristor; 1, Excitation transformer; 2, excitation current transformer; 3, voltage transformer; 4, compounding current transformer; 5, voltage-measuring circuit; 6, reference voltage; 7, amplifier; 8, firing gear; 9, voltage-setting rheostat; 10, compounding adjustment

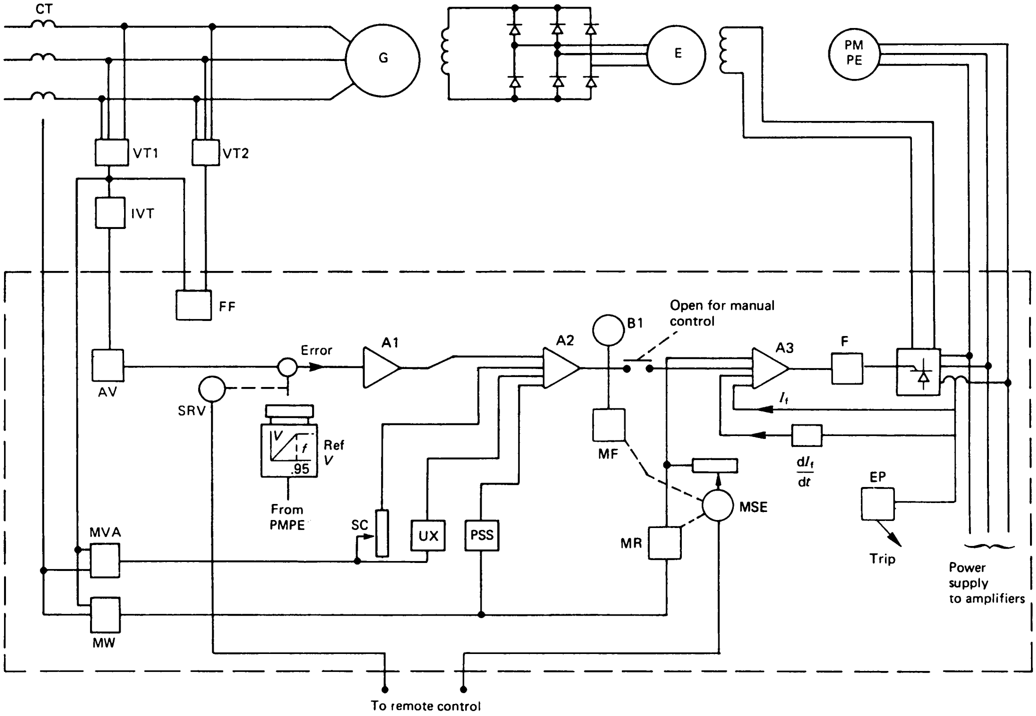

CT
G
E
PM
PE
VT1
VT2
IVT
FF
Open for manual control
B1
AV
Error
A1
A2
A3
F
SRV
Ref V
.95
From PMPE
I_f
MF
dI_f/dt
EP
MSE
Trip
MVA
SC
UX
PSS
MR
Power supply to amplifiers
MW
To remote control

Figure 28.29 Automatic voltage regulator with additional control features: A1, A2, A3, amplifiers; AV, average-voltage circuit; B1, balance indicator; CT, current transformers; E, exciter; EP, excitation protection circuit (operates on high or low excitation current); F, thyristor firing gear; FF, voltage-transformer fuse failure detector and alarm; G, generator; IVT, isolating voltage transformer; MF, manual follow-up circuit (adjusts MSE); MR, manual restrictive circuit; MSE, manual set-excitation control; MVA, MW, circuits providing signals proportional to MVA and MW; PMPE, permanent-magnet pilot exciter; PSS, power-system stabiliser control; Ref V, reference voltage; SC, set-current compounding control; SRV, set-reference voltage control; UX, underexcitation (var limit) control; V_{T1}, V_{T2}, voltage transformers

28.14.7.1 Control range

Generators are usually designed to deliver any load from zero to rated output over a voltage range of ±5%, at any power factor between rated (usually 0.8–0.9 lag) and say 0.95–0.9 lead. The a.v.r. setting controls must provide the corresponding range of excitation, and also provide say 85% of rated voltage on no load. Accuracy of control is usually within ±2.5 or ±1.0% of the set value over the load range.

28.14.7.2 Manual control

Manual control is usually provided for use if the automatic control fails or for convenience when the set is being commissioned. In small units the manually controlled system may be entirely independent of the automatic one, but (especially for economy on large units) it uses the thyristor output stage and the associated firing circuits of the regulator.

Some regulators, when in auto control, drive the manual control so that, if it were in use, it would give the same excitation as the auto circuit. However, this follow-up must be prevented from driving the manual control down to excitation levels that are stable with continuously acting control but unstable with fixed excitation.

Some systems use the manual circuits continuously to control the steady-state excitation. The auto circuits continuously trim this to suit minor fluctuations of load, voltage, etc.; larger disturbances will cause rapid automatic changes of excitation voltage, up to full boost or full buck if necessary. If changed conditions persist for more than a few seconds, the follow-up circuit adjusts the manual control to the new steady state and the auto-circuit output falls to its usual low level.

28.14.7.3 Manual-to-auto change-over

Whichever system is in control, it is necessary to adjust the other automatically or manually, so that a change-over can be made without causing a significant change in excitation. A balance meter is provided so that the outputs of the two systems can be matched before making the change. The manual rheostat and the voltage-setting rheostat are often motorised for control from a remote control room, and the balance-meter reading must be repeated there too, unless automatic matching is provided.

28.14.8 Additional control features

28.14.8.1 Parallel operation

To ensure satisfactory sharing of reactive load between generators paralleled at their terminals, the a.v.r. can arrange for the terminal voltage to fall with increasing reactive load, usually by 2.5–4.0% at full load. For generators paralleled on the h.v. side of step-up transformers, the a.v.r. can either add to or partly compensate for the transformer impedance drop, as desired.

28.14.8.2 Excitation limits

Fault conditions on the power system will cause the excitation to rise to ceiling value to try to maintain normal voltage. An adjustable timer is used to return to normal excitation after several seconds in order to avoid overheating the machines if the fault persists.

A reactive-power-limiting circuit can be used to prevent the excitation falling so low that the generator will not remain in step with the system. The reactive power (under-excited) at which this circuit operates is automatically varied in response to machine voltage and power output to maintain an adequate stability margin.

28.14.8.3 Overfluxing protection

It is operationally desirable to be able to leave the a.v.r. in control when the generator is shut down or run up. To avoid overfluxing the machine and its associated transformer (if any) the reference voltage is arranged to decrease in proportion to frequency at speeds below about 95% of normal. This ‘constant volts-per-cycle’ control is needed also for generators that have to operate over a speed range, e.g. for ship propulsion.

28.14.8.4 A.v.r. fault protection

Failure of a.v.r. components, or of other components in the excitation system, may cause excessive or insufficient excitation for safe operation. Either condition trips the a.v r. to manual control and alerts the operator.

Voltage-transformer fuse failure is detected by comparing the voltages from two voltage transformers. If the a.v.r. voltage transformer fails, the system trips to manual control; if the comparison circuit fails, an alarm shows that fuse failure protection is no longer working.

28.14.8.5 Double-channel a.v.r.

To enhance reliability of operation, large or vital generators frequently use regulators in which the automatic and manual control circuits are duplicated; often the thyristor output stage supplying the exciter field is duplicated too. Occasionally the much larger thyristor bridge that feeds the generator field directly may be duplicated. Each channel is able to perform the full excitation duty; the two channels may operate in parallel or in main and stand-by mode. If one channel fails, the other maintains the excitation unchanged. If the second channel fails subsequently, it trips to manual control. Alarms indicate the abnormal conditions.

28.14.9 Overall voltage response

The overall voltage response is defined in terms of steady-state and transient behaviour with the generator on open circuit and on no load, and under the control of the excitation system. For a generator operating *alone*, the following conditions may be relevant, their importance depending on the duty required of the generator:

(1) *Steady state*: accuracy of voltage control over the range of load and power factor.
(2) *Transient*: 1. Response of the open-circuit generator voltage to a step change in reference voltage; 2. Voltage response when a sudden increase or decrease of load occurs.

Conditions of importance under transient conditions are the voltage rise and recovery time of the generator voltage when load is suddenly removed, and the voltage dip and recovery time when a large motor is switched on to the generator terminals. For a generator running alone, BS 4999-140:1987, ‘Specification for voltage regulation and parallel operation of a.c. synchronous generators', specifies various grades of voltage regulation, for steady state and transient conditions. Accuracy of voltage control may conform to ±1% or ±2.5% or ±5%. Voltage dip must not exceed 0.15 p.u. when a current of 0.35 or 0.6 or 1.0 p.u. of rated generator current is suddenly demanded. The voltage is required to recover to 0.94 or 0.97 p.u. within 1.5 or 1.0 or 0.5 s. Values of the temporary voltage rise that occur when rated load at p.f. 0.8 is thrown off are specified with a range 0.35–0.15 p.u. The more severe the conditions the more powerful the a.v.r. control must be. Also, the lower must be the generator X''_d and X'_d in order to reduce the immediate fall or rise in voltage that the a.v.r. cannot affect. There is a consequential increase in the short-circuit current and in the generator frame size; both these increases raise the cost of the generator and, perhaps, of its switchgear. The response under transient conditions is a convenient way of expressing the overall performance and of testing it during commissioning. The terms used are

V_1	initial voltage
V_2	final voltage
$V_1 - V_2$	voltage step
v	voltage overshoot beyond V_2
t_1	rise time, that in which V_2 is first reached (and passed)
t_2	settling time, the shortest time after which the voltage remains within say ±0.5% of a steady value (which should be V_2). *Figure 28.30* illustrates a step up of voltage.

Normally, either V_1 or V_2 is the rated voltage V_r, depending on whether a 'step-up' or a 'step-down' change is being tested. The voltage–time curve is a well-damped transient, settling at V_2 after a few oscillations. Typical values of the quantities defined above are

$V_2 - V_1$	0.1 p.u. of V_r	v	not more than $(V_2 - V_1)/2$
t_1	0.2–0.6 s	t_2	1–5 s

For a given step change $(V_2 - V_1)$, t_1 is reduced by increasing the ceiling voltage V_c, by increasing the excitation system gain, and/or by reducing the system time constants. The parameters of the generator and exciter cannot be changed once the machines are made, but the a.v.r. parameters are designed to be adjustable. Changes that reduce t_1 will increase the overshoot v and the settling time t_2; hence settings of a.v.r. gain, time constants and feedback signals (if any, e.g. exciter voltage or exciter field current) are calculated to achieve the desired compromise, and performance is checked over a range of values of step change during commissioning (say steps of 1%, 5%, 10% and 20% of V_r).

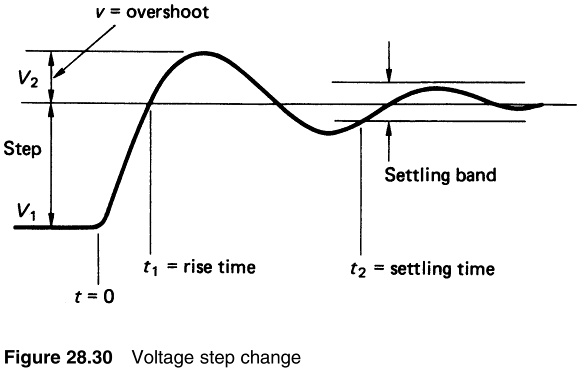


Figure 28.30 Voltage step change

When generators operate in parallel, as most do, the excitation systems must be designed and adjusted to achieve the best compromise between highly accurate voltage control, steady-state stability, and transient stability following a system disturbance. Thus some step-change tests, or tests by injecting low-frequency sinusoidal voltage into the reference circuit, are desirable, to confirm the calculated performance.

See IEEE Standard 421A, 'Guide for identification, testing and evaluation of the dynamic performance of excitation control systems'.

28.14.10 Digital control[170]

Increasing use is being made of a dedicated microprocessor to replace the analogue control system. The input signals (voltage, current, MW, MVAR, etc.) are converted to digital form, and the processor is programmed to respond to these to provide an analogue signal to the firing circuits. The characteristics needed in the excitation controller are reproduced digitally in the processor, and limits of relevant quantities are introduced. The whole can be set up and tested in the works, so that site commissioning is simpler and quicker. In service, settings do not drift or suffer from poor rheostat contacts, but can be readily changed, even while the set is in service, to suit changed operating conditions. If a component does fail, a new card can be inserted without repeating the commissioning tests.

There is a delay of up to 10 ms while the microprocessor scans the input signals and readjusts its output signal, if necessary, but this is very small compared with the machine time constants.

The thyristor firing pulses may also be generated digitally. Development is proceeding of adaptive controllers that will automatically tune their characteristics to suit the operating conditions.[169]

28.15 Turbogenerators[92–137]

The characteristic features of turbogenerators are their high speed to meet steam-turbine requirements and their large outputs to provide economy of capital and operating costs for the power station. Most are two-pole units running at 3000 or 3600 rev/min, but four-pole generators at 1500 or 1800 rev/min have become common for large outputs (1000 MW or more) from nuclear reactors of the boiling-water or pressurised-water type. These reactors deliver large volumes of steam at temperatures and pressures that are lower than those provided by fossil-fired boilers or some gas-cooled reactors. The low-speed turbine may handle these conditions with a greater efficiency that is sufficient to offset its higher capital cost compared with a high-speed unit. Design constraints are less exacting in the low-speed turbine and generator.

28.15.1 Main dimensions

The output coefficient C ranges typically from about 0.5 MVA s/m^3 for a rating of 20 MVA to about 2.0 MVA s/m^3 for a 1000 MVA unit. For 3000 rev/min machines these figures correspond to D^2L of 0.8 and 10 m^3, respectively. Economic diameters for these outputs range from approximately 0.75 m to 1.3 m, the latter being a limit set by centrifugal stresses in the endrings and in the rotor teeth. Hence outputs range from approximately 17 to about 170 MVA per metre of core length. The higher values of output coefficient are made possible by enhanced cooling techniques. Typical dimensions for a 660 MW two-pole hydrogen-cooled machine may be: rotor diameter, 1.15 m; core length, 6.8 m; overall shaft length, 13.5 m; core outside diameter, 2.7 m; outer casing, 4.8 m in diameter and 10.3 m long; total weight, 480 t.

28.15.2 Rotor body

The output available from a turbogenerator is largely determined by the excitation m.m.f. that can be carried on the rotor with acceptable winding temperatures. The high centrifugal stresses make cylindrical (i.e. non-salient-pole) construction essential.[92,93] Within the chosen diameter the number, shape, size and spacing of the winding slots have to be optimised to obtain the maximum m.m.f. capability with acceptable stresses in the teeth and slot wedges, with adequate insulation, with acceptable magnetic flux densities and with ducts for ventilation that enable temperature guarantees to be met. For air-cooled machines of medium output the manufacturing simplicity afforded by parallel-sided slots and solid copper conductors of rectangular cross-section may outweigh the loss of optimum performance and provide the cheapest design. For larger ratings tapered slots are used to accommodate more copper, while giving approximately constant mechanical stress and magnetic flux density along the radial length of the teeth.

A rotor is forged from a single steel ingot, the largest of which approach 500 t in weight; this would produce a rotor weighing 250 t, enough for a four-pole machine of about 1250 MW at 1500 rev/min. The forgings contain the alloying elements nickel, chromium, molybdenum and vanadium; according to size and speed, ultimate tensile strengths of the forgings range from 650 to 800 MN/m^2, while their 0.2% proof stresses range from 550 to 700 MN/m^2. The forgings are inspected with ultrasonics and magnetic-particle ink before use. Many generator makers now rely on these examinations and do not bore the forging axially along its centre line, except for large forgings or if ultrasonics reveals defects that can be removed by boring.

The endwindings (the parts projecting beyond the ends of the slots) must be supported against centrifugal forces by endrings (retaining rings) from which they are insulated by, for example, resin bonded fibreglass or aramid paper (Nomex) or combinations of synthetic insulating sheets. The endrings are steel forgings, usually shrunk on to the ends of the rotor body. In some older designs they were shrunk on to discs that were shrunk on to the shaft outboard of the windings, and they were not tight on the rotor body.

From the 1950s until 1982, the endrings[106,107] were of austenitic steel (18% Mn, 4–5% Cr, 0.3% C) warm worked to give high strength, up to 1100 MN/m^2 proof stress and 1220 MN/m^2 ultimate in the highest grade. This alloy is very susceptible to stress corrosion if it gets wet, e.g. by condensation from moist air or leakage of cooler water. In 1982 an austenitic alloy became available with 18% Mn, 18% Cr;[108,109] this is not subject to stress corrosion under any likely operating conditions, and has mechanical properties up to 1200 MN/m^2 proof stress (0.2% strain) and 1300 MN/m^2 ultimate strength. Endrings large enough for 1500 MVA two-pole or four-pole generators can be obtained, and higher strengths have been developed. Many endrings of the older alloy have been replaced, after some years in service, using the 18–18 material.[106,116]

Rotor vibration at running speed must be low—typically about 50 µm peak-to-peak measured on the shaft near the bearings, though up to twice this is commercially acceptable. Hence balance weights must be carefully positioned, axially as well as circumferentially, and the design of the rotor, its bearings and its supports must ensure that its critical speeds are sufficiently far from rated speed. Small 3000 rev/min rotors will have one critical speed below 3000 rev/min, say about 1700–2000 rev/min, but as ratings (and therefore the bearing span) increase, two or even three criticals will occur below 3000 rev/min (typically around 650, 1750 and 2500 rev/min). When the rotor is coupled to the turbine, the critical speeds are usually raised slightly, so that behaviour in both the coupled and the uncoupled condition must be acceptable, for site running and works testing (without the turbine) respectively.

Electrical faults on the power system or on the machine itself produce abnormally high oscillatory torques on the rotor, at system frequency and often at twice this frequency also. Where series capacitance is used in long lines to compensate for inductive reactance drop, electrical oscillation at the natural frequency f_n may cause torques on generator rotors at system frequency minus f_n. These torques may resonate with torsional natural frequencies of the shafts and produce unacceptable fatigue stresses. From all these causes, complex torsional oscillations develop in the shaft system, with components determined by the inertias and stiffnesses of the several shafts of the turbine, generator and exciter. The shaft dimensions must be chosen to avoid a serious loss of fatigue life during such incidents as well as to satisfy the critical speed criteria mentioned above. There are many articles on the subject; references 128–135 are a few of these.

Bearings are of the white-metalled cylindrical type, with forced oil lubrication and, except on small sets, high-pressure oil jacking also. Jacking allows the set to run slowly (typically 3–20 rev/min) on the turning gear to cool off the turbine and generator rotors before the unit is finally stopped. Without this, the rotors would bend because temperature gradients would occur across the diameter of each rotor, and vibration would occur on the next run-up.

All turbogenerator rotors bend under their own weight, and with two-pole rotors for outputs more than about 30 MW the amount of bend would be significantly more when the pole axis is horizontal than when it is vertical. Hence a vibration would occur at a frequency corresponding to twice the running speed; it is caused by the changing stiffness of the rotor in the vertical plane, and therefore cannot be removed by mechanical balancing. The stiffnesses about the polar (direct) axis and the slot (quadrature) axis must be made as nearly equal as possible under running conditions (when centrifugal force on the windings increases the stiffness in the plane of the quadrature axis). This is done either by cutting axial slots along the pole areas (and filling them with magnetic steel if necessary to avoid magnetic saturation) or by cutting narrow arcuate grooves circumferentially across the poles, sufficient grooves being spaced down the length of the rotor to reduce the stiffness to match that in the quadrature plane.[29,93]

28.15.3 Rotor winding

Excitation currents range from say 400 A at 200 V for a 30 MW generator to 5.7 kA at 640 V for a 1000 MW two-pole machine and up to around 7 kA at 650 V for a 1500 MVA machine. For rotors using indirect cooling each coil is wound with a continuous length of copper strap, bent on edge at the four corners. The copper contains about 0.1% silver and is hard drawn to increase its strength and so avoid the coil-shortening effect that occurred with plain soft copper as a result of heating while part of the copper was prevented (by centrifugal force) from expanding axially.

Directly cooled coils are usually made of larger section conductors of silver-bearing copper containing grooves and holes to provide gas passages. Half-turns are brazed together in the end regions after they have been positioned in the rotor slots. The slots may be parallel-sided, or tapered to contain more copper without increased tooth stress.

Insulation between turns is usually provided by interleaves of resin-bonded glass fabric or some other synthetic material. The coils are insulated from the rotor body by U- or L-shaped troughs of resin-bonded fibreglass, Nomex or melamine, or combinations of such materials. Similar insulating strips insulate the top conductor from the slot wedge; in direct-cooled rotors the strips must have through-holes to allow the cooling gas to escape from the rotor; they must be thick enough to provide adequate creepage distance to withstand the specified h.v. tests.

The end-windings are packed, partly or wholly, with blocks of insulating material to avoid distortion and the consequent risk of short circuits between turns.

28.15.4 Stator core

The stator core is built up of segments of electrical sheet steel, usually 2–3% silicon, 0.35 mm thick, cold rolled and non-oriented. To minimise weight, the core is worked at the highest flux density consistent with reasonable losses. In a two-pole machine the magnetic force across the airgap subjects the core to an elliptical distortion that rotates with the rotor, so producing a double-frequency ($2f$) vibration. The core depth must be chosen so that its natural frequency of vibration in this elliptical mode is well away from $2f$; usually $3f$ or more is practicable without excessive depth of core. Grain-oriented steel has better permeability and lower losses than non-oriented steel, but as it has a lower modulus of elasticity its other advantages cannot be realised without accepting higher vibration. This, and its higher cost, severely limit its use in turbogenerators.

28.15.5 Stator casing

Air-cooled machines may have bearing pedestals on a bedplate; the stator frame then merely supports the core and forms the ventilation enclosure. Alternatively, it may be a more rigid box frame with the rotor bearings carried in endbrackets.

A hydrogen-cooled machine must have a totally enclosed and gas-tight construction; the end-bracket bearing arrangement is adopted to minimise the bearing span and to raise the critical speeds. Hence the frame must be rigid enough to provide proper support for the bearings and to contain the gas pressure that might occur in the unlikely event of a hydrogen–air explosion inside the frame. This could produce pressures up to about 1400 kN/m^2; therefore this pressure, rather than the continuous working pressure of hydrogen, becomes the design criterion.

In large two-pole machines (which are invariably hydrogen cooled) some form of flexible mounting is needed between the core and the casing, and the casing should not have any natural frequency near to $2f$. This is to avoid excessive magnetic noise and the risk of unacceptable vibration on the casing, coolers or pipework.

Where transport facilities are inadequate for handling the complete stator, the core and windings must be made separately from the outer casing, separately transported, and assembled on site before the rotor is inserted. By contrast, smaller machines can be transported complete to some sites, with the rotor clamped in temporary supports; this arrangement facilitates erection.

28.15.6 Stator winding

For small machines the voltage is usually fixed at a standard network voltage (e.g. 6.6 or 11 kV), but for large machines, where a generator-transformer is used, the designer has a free choice. A high voltage avoids difficulties due to high currents, but valuable space in the slots has to be sacrificed to insulation; a compromise is thus about 15 kV for 100 MW and 200 MW machines, and up to 22 or 25 kV for the larger sets. Even so, generators of more than about 50 MW rating will have two circuits in parallel per phase; for more than 1000 MW it may be necessary to use special winding arrangements to have four parallel paths in a two-pole machine. In four-pole generators, four circuits occur naturally and can be in parallel or in series-parallel.

The winding is of the two-layer basket type, almost always with integral slots per phase per pole, as described in Sections 28.3 and 28.4. In the slots[27] and in the endwindings[101] the coils must be supported to resist electromagnetic forces.[93] These are, on normal load, continuous though fairly low vibratory forces at twice supply frequency. When a short circuit occurs close to the generator, transient oscillatory forces occur, 50–100 or more times greater than those on load.

In the slots the forces are radial, directed towards the bottom of the slot, except where different phases occupy the same slot; there the force on the top conductor is towards the wedge for part of each cycle. To support the coils along the whole core length, conformable packing strips of, for example, resin impregnated polyester fleece are placed beneath and between the coil sides. Slot wedges are fitted that apply a known radial load to the coils, greater than the electromagnetic forces. Packing may be fitted down the sides of the coils, and there may be a fibreglass ripple spring between the wedge and the top coil side to take up any small shrinkage in service.[27,29,93]

The endwindings are secured to a strong structure of insulating materials,[27,29,93,100] fibreglass rings carried on brackets of resin-bonded wood laminate have been used very widely. For larger ratings a solid cone of filament-wound resin bonded fibreglass is used for greater strength and long-term rigidity. The coils are bedded to the structure with conformable packing material. Usually the slot and endwinding packings are cured while the coils are held by temporary wedges and clamps, which are then replaced by the permanent ones. The complete structure is bolted to the end of the core; some axial movement may be allowed, to accommodate expansion of the coils relative to the core.

28.15.7 Cooling

Efficiencies are between 96.5% and 99%, increasing with the rated output. However, the losses will be 0.5–15 MW, appearing as heat that must be removed by circulation of an appropriate cooling medium: oil for removing bearing-friction losses, and air, hydrogen or water for other losses. Details of the cooling media used for stator and rotor windings are given in *Table 28.6*. The heat transfer coefficients

Table 28.6 Properties of cooling media

Medium heat	*Absolute pressure* (bar)	*Specific heat* (kJ/(kg K))	*Density* (kg/m^3)	*Relative volume flow*	*Relative thermal capacity*	*Relative heat transfer coefficient*
Air	1.0	1.0	1.1	1.0	1.0	1.0
Hydrogen	1.0	14.3	0.076	1.0	0.99	1.45
	2.0	14.3	0.152	1.0	1.98	2.52
	4.0	14.3	0.304	1.0	3.96	4.38
Water	1.0	4.2	1000	0.01	38	60

are typical, but depend considerably on velocity and duct size.

Traditionally, indirect air cooling was adopted for outputs up to 70 MVA, and indirect hydrogen cooling for outputs above about 50 MVA. Direct hydrogen cooling of rotors developed rapidly for ratings from about 100 MVA up to the largest, about 1500 MVA, that have been built. Some manufacturers used hydrogen cooled stator coils up to about 650 MVA: others adopted water cooling above about 150 MVA where indirect cooling was becoming difficult. In recent years cheap and simple designs have been developed up to 200 MVA rating using air cooling, direct in the rotor winding and indirect for the stator.

28.15.7.1 *Indirect cooling*

The cooling medium (air or hydrogen) is blown along the airgap, through ducts in the core and over the surface of the windings. Thus heat generated in the winding passes through the main insulation to the rotor and stator teeth, respectively, and is picked up by the cooling gas mainly from the iron surfaces.

With air cooling, closed-circuit ventilation is universal except in the very smallest sizes, and coolers are separately mounted—usually in a basement beneath the generator, but occasionally above or at the side of the machine. With hydrogen cooling, however, there is no alternative but to build coolers within the gas-tight explosion-proof structure of the machine itself. *Figure 28.31* shows a simplified diagram of a hydrogen-cooled machine with its gas system. It has directly cooled windings.

The cooling gas is usually circulated by a fan at each end of the rotor, though many air-cooled generators (30–60 MW or so) had motor-driven fans mounted in the basement with the air coolers. The rotor fans may be of the centrifugal or the aerofoil (axial-flow) type. Their main purpose is to establish the gas flow through the stator frame, core and coolers; the flow through the rotor results mainly from its own rotation. Most indirectly cooled rotors have axial ventilation slots in the teeth that are closed by wedges

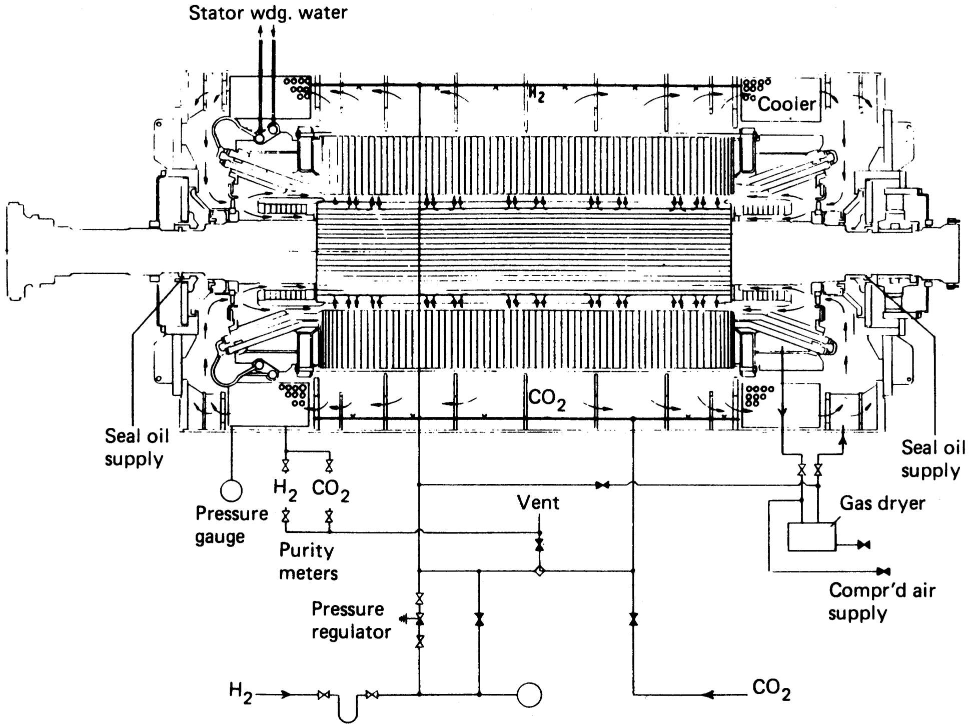


Figure 28.31 Section of a turbogenerator with simplified hydrogen cooling and water-cooled stator winding

except near the middle of the rotor body where the flows from each end emerge into the gap and then pass through the radial ducts in the stator core to the back of the frame.

Pure hydrogen has a density approximately 1/14 that of air, while its specific heat is 14 times that of air; it has a higher heat transfer coefficient and much better thermal conductivity. In service there may be about 1% impurity consisting of air and carbon dioxide; this increases the density by about 13%, but has no significant effect on the cooling properties listed in *Table 28.6*. Windage losses are proportional to density, but even at an operating pressure of 5 bar (absolute) they are still only 40% of what they would be in air at atmospheric pressure.

Early hydrogen-cooled machines were designed to operate at just above atmospheric pressure, but raising the pressure to 2 bar (absolute) and then to 3 bar raised the output available from a given frame size by approximately 15% and then by a further 10% respectively. No worthwhile improvement occurs above 4 bar (absolute) because the temperature gradient across the winding insulation is a large part of the permissible temperature rise.

The auxiliary equipment for hydrogen-cooled generators divides into two main groups: gas control and seal-oil treatment.

The gas control system provides means for filling and emptying the casing without risk of forming an explosive hydrogen–air mixture. Carbon dioxide is used as a buffer, and mixtures containing more than 5% hydrogen in air, or more than 5% air in hydrogen, are avoided. In service the rate of hydrogen loss, though small, is sufficient for the purity to settle at 98–99% hydrogen as make-up is added to maintain the desired operating pressure.

The shaft seals, which prevent leakage of hydrogen along the shaft to the bearings, are supplied with oil maintained at a pressure above the gas pressure. Ring seals which encircle the shaft are simple and allow free axial expansion of the shaft; however, they also allow a significant rate of oil flow towards the hydrogen side of the seal, and rather more to the air side. The gas-side flow absorbs hydrogen, and would release air or moisture into the machine if it contained these in solution. To avoid the consequent pollution, and loss, of the hydrogen, the oil is vacuum treated before being fed to the seals, and the hydrogen-side oil passes through detraining tanks to allow entrained hydrogen to return to the frame before the oil is vacuum treated and recirculated.

The face, or thrust, type of seal is a ring, usually of white metalled steel, which operates against the radial face of a collar on the shaft. The hydrogen-side oil flow is insignificant, so vacuum treating is unnecessary. The extremely thin hydrogen-side oil film (say 60 μm) makes the seal rather vulnerable to dirt particles, and if the ring does not slide freely to follow shaft expansion it will leave the collar and allow leakage or will suffer excessive face pressure and damage to the white metal. Both types of seal are in satisfactory operation. A doubly fed ring-type seal offers the advantages of both types, provided that the pressures of the two systems are kept accurately balanced so that the hydrogen- and air-side flows are kept separated.

A wide range of indicators fitted with audible and visible alarms is necessary to indicate any departure from normal operation of the various parts of the gas and oil systems.

28.15.7.2 Direct cooling[27,29,123–127]

Hydrogen The winding conductors are much more effectively cooled by passing the coolant through them in direct, or almost direct, contact with the copper. Hence much higher current density is possible with an acceptable temperature rise, and the output per unit volume of active material can be greatly increased. Furthermore, significant improvement in performance with hydrogen cooling is obtainable up to operating pressures of 5 or even 6 bar (absolute) (500–600 kN/m^2).

In the rotor, several flow arrangements are used:

(1) *Axial flow* The gas enters tubular, or axially grooved, conductors in the end-winding region and leaves radially through a group of holes through the conductors and wedges at the mid-length of the rotor.
(2) *Axial flow* This is as for (1), but for longer rotors. The middle quarter (approximately) is fed from subslots, while the two end-portions are fed as in (1).
(3) *Radial* The gas enters each end of a subslot cut beneath the winding slot and flows radially outwards through holes punched through the flat copper strips and insulation, distributed throughout the rotor length.
(4) *Axial–radial* This is a combination of (1) and (3) using axially grooved conductors in which the radial exit holes are displaced axially from those that feed gas from the subslots.
(5) *Gap pick-up* Specially shaped holes in the wedges ‘scoop up’ gas from the gap, and others eject it after it has passed through cooling ducts formed in the copper by punched holes or transverse grooves.

Types (1) and (2) require high-pressure axial-flow blowers, which may have three to seven stages of blades; (3) and (4) rely almost wholly on the self-ventilating action of the rotor; (5) may be applied by dividing the rotor into axially adjacent inlet and outlet zones, which may be co-ordinated with the stator ventilation zones. All these schemes are used for hydrogen-cooled machines up to the largest ratings. Type (3) is now common for air-cooled machines with totally enclosed air circuits.

Directly hydrogen-cooled stator coils are used by some makers up to ratings of 600 MW or more. The gas flows down thin-walled bronze or stainless steel tubes that are bonded among the conductor strips and lightly insulated from them. Entry and exit has to be at the ends of the coils, so a moderately high pressure differential is needed. The system co-ordinates well with rotor ventilation of type (1) or (2).

Water The high heat-removal capacity of water and its low viscosity allow it to be used in tubular subconductors that are still small enough to keep the eddy-current losses low.

At each end of the conductor the subconductors (tubes and strips, or all tubes) are brazed into a water box. The coil-to-coil current-carrying connection may be tubular to carry the water also, or separate connections may be made. The water boxes are connected to the inlet and outlet manifolds by insulating pipes of polytetrafluoroethylene (PTFE) or some other synthetic material. Water taken from the boiler make-up system is circulated by pumps around a closed pipework system containing the winding, coolers, filters, control valves and monitoring instruments. Water conductivity is easily maintained by using a demineraliser unit, usually of the resin-bead type; less than 10 μS-cm is easily attained, and the leakage currents down the water columns are insignificant.

The temperature difference between copper and water is only about 2°C, and the water temperature rise between inlet and outlet is usually about 25–30°C. The inlet water pressure is kept below the hydrogen pressure; if leakage does occur, it is leakage of hydrogen into the water (easily detected), and water does not enter the machine.

Water-cooling of the rotor winding increases the output available from given frame size; the slots can be smaller, leaving more room for magnetic flux, so the machine can be shorter provided the stator design is adjusted to suit. Constructional problems occur because of the need to convey water to and from the rotor, to accommodate water manifolds on the shaft, to support the water-pipes against centrifugal force, and to design them and the winding to withstand internal pressures (produced by rotation) up to around 15 MN/m^2. Nevertheless, such rotors are in successful service though they have not yet been widely adopted.

Direct cooling allows the specific electric loading to be increased without exceeding the permitted limits of temperature rise. A smaller frame size can be used for a given output, but the leakage reactances are increased by the increase of leakage flux compared with main airgap flux. The values of X_d and short-circuit ratio can be maintained by lengthening the airgap, until the increased excitation on load pushes the rotor winding temperature rise to its limit.

28.16 Generator–transformer connection

For small ratings up to 3 MW, single-core cables are used. For ratings above this where the number of cables needed would be excessive or cannot be accommodated, solid copper or aluminium bus-bars are used. The bus-bars are supported on insulating cleats or ceramic post insulators and are enclosed in a surrounding duct which provides mechanical protection and sealing. The bars are spaced and supported to suit the operating voltage, the current to be carried (taking into account skin and proximity effects which raises the a.c. resistance and leads to extra heating) and the electromagnetic forces produced by short-circuit currents. The phases are sometimes segregated by insulating barriers to achieve the creepage and clearance distances required. For higher current ratings, each phase conductor usually consisted of two angle- or channel-section bars mounted face to face to form an open diamond or box-shaped section. These sections and arrangements give lower skin-effect losses than flat rectangular bars of the same weight, and natural air cooling may be adequate up to at least 200 MW (with a corresponding current of the order of 10 kA).

To avoid the possibility of phase-to-phase faults, however, phase-isolated bus-bars were adopted for ratings above about 200 MW, and now above 60 MW. Each line connection consists of two angle- or channel-section bars supported by post insulators inside an aluminium tube which is physically and electrically continuous along its length. The tubes of the three phases are connected together electrically at each end of the run and are joined mechanically to the generator and transformer frames. Thus only phase-to-earth faults are possible. Eddy currents induced in the aluminium tubes cause additional losses, but they confine the magnetic field largely within the tube, so that heating of the foundation steelwork or other structures is avoided.

For the higher ratings, say above 12 kA, each conductor bar may be of semi-hexagon or semi-octagon shape, so that the complete conductor approximates to the circular cross-section that gives minimum skin effect and minimum loss.

Table 28.7 Typical dimensions of generator–transformer connections

	Cooling		
	Natural air	*Forced air*	*Pumped water*
Conductor shape	[ ]	[ ]	○⇐
Diameter of circumscribing circle (m)	0.9	0.5	0.13
Diameter of isolating trunking (m)	1.5	1.1	0.75
Total loss in conductor and trunking, for 3 phases (kW/m)	2.0	4.5	7.5

Natural air cooling is practicable up to about 20 kA, but a significant reduction in cross-section or an increase in current rating is made possible by forced-air cooling. A maximum conductor temperature rise of 55°C above the ambient air is usual. With this, and a continuous rating of 19.5 kA (660 MW at 23 kV, 0.85 p.f.), typical dimensions are as given in *Table 28.7*.

For forced-air cooling, the flow is usually from one end of the run to the other along one phase, half the flow returning along each of the other two phases. The air circuit is totally enclosed, with an air-to-water heat-exchanger extracting the loss. If the air circulation fails, the naturally cooled rating is about 60% of the forced-flow rating.

The use of water-cooled stator windings led to the development of water-cooled connections in solid tube or cable form. The water-cooling circuit is similar to, but usually separate from, that for the generator stator. The smaller dimensions are an advantage where space is limited, and the greater loss is not usually economically unacceptable. The system has not been widely adopted, however, because there is usually room for air-cooled connections, while the extra water-cooling auxiliaries introduce additional maintenance and extra complications of duplication and control to guard against shut-down of the generator if an auxiliary item fails.

28.17 Hydrogenerators

28.17.1 Introduction[26–28,30,53,138–156]

The design of hydrogenerators is determined mainly by mechanical considerations. Outputs range from less than 1 MVA to over 800 MVA, and speeds from 50 to 1000 rev/min, depending on the water head available and the type and size of the turbine. The low speeds require the generators to be physically large, and it is often necessary to transport them to site in sections. The inertia required in the set is determined by turbine governing or speed regulation requirements, or by the transient stability of the associated power system. The turbine contributes little flywheel effect, so the generator inertia must often be more than that of a design that satisfies the electrical specification in the least expensive way. The diameter must be increased, or in extreme cases a separate flywheel coupled to the shaft. This is often the best arrangement in fairly small horizontal shaft units.

A water turbine runs up to a high overspeed when load is suddenly removed, because the flow of water cannot be suddenly stopped without causing a high and probably damaging rise of pressure at the turbine gates or valves. The ratio of overspeed to normal speed is, approximately, for impulse turbines (Pelton) 1.7 to 1.9, for reaction (Francis) 1.8 to 2.1, and for propeller (Kaplan) 2 to 2.2. If the governor should fail during load rejectien, a Kaplan turbine could run up to 3 times normal speed.

The rotors must be designed to be safe at overspeed, where the factor of safety on the proof stress of the material used is normally not less than 1.5. A figure closer to 1.1 is acceptable for the very rare runaway condition of a Kaplan unit.

The first critical speed is required to be above the overspeed.

28.17.2 Construction

28.17.2.1 General arrangements[8,26,30]

Horizontal and vertical shaft arrangements are used, the former usually for impulse turbines and small reaction turbines, the latter for large reaction and propeller turbines. Nevertheless, the vertical arrangement has been used up to 36 MVA and 1000 rev/min, and the horizontal shaft up to more than 100 MVA at 428 or 600 rev/min.

Horizontal generators are similar to those made for diesel engine or geared turbine drive, except that they must accept the higher overspeed, may carry a flywheel, and need a thrust bearing to carry any unbalanced hydraulic thrust. A turbine may be overhung at each end, or at one end only.

With a vertical shaft, five bearing arrangements are possible. These are shown in *Figure 28.32*. The umbrella arrangements of (*d*) and (*e*) are suitable for low-speed sets in which the ratio of core length to stator bore diameter does not exceed say 1/4. With a top guide bearing added as at (*c*), higher speeds and an *L*/*D* of 1/3 are practicable. The top thrust bearing arrangement of (*a*) and (*b*) is used, where necessary, for higher *L*/*D* ratios and speeds usually above 400 rev/min. The top bracket and the stator frame must be rigid enough to carry the thrust load, and so are more expensive than with a bottom bracket assembly. However, as the bearing is on a smaller diameter part of the shaft, its losses are less; if the contract places a high value on losses (as with pumped storage for example) the capitalised value of the lifetime losses may be enough to pay for the top-thrust layout. If the bracket is supported directly from the pit walls, the stator frame is relieved of the thrust load, and some problems of differential expansion and electromagnetic vibration are avoided.

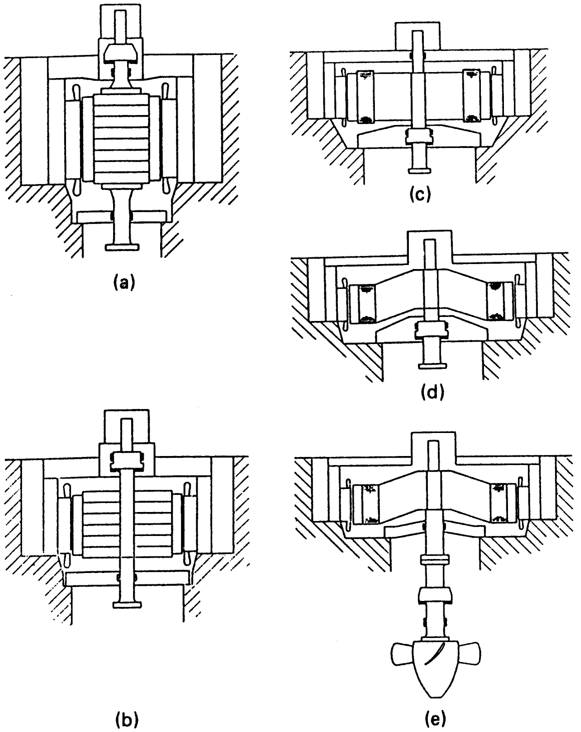


Figure 28.32 Hydroelectric generator bearing arrangements: (a) thrust bearing above rotor, two guide bearings—upper guide bearing separate from and below thrust bearing; (b) thrust bearing above rotor, two guide bearings—upper guide bearing combined with thrust bearing; (c) thrust bearing below rotor, two guide bearings (semi-umbrella machine); (d) thrust bearing and single guide bearing below rotor (umbrella machine in English, American and Continental literature); (e) thrust bearing mounted on turbine casing, single guide bearing below generator rotor (umbrella machine in German literature). (Reproduced, with permission, from Anscombe[30])

28.17.2.2 Thrust and guide bearings

In large vertical machines the dead weight plus hydraulic thrust is several hundred tonnes. The thrust collar rests on segmental pads (usually of steel) faced with white metal. They are supported on a thrust ring at the bottom of the annular oil chamber that surrounds the shaft. The pads and the lower part of the collar are submerged, and careful design and assembly are needed to avoid leakage of oil and vapour.

The pads are supported so that they can tilt slightly to develop hydrodynamic lubrication. The Kingsbury pad is supported on a single spherical pivot: the Michel pad pivots on a radial ridge. Both supports are offset from the centre line of the pad if rotation is to be in only one direction: a pivot on the centre line is needed for a reversing pumped storage set. Load sharing among the pads depends critically on the accuracy of dimensions or adjustment of the pad heights. This difficulty is much reduced if the pad sits on a mattress of closely spaced helical springs instead of a solid pivot. The specific bearing pressure can be quite high, up to 3.5 MPa average over the pad surface. Transient conditions on a pump/turbine unit may increase this at some times by a factor of about 1½. The peak oil pressure on the pad is about 1.5 times the average. The oil film thickness is usually between 0.1 and 0.05 mm. On such highly loaded bearings a high pressure oil supply is often fed through one or more small holes in the surface of the pad to provide oil jacking and hydrostatic lubrication at rest. This is especially needed to reduce the starting torque required when the unit is being motored for pumping duty.

The pumping action of the bearing is often enough to circulate the oil through water-cooled heat exchangers placed in the oil pot. If the pot is too small for this, or to improve accessibility to the pads, the coolers may be mounted separately below the bearing. The oil is then circulated by main (a.c.) and stand-by (d.c.) motor driven pumps.

With high bearing pressures and surface speeds cooling water may be circulated through channels in the pads themselves.

Vertical shaft machines have friction brakes operated by air pressure to bring the set to rest without protracted running at low speeds, when the lubrication of the thrust pads may become inadequate. To reduce wear on the brake pads they are not applied at or near the full speed except in an emergency. Dynamic braking is achieved by circulating up to full-load current through the stator winding short-circuited by a braking switch. The friction brakes are then applied at 15–20% of full speed.

The brakes can also be operated by oil pressure when the machine is stationary, to lift the thrust collar off the pads, either to flood the surfaces with oil after prolonged standing, or for maintenance such as inspection or removal of the thrust pads. If the shaft is to remain lifted for a long time, the jacks are locked mechanically to allow the oil pressure to be removed.

Guide bearings are usually of the pivoted-pad type, running on the outside of the thrust collar or a smaller diameter of the thrust block. Where the guide is remote from the thrust bearing, it runs on the shaft in its own oil enclosure.

If there is a bearing above the rotor, it needs to be insulated to prevent the flow of shaft currents caused by some asymmetry in the core flux, e.g. differences in reluctances at the joints in the core. With a top thrust bearing the thrust face, which is normally a renewable ring bolted to the collar, can be insulated from the collar.

28.17.2.3 Rotor

For large machines the cheapest construction is to secure a laminated rim[143,149] to a fabricated spider that is shrunk on the shaft. The rim must carry its own hoop stress plus the load of the poles and coils. Laminations of rolled steel plate commonly have 0.2% proof stress of 450 MPa, suitable for rim speeds up to about 170 m/s at overspeed. Proof stress up to 700 MPa can be obtained, if necessary.

Sheet can be obtained of quality and width suitable for rims up to approximately 1.5 m diameter, so these can be built of complete rings, shrunk and keyed to the spider with enough interference to maintain contact at overspeed. Larger diameter rims have to be segmental; they will come free at just above running speed, but are kept true to the spider by keys that allow radial growth but not tangential movement. Some designs rely upon this key location, and do not have a shrink fit.

Two kinds of rim are used. The so-called *chain rim* uses segments about 5 mm thick spanning two or more pole pitches, clamped axially by close-fitting bolts. By suitably overlapping the segments and distributing the bolts, the rim can have a hoop strength up to 75% of that of a solid rim. The *friction rim* is being increasingly used. It has segments about 2 mm thick, to increase the number of surfaces in contact. They are clamped with high pressure by high tensile bolts in holes with some clearance. To build a good rim of either type the segments must be flat and of uniform thickness, and accurately aligned to ensure accurate dimensions of the keyways and the axial T-slots for the pole fixings.

Smaller rotors for 600 rev/min or more, especially horizontal ones, are built of discs up to 150 mm thick, shrunk on to a shaft, or spigotted together and held by through bolts, with bolted-on stub shafts. The diameter is limited by the size of plate available of suitable thickness and quality.

Poles are now most frequently built up of laminations about 2 mm thick. They, like the rim, carry unidirectional flux, so tensile strength, flatness, uniform thickness and magnetic permeability are important. It is not usually thought necessary to insulate the laminations, though this would reduce the surface losses caused by harmonic fluxes in the airgap. The laminations are clamped tightly between heavy steel endplates, cast, forged or fabricated depending on the load imposed by their own centrifugal force and the end turns of the field coil.

Poles are fixed to the rim by dovetail, or more often by T-head, projections. These are secured in corresponding axial slots in the rim by taper keys driven in from each end. These pull the base of the pole against the rim. Large rotors may use as many as nine T-heads per pole; detailed analysis of the stress distribution is needed, e.g. by a finite element technique, to ensure that they share the total centrifugal force reasonably equally.

Copper damper bars, usually uninsulated, are fitted reasonably tightly in semi-closed axial slots in the pole face. They are brazed at each end to a copper segment to form a closed grid. Currents induced in them help to reduce the oscillations of load angle when a disturbance occurs on the power system, and to counteract the negative sequence stator current of unbalanced loads. Pole-to-pole links are sometimes fitted to make the grids into a complete cage, but in large machines centrifugal force and fatigue stresses can cause mechanical failures. If interconnection is needed, it is usually sufficient to braze the bars to a few thick copper laminations next to the end clamps of the poles. These provide satisfactory contact to the rim, which completes the circuit between poles.

Field coils on very small machines have several layers wound with round or roughly square-section copper, enamelled and wrapped with resin impregnated insulating fibre or braid. Larger rotors have coils wound with bare copper strip, bent on edge, and larger rotors still[143] use coils fabricated from straight lengths of rectangular section copper strip, brazed together at the corners. The cooling surface is increased by making some strips wider than the others, and placing these, singly or say in pairs, every few turns. With continuously wound strip-on-edge coils the finned effect can be produced conveniently on the sides or the ends of the coil, but not on both.

Insulation between turns consists of two layers of polyimide or aramid paper, or woven glass, with a thermosetting resin. After the coil has been consolidated by baking it under pressure, the interturn insulation is approximately 0.3 mm thick.

Wire-wound coils may be wound directly on to the insulated pole. Strip-on-edge coils may go directly on to the insulated pole, or on to a separate insulated spool which is then secured to the pole. Thick washers of resin–synthetic paper or resin–glass insulate the coil from the rotor rim and from the pole shoe, unless the spool has insulated flanges that do this.

Coil-to-coil connections are secured to the rim, the connection to the top turn (i.e. the outermost one) of a strip-on-edge coil may be brought down inside the coil, between it and the pole, then out beneath the bottom turn. Connections should preferably be arranged to carry the field current equally clockwise and anticlockwise round the shaft. Otherwise it will be magnetised axially, and the bearing surfaces may be damaged by induced currents.

Centrifugal force on the sides of the coil has a component tending to bow the turns, i.e. to bend them on edge away from the pole. To resist this one or several V-blocks (depending on the axial length) are secured to the rim between the poles. Pieces of insulating board of course are fitted between the bare copper and the metal V-block. It is often required that poles may be removed without removing the rotor from the stator bore, and V-blocks make this difficult to arrange. Therefore in some designs, the coils are held by clamps that are secured to the pole itself; the two parts of each clamp are tightened together after the coil and its insulation have been fitted. Then the pole, coil and clamps can be withdrawn as a unit after the pole keys have been released.

28.17.2.4 Stator frame and core[26, 143]

The frame is a fabricated assembly of rings of steel plate connected by axial members of tubular, angle or channel section. It is often necessary to split the frame and core of a large vertical shaft machine into several segments for transport to site. These must be bolted together accurately to form a true circle, with the joint faces of the core fitting extremely closely together.

The core joints are avoided if the core is built into the assembled frame sections on site.

The lower endplate of the frame is bolted to the foundations; if the machine has a top thrust bearing the frame may have to support the thrust load as well as the core and windings.

The wrapper plate round the outside of the frame will usually have openings to receive the air-to-water heat exchangers.

The core is built of cold-rolled silicon steel, usually 0.35 or 0.5 mm thick. The reduction of loss that could be obtained with oriented-grain sheet is small, and rarely enough to justify the higher cost of it. The coreplates are assembled with spacers to form radial ventilating ducts, and are clamped between strong endplates.

28.17.2.5 Stator winding (see Section 28.4)[26, 53, 143]

This is usually of the two-layer, fractional-slot type, using diamond coils with either lap or wave connections. The number of slots is often chosen to permit two or more parallel paths per phase, depending on the number of poles. Coils may be pulled, formed or made as single bars, depending on size. In large high-voltage machines they must be firmly secured in the slots and endwindings, especially if they are to suffer much thermal cycling, e.g. in pumped-storage units. Proper earthing in the slots and stress grading for a distance beyond the slot ends are essential to avoid surface discharges. Large and important machines now often have permanent instruments to give warning of any increase in electrical discharges in service.

28.17.3 Cooling

Except for small machines, closed-circuit air cooling is used. Most hydrogenerators rotate in one direction only, and radial-flow or axial-flow fans mounted on the shaft are commonly used. The heat is removed by air-to-water heat exchangers mounted on the back of the stator frame. About 4 m^3/s of air is required per kilowatt of loss removed. In cold climates, some of the hot air may be taken from the machine and used for heating the station.

Reversing sets for pumped storage would have to use radially bladed fans, which have poor performance and efficiency. For these machines, and for highly rated machines needing carefully controlled ventilation, several motor driven fans are used.

Hydrogen cooling has not been applied to hydrogenerators, mainly because of the cost and practical difficulties of making an explosion-proof casing. However, when inertia is not a controlling factor, a useful reduction in physical size can be achieved by water cooling the stator or rotor windings, or both.[26] Water cooling of only the stator winding introduces a multiplicity of joints which must not leak, but greatly increases the current capacity, or conversely allows the slot size to be reduced for a given rating; this slightly reduces the outside diameter of the core and significantly reduces the leakage reactance.

Water cooling of the rotor winding introduces more constructional difficulties and affects the design more profoundly. The excitation capability is increased without risk of exceeding the specified temperature limits, so a higher short-circuit ratio (longer airgap) is possible, improving the underexcited (line charging) capability. Alternatively a smaller machine with higher electric loading is possible with the stator design suitably adjusted. However, it may be necessary to adopt a size larger than the smallest determined from purely thermal considerations. This larger size may be needed to attain the desired inertia constant, or the lower capitalised cost of its lower losses may be enough to offset the lower first cost of the smaller frame. In some circumstances, water-cooled stator and rotor windings may be economically justified at ratings as low as 150 MVA; conversely, water-cooled stator windings with some form of improved air cooling for the rotor may be preferred on grounds of adequacy and simplicity for ratings as high as 700 MW.

28.17.4 Excitation

Vertical-shaft generators may use main and pilot a.c. exciters mounted above the generator, or separate motor-driven exciters. These often have a flywheel to maintain the exciter speed and output during momentary interruptions or reduced voltage of the motor supply.

Static thyristor equipment is now more often used, because it simplifies the mechanical arrangement of the unit, reduces the height, and avoids the possibility of an unacceptable run-out at the top of a tall shaft assembly. Thyristors also have the advantage of inherently high response, which is valuable when the generator feeds along transmission lines.

Slip-rings and brushgear are rarely troublesome, as the peripheral speed can be fairly low; 40 to 50 m/s is a usual limit.

Horizontal-shaft generators may use brushless exciters, overhung or two bearing depending on size and speed. Again static thyristors simplify the mechanical layout, especially if the generator has a turbine at each end.

28.17.5 Pumped storage units

Pumped storage units were originally installed as peak-levelling units, running as generators at times of high system load, and as motors pumping water up to the top reservoir during light-load periods. A unidirectional set has separate pump and turbine, whereas a reversing set uses the same hydraulic machine either as a pump or as a turbine, depending on the direction of rotation.

The original purpose has been extended to provide spinning reserve and to deliver power into the system at a ‘few seconds’ notice to assist in maintaining stability if other generation, or a system interconnection, is suddenly lost. This introduces particular problems of thermal cycling and mechanical fatigue, especially with reversing sets, which may be required to go from full-speed pumping to reversed full-speed generating within seconds, and to do this perhaps several times a day. For reversing units, separately driven fans are usually provided because rotor-mounted fans designed for both directions of rotation have low efficiency. Water cooling may be applied, as for generators, and may be particularly valuable for the damper cage if this is used for 'induction-motor' starting.

Figure 28.33 shows a cross-section of one half of the motor/generator of a reversible pumped-storage unit rated at 330 MVA, 18 kV, 0.95 p.f., 500 rev/min. The design at this output and speed approaches the limit achievable with present materials and air cooling. The outside diameter of the stator core is 6.2 m, the rotor diameter 4.5 m and the active core length 3.6 m. The mechanical design of the rotor is dominated by centrifugal stresses and the fatigue effects of reversals. The rotating parts weigh about 440 t, and the hydraulic thrust raises the load on the thrust bearing to almost 600 t.

28.17.5.1 Starting[26,146]

Methods available for starting and run-up of a machine in the pumping mode are: (i) by a direct-coupled auxiliary starting turbine or pony motor, (ii) by back-to-back connection with another machine driven by its own turbine and acting as a generator; (iii) from the power network through a step-down transformer, using the rotor solid pole-shoes or the damper winding as a cage for an 'induction' start; and (iv) from the power network through a variable-frequency thyristor converter controlled to give an output over the range of a few hertz up to normal system frequency. In (iii) the pole-shoes or damping windings must be designed to carry the induced currents without excessive rise in temperature. In (iv) the machine moves from rest by induction torque, but at a low frequency it synchronises with the converter and thereafter remains in synchronism up to normal frequency, to be then synchronised with the power network. The starting equipment is expensive, but run-up is more readily supervised, and the damping cage (or pole-shoe) design is not constrained. The method is preferred for large units.

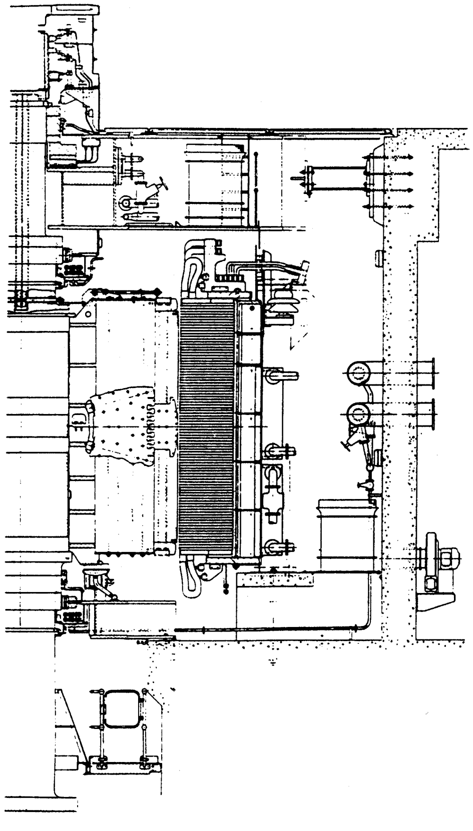


Figure 28.33 Section of a pumped-storage motor/generator

28.18 Salient-pole generators other than hydrogenerators

These are made for synchronous speeds ranging from less than 100 rev/min to 1500 rev/min for 50 Hz and 1800 rev/min for 60 Hz supplies: hence they have from 4 to more than 72 poles. Small 2-pole generators are also made in large numbers. The outputs for salient pole generators range from a few kilovolt-amperes up to about 60 MVA. Prime movers range from internal combustion engines burning petrol, gas (either natural gas or methane obtained from land-fill sites or biomass schemes) or diesel fuel and steam or gas turbines. Direct drive by an internal combustion engine is practicable over the whole range, though at 1500 and 1800 rev/min the output is limited by the maximum engine available power of about 5 MW. Typical ratings for low-speed two-stroke diesel sets are 15–60 MVA at 150–100 rev/min. Diesels, often four-stroke, in the speed range 428–1200 rev/min (14–6 poles) are particularly common, with outputs of say 10–45 MW. Four-pole generators with outputs up to about 60 MVA are driven via a suitable gearbox by high speed (up to 15 000 rev/min) steam or gas turbines.

28.18.1 Applications

Salient-pole generators are used both for stand-by applications and to continuously supply power. They can therefore be connected to public power systems, incorporated in marine installations (ships and oil rigs) or in a great variety of industrial plants. Where a *public system* requires unit ratings up to about 40 MVA, high speed, medium speed or low speed diesel sets can be installed more cheaply and more quickly than a gas or steam turbine set with its boiler and auxiliaries. The internal combustion engine set can deliver full load within a few minutes of starting, which can be done remotely if necessary; it will respond rapidly to changing load demand, and will have better efficiency at part load than the turbine. The slow two-stroke diesel is particularly suitable for the larger outputs (15–60 MVA), running economically on low grade fuel and with low maintenance costs. The recent focus of attention on emissions has had a marked effect on the choice of type of i.c. engine to be used. Engines burning gas are able to achieve lower level emissions of nitrous and sulphur oxides and are therefore proving very popular for new installations. In contrast, achieving acceptable levels of emissions with 2 stroke diesel engines burning low grade fuel is more difficult and costly.

Combined cycle installations are being increasingly used to attain higher thermal efficiency. Exhaust heat from a gas turbine that drives one generator is used to supply steam to a turbine that drives another. The turbines are high speed, geared down to, usually, the four-pole speed. For outputs greater than the gearbox limit of about 60 MW, two-pole cylindrical rotor generators are directly coupled to the turbines.

For *marine applications*, high-speed or medium-speed generators are used. These can either be driven by dedicated diesel engines, shaft driven from the engine used for propulsion or, in larger installations by gas turbines.

For *ship propulsion*, medium-speed diesel or gas turbine driven generators supply d.c. through rectifiers, or variable frequency a.c. to motors coupled to the propellers. The domestic load may be supplied from the same generators, but often separate generators are used—this has the advantage of isolating the domestic supply from the converter harmonics.

Oil rigs can either use internal combustion engine-driven generator sets or gas-turbine-driven generators for their power depending on their size. The generators are required to supply the large drilling motors and mud-pump motors, and the domestic supply. The number, rating and reactances of the generators are chosen to avoid excessive voltage dip when a large motor drive is started, so that other loads are not badly affected. Rated outputs can be around 15 MW where gear-driven generators running at 1800 rev/min would be used or up to 25 MW directly coupled to the power turbine at 3600 rev/min.

Industrial power stations commonly use steam turbines if there is need for process steam, which is delivered through a back-pressure or pass-out turbine; or if waste heat is available to raise steam for generating electricity.

Gas and diesel engines are often used, alone or in combined cycle installations. Gas turbines are expensive to run, and are used only if rapid response to load is essential, or the fuel is locally not so expensive, e.g. in an oil refinery.

28.18.2 Construction

The *rotor construction* for these salient pole machines is determined by the rating of the unit and its peripheral speed. High-speed rotors with four or six poles and ratings up to about 30 MW can have integral poles which are made up of a series of laminations. These can be punched from sheet material using a die or a nibbling type press; alternatively, they can be manufactured using a laser cutting machine. The thickness of punchings is normally limited to less than 3 mm if one of these processes is used, which has the advantage of limiting the pole-shoe losses. Thicker plates up to 100 mm thick can also be used, in which case the pole profile is cut using a numerically controlled flame cutting machine. With this type of construction, due consideration must then be given to the extra surface losses which will occur in the pole-shoes.

For larger machines, i.e. above 30 MW, with bigger diameter rotors, the peripheral speeds will result in rotational stresses in excess of the safe limits for the laminated type of rotor materials. In this case, forged solid rotors with pole pieces integral with the shaft are used. Solid forged pole-shoes are secured by high tensile strength bolts after the coils have been put on.

For the lower peripheral speeds associated with multipole rotors, separate laminated pole pieces may be secured by dovetails or T heads in slots in the square or hexagonal middle portion of the forged shaft. Alternatively, the laminated poles may be bolted or dovetailed to the wide cylindrical rim of a flywheel on the shaft; or they can be fixed by dovetails or T-head slots in a laminated rim.

The laminated-pole construction is cheaper than a forged rotor, and reduces the cost of the stator winding too, because fewer, wider, stator slots can be used with acceptably low eddy current losses in the pole faces.

Generators may have one or two *bearings*. Many diesel driven generators have only an outboard bearing, the driving end of the shaft being solidly coupled to the engine crankshaft flange. This shortens the set and saves the cost of a drive end bearing and the flexible coupling. It also allows the generator to be partly supported on a Society of Automobile Engineers (SAE) flange on the engine crankcase. The crankshaft bearing must carry half the weight of the generator rotor. A gearbox bearing can rarely do this, so gear-driven generators normally have two bearings.

The generator may be supported on sole plates grouted into the foundations, or may be bolted directly to the engine base plate, where this has been designed for the purpose. If transport and site lifting facilities permit, the generator can be delivered complete on its own bed plate; this is to be preferred as it saves erection time on site.

Often it is desirable to limit the vibration transmitted to the foundations, especially when the driver is a diesel engine. A common method is to put flexible mountings beneath the combined bed plate. These need to be chosen, or tuned, to isolate the vibration and not to permit resonance at any of the disturbing frequencies.

Class F *insulation* is always used on high voltage stator and some rotor windings, although class B temperature limits are often specified to extend the life of the machine and give some margin for an overload capability. Class H systems are common on low voltage generators and also on some medium voltage machines. Often class F temperature rises will still be specified even for machines with class H insulation systems in order to achieve an extended life.

Stator coils are usually of the pulled diamond type (see Section 28.4.1). Rotor coils are generally similar to those of hydrogenerators, but as they are smaller they are more often of the wound on edge construction using rectangular strip copper rather than fabricated. Where the rotor has integral poles, the field copper is wound directly onto the pole pieces by rotating the complete rotor or the laminated rotor core-pack about its longitudinal axis.

28.18.3 Ventilation and cooling

These generators are always air-cooled. If an adequate supply of clean air is available, open ventilation with an appropriate class of protection is used. If the air is only slightly dirty, it may be ducted to and from the machine, with filters on the inlet side. In a short machine the air is drawn through from one end to the other by a single centrifugal shaft-mounted fan. Longer machines have radial ducts in the stator core; the air enters the stator bore at both ends and is expelled radially through the ducts to the back of the core. In this case, shaft mounted axial-flow fans are used to drive the air through the machine.

In dirtier surroundings a totally enclosed machine is used. The primary (internal) air is circulated by shaft fans through an air to air heat exchanger (c.a.c.a. arrangement). The secondary (external, ambient air) is driven through the heat exchanger by one or more separate motor driven fans. Alternatively a water-cooled heat exchanger is used (c.a.c.w. arrangement). The materials used in the heat exchanger must be chosen to suit the quality of the water, e.g. tubes of aluminium brass or cupro-nickel if the water is salty. The limits of temperature rise may be adjusted in accordance with the maximum temperature of the secondary coolant. (see Section 28.5.)

28.18.4 Particular design requirements

28.18.4.1 Machines for operation in hazardous atmospheres

As a result of a vast increase in power generation for the oil industry, and a few catastrophic accidents, generators are required to be designed to operate safely in potentially hazardous atmospheres. BS EN 60079-10:1996 defines the different levels of hazard. Generators, especially when driven by gas turbines, are considered to operate in a zone-2 area. They must be designed and made to prevent any gas that is present being ignited, and compliance with the requirements must be certified by a type-N certificate granted by one of the nominated inspecting authorities. The requirements are specified in BS 5000-16:1997. They are many and complicated, but the major ones can be summarised thus. For a particular group of gases, the surface temperatures of all bare live parts must not exceed an ignition temperature defined for that group. Mechanical clearances on fans, including separate motor-driven fan units, must be greater than usual to prevent any contact between the stationary and the rotating parts, to avoid sparking or excessive local temperatures. Auxiliary electrical devices, including terminals, heaters, fan motors, etc., must all be of certified origin, with certificates of conformity to demonstrate that they comply with the requirements.

28.18.4.2 Reactances

X_{d}, X'_{d} and X''_{d} are those that have most influence on the design and the operation of the generator. A maximum X_{d} or minimum short-circuit ratio is usually specified. A minimum value of X''_{d} or a maximum X'_{d} are either specified or implied by specifying the permissible fault megavolt-amperes or the grade of voltage regulation required. For example, BS 4999-140:1987 specifies voltage regulation grades in terms of

(1) accuracy of voltage control under steady load;
(2) voltage dip when specified loads are suddenly applied; and
(3) voltage rise when full load is suddenly removed.

This performance is with the automatic voltage regulator in control of course, but implies lower X'_{d} to achieve the smaller voltage dips and rises. As implied in, Section 28.6.1.4 it becomes expensive to specify a closer grade of voltage regulation than is really needed. The grade may be unattainable if X''_{d} is high to limit the fault megavolt-amperes.

28.18.4.3 Generators driven by internal combustion engines

ISO standard 8528 (equivalent to BS 7698:1993) entitled ‘Reciprocating internal combustion engine driven a.c. generating sets’ is applicable to generators used in these applications. Most of the standard is concerned with the engines, controls, etc., but Parts 3, 5 and 9 contain requirements for the generator, including definitions and limit values of parameters to do with voltage control.

The mechanical design of such sets must consider the transverse vibration and critical speeds of the shaft assembly, and also the possible torsional modes. BS 5000-3:1980 and ISO 8528 place the responsibility for seeing that the torsional behaviour is investigated with the supplier of the set, assisted by the makers of the engine and the generator. Some cyclic irregularity of the generator speed is inevitable: the rotor inertia may be enough to keep it acceptably small, but an extra flywheel may be needed.

Cyclic irregularity of torque will occur at frequencies determined by the number of firing impulses per second. In low-speed sets (100–150 rev/min) one of these harmonic frequencies may lie near to 10 Hz, and around this frequency the eye is very sensitive to light flicker caused by fluctuation of voltage as small as 0.5%. The fluctuation may affect the operation of electronic equipment, mains ripple control systems, etc. If it is impracticable, or too expensive, to reduce the speed fluctuation by adding more inertia, the voltage swing can be reduced with a high response excitation system, phase-controlled to swing the generator flux in opposition to the speed.

If the cylinder torques are not all equal, there will also be torque fluctuations at rotational frequency or low multiples of it, i.e. in the frequency range 1 ½ to 4 ¼ Hz. For most sets the natural frequency of electromechanical oscillation relative to the power system lies in this range. The resonant, or near-resonant, swings of load angle, power, voltage, etc., may nevertheless be tolerably small if the generator has a sufficiently effective damper winding.

28.19 Synchronous compensators

Synchronous compensators are synchronous motors running without mechanical load; they are used to generate or absorb reactive power, in order to control the voltage of a power system. Hence they are usually installed near a load, or part way down a long transmission line to support the voltage at the intermediate point.

At times of heavy load, compensators run overexcited to supply the magnetising power demanded by the load (transformers and induction motors) or the inductive I^2X losses of the line. At light load they must run underexcited to take reactive power from the line to offset the capacitive line-charging current and so avoid excessive voltage rise. In many h.v. systems (200 kV and above), the line capacitive power exceeds the load magnetising power, even at times of heavy load.

Static inductors and capacitors, switched to suit the system conditions, can be used for the same purpose, but the synchronous compensator has the advantage of providing continuously variable control, and with thyristor excitation it can have a response fast enough for many contingencies.

Thyristor-controlled static compensators give rapid and continuously variable control, but require filters to limit harmonic generation in the power system. Lower maintenance and running costs give them an advantage over rotating machines in most new installations.

Synchronous compensators up to 300 MVAR are in service. Air cooling has been used up to ratings of 40 MVAR, but hydrogen cooling is now normal to reduce the size and the light-load losses, the latter by reducing windage. As the shaft-end need not emerge from the hydrogen-tight casing, no shaft seals are needed. Losses at full load (overexcited) are in the range 0.01–0.016 MW per MVAR of rating.

The underexcited capability is usually about half the overexcited rating; for this a short-circuit ratio of about 0.75 is desirable, to ensure that at the underexcited capability the rotor has sufficient positive excitation to maintain stability. A short-circuit ratio of 1.3–1.5 will provide an underexcited capability level equal to the overexcited level, but the machine is larger and has higher losses than the design with the lower short-circuit ratio. Water cooling has been used for stator and/or rotor windings at ratings of 200 MVAR or more.

The compensator may be run up to speed as an induction motor through a step-down transformer, by means of a direct-coupled pony motor, or by using a variable-frequency inverter.

28.20 Induction generators[157–164]

If an induction motor is driven above synchronous speed it will deliver power to the system, with a slip of about −0.05 at full load. It has the advantage of simple construction, and needs no excitation, speed governing, or synchronising. This makes it cheaper than a synchronous machine and operationally more convenient, e.g. for unattended hydrostations or wind-driven generators. The disadvantage is that it must draw from the power system magnetising power of 0.5–0.75 of its rated active power output, and this has limited the size to about 5 MW.

Research in the USA has shown that, by using static var compensators to supply the reactive power, it should be practicable to run induction generators of a few hundred megawatts output either in parallel with synchronous generators or even as a separate supply system. Speed control would then be essential to fix the frequency of the separate system.

Connection to the power network can be made merely by closing the breaker when the machine is up to synchronous speed. To reduce the current surge in large machines, the machine can be allowed to build up to normal voltage by first connecting a capacitor and then synchronising in the usual way. By suitable design of the machine and the static compensator, efficiency and stability can be comparable to those of a synchronous machine.

The most suitable locations appear to be where transmission by h.v. cable is required, for the cable capacitance will contribute to the reactive-power requirement, and at points in the system where substantial var support is installed anyway, the generator being run when active power is also required.

To maintain stability following system faults requires considerable reactive-power capacity beyond that needed for steady full-load operation. Where the system is strong enough, however, this is not too costly, and the total cost of the induction generator installation can be less than that of a synchronous unit.

28.21 Standards

A selection of standards relevant to generators published by ANSI, BSI, Cenelec, IEC, IEEE and NEMA is listed below. Revisions are made every few years, so care must be taken to use the most recent issue, or an earlier one if that is relevant. Member countries of the European Economic Community are required not to have national standards that conflict with Cenelec Euronorms or Harmonised Documents. Some of these documents are technically equivalent to, or closely similar to, IEC standards that were adopted as the bases for harmonisation. British Standards are technically equivalent to relevant Cenelec documents, or to IEC standards if a corresponding Cenelec standard has not been published. BS EN 60034 parts 1 to 22 are numbered consistently with IEC 60034–1 to 60034–22. For a full comparison, see BS 4999: Part 0. The parts of BS 5000 relate to machines of particular types or for particular applications, and call for parts of BS 4999 where appropriate.

No.	*Title*	*Corresponding*	
		IEC 60034	*CENELEC 34HD*
BS 4999			
Part 102	Methods of determining losses and efficiency from tests	Part 2 and 2A	—
Part 104	Methods of test for determining synchronous machine quantities	Part 4	—
Part 144	Specification for the insulation of bars and coils of h.v. machines	—	345
BS EN 60034			
Part 1	Rating and performance	Part 1	
Part 3	Specification for turbine type synchronous machines	Part 3	—
Part 4	Methods for determining synchronous machine quantities from tests	Part 4	—
Part 16	Excitation systems for synchronous machines	Part 16	
Part 22	A.c. generators for reciprocating internal combustion (RIC) engine driven generators	Part 22	
ANSI C50			
Part 10	General requirements for synchronous machines		
Part 12	Requirements for salient-pole synchronous generators and generator/motors for hydraulic turbine applications		
Part 13	Requirements for cylindrical-rotor synchronous generators		
Part 14	Requirements for combustion gas turbine–driven cylindrical-rotor generators		
Part 15	Requirements for hydrogen-cooled combustion gas turbine-driven cylindrical-rotor generators		
ANSI/IEEE			
115	Test procedures for synchronous machines		
115A	Standard procedures for obtaining synchronous machine parameters by standstill frequency response testing		
421.1	Definitions for excitation systems for synchronous machines		

Cont'd

421A	Identification, testing and evaluation of dynamic performance of excitation control systems
NEMA MGI	
Part 32	Synchronous generators
Part 33	Definite purpose synchronous generators for generating set applications

Copies of standards can be obtained through the addresses shown for the standards authorities in Chapter 49; BSI, ANSI and IEEE documents are available directly from, respectively:

BSI Sales, Linford Wood, Milton Keynes MK14 6LE, UK

ANSI, 1I West 42nd Street, New York, NY 10036, USA

NEMA 1300 North 17th Street, Suite 1847, Rosslyn, VA 22209, USA

IEEE Service Centre, 445 Hoes Lane, Piscataway, NY 08854, USA

Most IEC standards are available from BSI Sales.

Acknowledgements

Acknowledgements are given to the following.

IEEE for permission to reproduce (a) Figure I from reference 19 as *Figure 28.32*, and (b) portions of Tables 1 and 2 from reference 22 as *Table 28.2*.

BSI for information from BS EN 60034-1 used in Sections 28.4.4, 28.5 and 28.14; from BS EN 60034-3 used in Section 28.5 and from BS 4999: Part 0 used in Section 28.21.

ANSI for information from Standards C50: Parts 10–15 used in Sections 28.4.4 and 28.5

IEEE for information from standard Standard 421.1 used in Section 28.14.

Copyright in these publications is held by the organisations listed.

The author expresses his thanks to the Directors of Newage International Ltd for their permission to carry out the necessary editing and updating of this chapter.

His thanks are also given to Mr Albert Hunt and his many other ex-colleagues at ALSTOM for the assistance given at the time of preparation of the text contained in the previous edition of this book, from which much of the current material is derived. In particular, Messrs. G. K. Ridley and I. McShane in connection with Sections 28.17 and 28.18, Mr H. S. McNaughton (Section 28.4); Dr G. K. M. Khan and Dr T. W. Preston for the preparation of *Figures 28.1* and *28.6*; Mr N. C. W. Grocott, Prof. A. B. J. Reece and Dr R. D. M. Whitelaw for information and discussion on many subjects; and Mr P. H. Conceicao and Dr R. A. Hore for information and discussion.

References

Books (arranged in reverse order of publication date)

1 CHALMERS, B. J. and WILLIAMSON, A. C., *A.c. Machines: Electromagnetics and Design*, Research Studies Press, Taunton (1991)

2 AMES, R. L., *A.c. Generators: Design and Application*, Research Studies Press, Taunton (1990)

3 SMITH, J. R., *Response Analysis of a.c. Electrical Machines: Computer Models and Simulation*, Research Studies Press, Taunton (1989)

4 TAVNER, P. J. and PENMAN, J., *Condition Monitoring of Electrical Machines*, Research Studies Press, Taunton (1989)

5 CHALMERS, B. J. (ed.), *Electrical Motor Handbook*, Butterworth, London (1988)

6 SAY, M. G., *Alternating Current Machines*, 5th edn Pitman, London (1983)

7 FITZGERALD, A. E., KINGSLEY, C. and UMANS, S. D., *Electric Machinery,* 4th edition, McGraw Hill, New York (1983)

8 WALKER, J. H., *Large Synchronous Machines: Design, Manufacture and Operation*, Oxford Scientific Publications, Clarendon Press, Oxford (1981) (Paperback edition from University Microfilm International via White Swan House, High St., Godstone, Surrey, England.)

9 SARMA, M. S., *Synchronous Machines: Their Theory, Stability and Excitation Systems*, Gordon and Breach, New York and London (1987)

10 GUILE, A. E. and PATERSONS, W. *Electric Power Systems*, Vol. 1, Chap. 7, Synchronous machines, Pergamon Press, Oxford (1977)

11 KOSTENKO, M. and PETROVSKY, L., *Electrical Machines*, Vol. II, A.C. machines, 3rd edition, Mir Publishers, Moscow 1977 (translated from the Russian)

12 ADKINS, B. and HARLEY, R. G., *The General Theory of A.C. Machines*, Chapman and Hall, London (1975)

13 SAY, M. G., *Introduction to the unified theory of electromagnetic machines*, Pitman, London (1971)

14 BRITISH ELECTRICITY INTERNATIONAL, *Modern Power Station Practice*, 12 vols, Turbines, Generators and Associated Plant, Vol. C, 3rd edition, Pergamon, Oxford (1992)

15 BROWN, J. G. (ed.), *Hydroelectric Engineering Practice*, Vol. 2 Chapters 7–9, Blackie and Son, Glasgow (1964)

16 JAIN, G. C., *Design, Operation and Testing of Synchronous Machines,* Asia Publishing House, London (1959)

17 LEWIS, W. A., *The Principles of Synchronous Machines*, Illinois Institute of Technology, Chicago, IL (1959)

18 SAY, M. G., *Performance and Design of a.c. Machines*, Pitman, London (1958)

19 KIMBARK, E. W., *Power System Stability*, Vol. 3 *Synchronous Machines*, Chapman and Hall, London (1956)

20 LANGSDORF, A. F., *Theory of a.c. machinery*, 2nd edition, McGraw Hill, New York (1955)

21 CONCORDIA, C., *Synchronous Machines: Theory and Performance*, General Electric Co., Schenectady, NY (1951)

22 LIWSCHITZ-GARIK, M., *Winding a.c. Machines*, van Nostrand, New York (1950) McMillan, London

23 LIWSCHITZ-GARIK, M. and WHIPPLE, C. C., *Electrical machinery*, Vol. II, *A.c. Machines,* van Nostrand, New York (1946)

24 VARIOUS AUTHORS, Transmission and distribution handbook, Westinghouse, Pittsburgh, PA

Technical papers, etc.

General reviews

25 HAMMONS, T. J. and GEDDES, A. G., Assessment of alternative energy sources for generation of electricity in the UK following privatization of the electricity supply industry, *IEEE T En. Conv*., **5**, No. 4, 609–615 (December 1990)

26 FOSTER, E. N. and PARKER, F. J., 'Hydroelectric machines', *IEE Proc*., *Part C*, **133**, No. 3, 126–136 (April 1986)

27 FRITSCH, T. A., 'Recent trends in large thermal and hydroelectric power production generators', *Trans. The South African Inst. of Elec. Eng*., 318–334 (November 1978)

28 SEONI, R. M. *et al*., 'Review of trends of large hydroelectric generating equipment', *Proc. IEE,* **123**, No. 10R (1976)

29 VICKERS, V. J., 'Recent trends in turbogenerators', *Proc. IEE*, **121**, No. 11R, 1273–1306 (November 1974)

30 ANSCOMBE, L. D., A.c. generators for hydro-electric stations', *Proc. IEE*, **110**, No. 7, 1223–1234 (July 1963)

Flux and e.m.f. waveforms; a.c. windings

31 LIWSCHITZ-GARIK, M., *Winding a.c. Machines*, van Nostrand, New York (1950) McMillan, London. KOSTENKO, M. and PETROVSKY, L., *Electrical Machines*, Vol. II, A.C. machines, 3rd edition, Mir Publishers, Moscow 1977 (translated from the Russian). CHALMERS, B. J. and WILLIAMSON A. C., *A.c. Machines: Electromagnetics and Design*, Research Studies Press, Taunton (1991). LIWSCHITZ-GARIK, M. and WHIPPLE, C. C., *Electrical Machinery*, Vol. II, A.c. Machines, van Nostrand, New York (1946). SAY, M. G., *Alternating Current Machines*, 5th edn, Pitman, London (1983). WALKER, J. H., *Large Synchronous Machines: Design, Manufacture and Operation*, Oxford Scientific Publications, Clarendon Press, Oxford (1981) (Paperback edn from University Microfilm International via White Swan House, High St., Godstone, Surrey, England.)

32 WALKER, J. H., 'Parasitic losses in synchronous machine damper windings', *J. IEE*, *P* II, **94**, 13–25 (1947)

33 KARMAKER, 'A.c. tooth ripple losses in slotted laminated machines with amortisseur windings', *IEEE PAS*, **101**, No. 5, 1122–1128 (May 1982)

34 STEPHEN, D. D., 'Evaluation of characteristics of a.c. stator windings'. *GEC J. Sci. Technol.* **40**, No. 1, 25–32 (1973)

35 BURBIDGE, R. F., 'A rapid method of analysing the m.m.f. wave of a single or polyphase winding', *IEE Monograph No. 2805* (January 1958)

36 LIWSCHITZ-GARIK, M., 'Distribution factors and pitch factors of the harmonics of a fractional-slot winding', *Trans. AIEE*, **62**, 664–666 (October 1943)

37 CHALMERS, B. J., 'A.c. machine windings with reduced harmonic content', *Proc. IEE*, **111**, No. 11, 1859–1863 (5 refs). (November 1964)

38 WALKER, J. H. and KERRUISH, N., 'Design of fractional slot windings', *Proc. IEE*, **105**, paper No. 26785, 428–440 (12 refs) (August 1958)

39 LIWSCHITZ-GARIK, M., 'Harmonics of the salient pole machine and their effects. Part 1 **MMF** harmonics produced by armature and damper windings', *Trans. AIEE*, PAS, **75**, 35–39 (1956)

40 STROMBERG, T., 'Alternator voltage waveshape with particular reference to the higher harmonics, *ASEA J.* 139–148 (1947)

41 CALVERT, J. F., 'Amplitudes of magneto-motive force harmonics for fractional-slot windings', *Trans. AIEE*, **57**, 777–785 (1938)

42 ANGST, G. and OLDENKAMP, J. L., 'Third harmonic voltage generation in salient pole synchronous machines', *Trans. AIEE*, PAS (June 1956)

43 WALKER, J. H., 'Slot ripples in alternator e.m.f. waves', *Proc. IEE*, Part II, **96**, Paper No. 777, 81–92 (17 refs) (1949), **97**, Part II, 45–46 (1950)

44 WIESEMAN, R. W., 'Graphical determination of magnetic fields', *Trans. AIEE*, February, 141–154 (1927)

45 GINSBERG, D., JOKL, A. L. and BLUM, L. M., 'Calculation of no load waveshape of salient-pole AC generators', *Trans. AIEE*, PAS, October, 974–980 (4 refs) (1953)

46 GINSBERG, G. D. and JOKL, A. L., 'Voltage harmonics of salient-pole generators under balanced 3-phase loads—I', *Trans. AIEE*, February 1960, 1573–1580 (6 refs) (1959)

47 GINSBURG, G. D. and JOKL, A. L., 'Voltage harmonics of salient-pole generators under balanced 3-phase loads—II', *Trans-AIEE*, PAS, August, 560–565 (7 refs) (1960)

Coils and insulation

Coil construction, materials and windings

48 CHALMERS B. J. (ed.), *Electrical Motor Handbook,* Butterworth, London (1988)

49 LIWSCHITZ-GARIK, M., *Winding a.c. Machines*, van Nostrand, New York (1950) McMillan, London.

50 KOSTENKO, M. and PETROVSKY, L., *Electrical Machines*, Vol. II, A.C. machines, 3rd edn, Mir Publishers, Moscow 1977 (translated from the Russian).

51 BENNINGTON and BRENNER, 'Transpositions in T.G. coilsides—short circuit at ends', *IEEE PAS*, **89**, No. 8, 1915–1921 (November–December 1970)

52 SUMMERS, 'Reduction of armature copper losses', *Trans. AIEE*, **46**, 101–111 (Febuary 1927)

53 MAINS, A. J. and McNAUGHTON, H. S., 'Design and manufacture of a.c. stator windings for hydrogenerators: a review of the 18 kV generator-motor units at Dinorwig power station', *Proc. 6 BEAMA Intl. Electrical Insulation Conf*., British Electrical and Allied Manufacturers Association, London (1990)

54 NEAL, J. E. and WHITMAN, A. G., 'The role of backing materials in mica-paper based insulation for h.v. rotating machines'. *IEEE Elec. Insulation Magazine*, **2**, No. 4, 30–34, July (1986) (Presented at 5th BEAMA Intl. Conf. on Electrical Insulation, May 1986)

55 SMITH, G. F., 'Mica Film'. *Proc. 5th BEAMA Conf.* (1986)

56 SCHULER, R., 'H.v. rotating machines: turn insulation for stator windings with form wound coils', *Proc. 2nd IEEE Intl. Conf. on properties and applications of Dielectric materials*, Beijing (1988)

57 NURSE, J. A. and KENNEDY A. G., 'Global vacuum impregnation of large high voltage stator windings', *Proc. of 6th BEAMA Intl. Conf. on Electrical Insulation*, British Electrical and Allied Manufacturers Association, London (1990)

58 NURSE, J. A., 'Resivac—an insulation system with extended life, *GEC Rev.* **2**, No. 2, 111–116 (1986)

59 HUTTER, W., LIPTAK, G. and SCHULER, R., 'Micadur—compact insulation system for rotating h.v. machines up to medium ratings—behaviour under extreme operating conditions', *Brown-Boveri Rev.*, **6/7**, 294–298 (1984)

60 McNAUGHTON, H. S. and NURSE, J. A., 'Vacuum-pressure impregnation and resin-rich insulation systems for high voltage industrial machines—a comparison', *IEE Intl. Conf. on Electrical Machines—Design and Application,* IEE, London (1982)

61 JONSSON, K., 'Micapact II coils for h.v. rotating machines', *ASEA J.*, **54**, No. 2, 27–35 (1981)

Insulation testing and evaluation

62 IEEE, 'Recommended practice for voltage-endurance testing of form wound bars and coils', *IEEE Standard* **1043** (1989)

63 IEEE, 'Proposed Test procedure for evaluation of systems of insulating materials for a.c. electric machinery employing form wound pre-insulated stator coils', *IEEE Standard 275*; these two were presented at the 1990 IEEE International Symposium on Electrical Insulation

64 IEEE, 'Recommended practice for insulation testing of large a.c. rotating machinery with high direct voltage', *ANSI/IEEE Standard 95* (67 refs) (1977)

65 REYNOLDS, P. H. and LESZCZYSWKI, S. A. 'Direct current insulation analysis—a new and better test method'. *IEEE PAS 104*, 7, 1746–1749 (1985)

66 MEYER, H. and WICHMANN, A., 'Experience and practice with standardised acceptance test procedures for windings of rotating machinery', *Proc. 16th Electrical/Electronics Insulation Conference*, Chicago, **IL**, 146–151 (1983)

67 BRANCATO, E. L., 'New diagnostics for rotating machinery (an EPRI report)', *IEEE Insulation Mag.*, 40–41 (January–February 1989)

68 YESHIDA, H. and UMEMOTO, K., 'Insulation diagnostics for rotating machines. *IEEE Trans. on Elec. Insulation*, **E1 21**, 1021–1025 (1986)

69 SIMONS, J. S. 'Diagnostic testing of h.v. machine insulation', *Proc. IEE*, *Part B*, **127**, No. 3, 139–154 (May 1980)

70 WICHMAN, A., 'Two decades of experience and progress in epoxy-mica insulation systems for large rotating machines'. *IEEE Trans. PAS*, **102**, 74–82 (1982)

71 DACIER, J. and GOFFAUX, R., 'Contribution to the overall and local characterisation of the condition of electrical ageing of h.v. insulation in large rotating machines'. *Proc. 3rd Intl. Conf. on Conduction and Breakdown in Solid Dielectrics*, *IEEE Catalogue No. 89* CH 2726–8 (July 1989)

72 KAKO Y. *et al.*, 'An analysis of multifactor ageing of mica-epoxy insulation systems by the infinite sequential stress method', *IEEE Trans. on Elec. Ins.*, **22**, 69–76 (1987)

73 SIMONI, L., 'An analysis of combined stress degradation of rotating machine insulation'. *IEEE Trans. on Elec. Ins.*, **19**, 364–367 (1984); see also 45–52

74 RENGARAJAN, S. *et al.*, 'Accelerated ageing of h.v. machine insulation under combined thermal and electrical stress', Annual report, Conf. on Elec. Ins. and Dielectric Phenomena, *IEEE Trans. on Elec. Ins.*, 129–136 (1983)

75 KIM, Y. J. and NELSON, J. K. 'Voltage dependence of corona signature from defect stator bar insulation during ageing', *Conf. report IEEE Trans. on Elec. Ins.*, 502–507 (1986)

76 VARIOUS AUTHORS, Five papers on testing and ageing of insulation. *CIGRE conference* on large h.v. systems 1976, Paper Nos: **15–00, 15–03, 15–05, 15–06** and the Minutes of Group 15. Pages 1–48 of the Proceedings

77 WICHMANN, A., 'Accelerated voltage endurance testing of micaceous insulation systems for large turbo-generators under combined stresses', *IEEE PAS*, **96**, 255–260 (1977)

78 WICHMANN, A. and GRUNEWALD, P., 'Statistical evaluation of accelerated voltage endurance tests on mica insulation for rotating electrical machines', *IEEE Trans. on Elec. Ins.*, **25**, 319–323 (1990)

79 PIERRAT, L., STEINLE, J. L. *et al.*, 'On load methods for dielectric diagnosis of large rotating machines', *CIGRE Paper*, **11–14** (1988)

80 KRECKE, M. and GOFFAUX, R., 'Attempt at estimating the residual life of the h.v. insulation of a.c. rotating machines', *CIGRE Paper,* **11–12** (1988)

81 GOFFAUX, R. *et al.*, 'A novel electrical methodology of diagnosis for the h.v. insulation of a.c. generators', *CIGRE Paper*, **11–12** (1986)

82 JONSSON, K. and RODOLFSSON, D., 'Diagnostic test of insulation. A test package to determine the condition of the generator stator winding insulation' *CIGRE Paper*, **11–11** (1986)

Determination of machine parameters

83 See Standards in section 28.21: IEC 34–4; BS 4999: Part 104; IEEE 115 and 115A.

84 KILGORE, L. A., 'Calculation of synchronous machine constants—reactances and time constants', *Trans. AIEE*, 1201–1214 (December 1931)

85 KOSTENKO, M. and PETROVSKY, L., *Electrical Machines,* Vol. II, A.C. machines, 3rd edition, Mir Publishers, Chapter 5, Moscow (1977) (translated from the Russian)

86 LAWRENSON, P., 'Calculation of machine endwinding inductances with special reference to turbogenerators', *Proc. IEE*, **117**, No. 6 (1970)

87 DE MELLO *et al.*, 'Derivation of synchronous machine stability parameters from pole slipping conditions. *IEEE T-PAS*, **101**, No. 9, 3394–3402 (September 1982)

88 BALDA, J. C., *et al.*, 'Measurement of synchronous machine parameters by a modified frequency response method', *IEEE T-EC*, **2**, No. 4, 646–651 (December 1981)

89 JACK, A. G., and BEDFORD, T. J., 'A study of the frequency response of turbine generators with particular reference to Nanticoke', *IEEE T-EC 2*, No. 3, 495–505 (19 refs) (September 1987)

90 KAMURA I., VIAROUGE, P. and DICKINSON, J., 'Direct estimation of the generalised equivalent circuits of synchronous machines from short-circuit oscillographs', *Proc. IEE*, *Part C*, **137–6**, 445–452 (28 refs) (November 1990)

91 CANAY, M., 'Identification and determination of synchronous machine parameters', *Brown-Boveri Rev.*, **6/7**, 299–304 (1984)

Turbogenerators

Design and construction

92 CREEK, F. R. L., 'The design of 985 MW 2 pole 3000 rpm turbine generators for Daya Bay nuclear P. S.', *GEC Rev.*, **4**, No. 3, 176–186 (1988)

93 MARLOW, B. A., 'The mechanical design of large turbo-generators 51st Parsons memorial lecture', *Proc. Inst. Mech. Eng*., **200** (1986)

94 GUILLARD, J. M. and DAMIRON, R., '300 MW Modular design generators', *Alsthom Rev*. No. 7, 19–30 (1987)

95 HASSE, H. and LARGIADER, H. G., 'Air cooled turbine generators in the 200 MVA class', *Brown-Boveri Rev*., No. 3 (1986)

96 HASSE, H. and LARGIADER, H. G., 'Design and operation on test bed of a 200 MVA air cooled turbine generator', *Proc. CIGRE*, **11–09** (1984)

97 VORONOWSKI, G. P. *et al*., 'Standard structural design solutions for turbogenerators', *Elektro teknika*, **46**, No. 1 (1975) (In English)

98 GLEBOV, I. A. *et al*., 'A 1200 MW 3000 rpm turbo-generator', *Elektrotektronika*, **49**, No. 3 (1978)

99 FEDOROV, V. F. *et al*., 'A fully water cooled 800 MW 3000 rpm turbogenerator', *Proc. CIGRE*, **11–11** (1984)

100 LAMBRECHT, D. and BERGER, H., 'Integrated endwinding ring support for water-cooled stator winding', *IEEE PAS*, **102**, No. 4 (April 1983)

101 KHAN, G. K. M. *et al*., 'Calculating electromagnetic forces on endwindings of large turbogenerators', *IEEE T-EC,* 661–670 (December 1989)

102 MECROW, B. C., JACK, A. G. and CROSS, C. S., 'Electromagnetic design of t.g. stator end regions' *Proc. IEE 136-C*, No. 6, 361–372 (38 refs) (November 1989)

103 SINGLETON *et al*., 'Axial magnetic flux in synchronous machines', *IEEE PAS*, **100–3**, 1226–1233 (March 81)

104 KAHN, G. K. M. *et al*., 'An integrated approach to the calculation of losses and temperature in the end region of large turbogenerators', *IEEE T-EC*, **5–1** 183–194 (23 refs) (March 1990)

105 COULSON *et al*., 'Transient negative sequence capability of turbine generators: a rational assessment', *CIGRE Proc*., **11–02** (1980)

106 CIGRE Study Committee, 'CIGRE repot—A summary of replies to a questionnaire on the properties and design of t.g. rotor endrings', *Electra*, **80** (January 1982); *Electra*, **17** (March 1988)

107 EPRI *Workshop Proceedings*, 'Retaining rings for electric generators' Pub. No. EL 3209 (August 1983)

108 VSG Publication on 18Mn 18Cr retaining rings, from VSG; Altendorferstrasse 104, D-4300, Essen, Post 10225, Germany

109 McINTYRE, NESBITT and RILEY, 'Improved steels for non-magnetic generator endrings', *Conf. Proc.: Materials Development in Turbo-machinery Design*, 12–14, September 1988, Churchill College Cambridge, available from Institute of Metals, London

Operation, monitoring and testing

110 IEEE, 'Guide for operation and maintenance turbine type generators' (64 pp. 88 refs) *IEEE Standard 67–1990*

111 MAMIKONIANTS, L. G., for Study Committee 11, 'Draft guidelines on some of the synchronous generator abnormal operation conditions', *Proc. CIGRE*, **11–13** (1980)

112 HUTTNER, H. *et al*., 'Some aspects on diagnosis methods and operational monitoring for large a.c. generators', *Proc. CIGRE*, **11–01** (1986)

113 SANDHU, S. *et al*., 'Diagnostic methods for testing the electrical and mechanical integrity of stator end-windings of large turbogenerators', *Proc. CIGRE* **11–03** (1986)

114 CARLIER *et al*., 'Investigations into the mechanical behaviour of turbogenerator stator windings during faults', *Proc. CIGRE*, **11–14** (1980)

115 JACKSON, R. J. *et al*., 'Generator rotor monitoring in the UK', *Proc. CIGRE*, **11–04** (1986)

116 GRANGER, B. and LEHUEN, C., '*In situ* ultrasonic inspection of t.g. rotor endbells', *Proc. CIGRE*, **11–10** (1986)

117 VERMA, S. P. *et al*., 'The problems and failures caused by shaft potentials and bearing-currents in turbogenerators: methods of prevention', *Proc. CIGRE*, **11–10** (1980)

118 CANDELORI, C. *et al*., 'Shaft voltages in large t.g. with static excitation. Experimental investigations and protective devices', *Proc. CIGRE*, **11–04** (1988)

119 JOHO, R. *et al*., 'Shaft voltages in turbosets: a new grounding design to improve reliability of the bearings', *Proc. CIGRE*, **11–10** (1988)

120 HEARD, J. G., 'Summary report on large turbine generator maintenance practices', *CIGRE Electra*, **11** (March 1988)

121 EMERY, F. T. and HARROLD, R. T., 'On line incipient arc detection in large t.g. stator windings', *T-PAS*, **99–6** 2232–2238 (November/December 80)

122 SCHULER, R. H. and LIPTAK, G., 'A new method for high-voltage testing of field windings (interturn insulation)', *CIGRE Proc*., **11–04** (1980)

Cooling

123 SCHMITT, WILLYOUNG and WINCHESTER, 'Diagonal flow ventilation of gap pickup rotors', *IEEE PAS* (Febuary 1963)

124 GOTT, B. E. B., KAMINSKI, C. A. and SHORTRAND, A. C., 'Experience and recent development with gas directly cooled rotors for large steam turbine generators', *IEEE PAS*, **103**, No. 10, 2974–2981 (October 84)

125 CSILLAG, I. K., 'Studies in cooling of gap pick up turbine generators with cross flow ventilation', *T-PAS,* 871–882 (May/June 1979)

126 GRUNENWALD, J. *et al*., 'Rotor water cooling in turbogenerators', *Proc. CIGRE*, **11-07** (1980)

127 MANG, Y., 'Twenty-one year development in turbine generators with water-cooled stator and rotor windings', *IEEE PAS,* **101**, No. 3 (March 1982)

Shaft fatigue: vibration

128 IEEE SUBSYNCHRONOUS RESONANCE WORKING GROUP, 'Comparison of SSR calculations and test results', *IEEE T-PWRS*, 336–344 (Febuary 89)

129 IEEE SUBSYNCHRONOUS RESONANCE WORKING GROUP, 'Bibliography', *T-PAS*, **95–1**, 216–218 (January/Febuary 1976); 'First Supplement', *T-PAS*, **98–6**, 1872–1875 (November/December 1979); 'Second Supplement', *T-PAS*, **104**, 321–32 (February 1985)

130 JOOS, GEZA *et al*., 'Torsional interactions between synchronous generators and long transmission lines: supersynchronous and subsynchronous resonances', *IEEE T-PWRS*, 17–24 (February 87)

131 LAMBRECHT, D. *et al*., 'Evaluation of the torsional impact of accumulated failure combinations on

turbine generator shafts as a basis of design guidelines', *Proc. CIGRE*, **11–06** (1984)

132 RUSCHE, P. A., 'TG shaft stresses due to network disturbances: a bibliography with extracts', *PAS 99–6*, 2146–2152 (November/December 80)

133 DUNLOP, R. D. *et al.*, 'Torsional oscillations and fatigue of steam t.g. shafts caused by system disturbances and switching events', *Proc. CIGRE*, **11–06** (1980)

134 CUDWORTH, C. J. and SMITH, J. R., 'Steam turbine generator shaft torque transients: a comparison of simulated and test results', *Proc. IEE*, **137-C**, 327–334 (September 1990)

135 MASRUR, M. A., *et al.*, 'Studies on asynchronous operation of synchronous machines and related shaft torsional stresses', *IEE Proc-C*, **138–1**, 47–56 (48 refs) (January 1991)

136 HEATHCOTE, C., PETTY, D. J. and SMITH, R. J., 'Lifetime capability of turbogenerators to withstand vibration and other cyclic effects', Proc. *CIGRE*, **11–09** (1988)

137 GLEBOV, I. A. *et al.*, 'Vibratory behaviour study and control of large turbo and hydro-generators', *Proc. CIGRE,* **11–11** (1988)

Hydro-electric generators

138 HORN, '615 MVA generators for Grand Coulee: electrical and mechanical design features', *PAS*, **94**, 2015–2022 (November/December 1975)

139 MOORE, 'Large hydro-generators at Grand Coulee 3—design experience', *PAS*, **102**, 3265–3270 (October 1983)

140 KERKMAN, 'Pumped storage plants', *PAS*, **22**: 'Machine design and performance', 1828–1837; 'System analysis', 1838–1844 (September/October 1980)

141 VARIOUS AUTHORS, Bulb type generators: 'St. Onge, Rock Island 2', *PAS*, **96**, 1690–1696 (Sept./October 1977); 'McGilvery, Manitoba', *PAS*, **99**, 990–997 (May/June 1980); 'St. Onge, Columbia River', *PAS*, **101**, 1313–1321 (June 1982); 'Ruelle, Rock Island', *PAS*, 101, 639–643 (March 1982); 'Paine', *PAS*, 103, 2405–2409 (September 1984)

142 KERMIT, P., 'Design features of the Helms pumped storage project', *IEEE T-EC*, **9–15** (gives operating experience) (March 1989)

143 BEVC, F. P. and MEEHAN, R. J., 'Generator-motors for PG and E Helms pumped-storage project', *IEEE* PAS, **99–6**, 2021–2030 (November/December 1980)

144 Hydraulic Generators and Synchronous Compensators', *CIGRE—Proceedings of the Rio de Janeiro Symposium* (November 1983)

145 VARIOUS AUTHORS, Special issue on hydro-electric power, *IEE Proc-C*, No. 3 (April 1986)

146 HLAVAC, J. and GLEICH, K., 'Design and proving tests of generator motors of 121 MVA 136.5 rpm with asynchronous starting', *Proc. CIGRE*, **11–03** (1980)

147 KRANZ, R. D. for STUDY COMMITTEE 11, 'Selected aspects of salient pole machines mechanical problems', *Proc. CIGRE*, **11–15** (1980)

148 TALAS, P. *et al.*, 'On-line monitoring of airgap of hydro-electric generator using optical triangulation', *IEEE T-EC*, **526–533** (December 1987)

149 BAROZZI *et al.*, 'Laminated segmental rims', *PAS 95* (July/August 1976): 'Elastic behaviour', 1045–1053; 'Design criteria', 1054–1061

150 XU SHIZHANG, 'Magnetic vibration of hydro-generators stator core due to rotor eccentricity, rotor non-circularity and negative sequence current', *CIGRE Electria*, **86** 77–88 (January 1983)

151 TOOM, P. O. *et al.*, 'Application of precision airgap monitor for analysis of generator problems' *Proc. CIGRE*, **11–02** (1986)

152 PAIXAO, R. *et al.*, 'Research analysis of the effects of switching operations on hydro units. A diagnosis of unit life performance', *Proc. CIGRE*, **11–05** (1986)

153 MISTRY, D. K. *et al.*, 'Salient design features of brushless hydro-generators for mini/micro hydro-electric schemes', *Seminar Elroma 88*, Indian Electrical and Electronic Manufacturers' Association, Bombay (11 pages) (January 88)

154 SMITH, J. R. *et al.*, 'Prediction of forces on the retaining structure hydrogenerators during severe disturbances', *Electric Power Systems Res.* (Switzerland), **14**, No. 1, 1–9 (February 1988)

155 OHISHI *et al.*, 'Radial magnetic pull in salient pole machines with eccentric rotors', *IEEE T-EC*, **2**, No. 3, 439 113 (September 1987)

156 IEEE WORKING GROUP REPORT, 'Hydro generator thermoset insulation systems—premature failures bibliography', *T-PAS*, 3284–3303 (July 1981)

Induction generators

157 LEITHEAD, W. E. *et al.*, 'The role and objectives of control for wind turbines', *Proc. IEE*, Pan C, **38**, No. 2, 135–148 (22 refs) (March 91)

158 MALIK, N. H. and AL-BAHRANI, A. H., 'Influence of the terminal capacitor on the performance characteristics of a self-excited induction generator', *IEE Proc-C*, **137**, No. 2, 168–173 (13 refs) (March 1990)

159 MURTHY, S. S. *et al.*, 'Grid connected induction generators driven by mini-hydro or wind turbines, operational behaviour', *IEEE 7-En Conv.*, 1–7 (March 1990)

160 JABRI, A. K. and ALOLAH, A. 1., 'Limits on the performance of the 3-phase self excited induction generator', *IEEE T-En Conv.*, **5**, No. 2, 350–356 (5 refs) (June 1990)

161 DE MELLO *et al.*, 'Application of induction generator in power systems', *IEEE PAS*, **101–9**, 3385–3393 (September 1982)

162 DEMOULIAS, C. S. *et al.*, 'Transient behaviour and self-excitation of wind driven induction generator after disconnection from the power grid', *IEEE T-EC*, 272–278 (June 1990)

163 WOODWARD, J. L. and BHATTACHARYA 'Induction generators in micro hydro-electric systems', *Seminar Elroma 88*, Indian Electrical and Electronic Manufacturers' Association, Bombay (11 pages) (January 1988) Bombay

164 GRANTHAM C. *et al.*, 'Steady state and transient analysis of induction generators', *Proc. IEE*, Part B. No. 2, 61–68 (18 refs) (March 1989)

Excitation and stability

165 DILLMAN *et al.*, 'A high initial response brushless excitation system', *IEEE PAS 90*, 2089–2094 (September/October 1971)

166 COTZAS, G.M., HESSE, H.M. and LANE, L.J., 'Electrical design and steady state performance of

Generrex* (*Trade Mark of GE) excitation system', *T-PAS* **98–6** 2251–2261 (November/December 1979)

167 CHABOT, E. and TRAN THANH TAM, 'New developments in brushless bearingless integral hydrogen cooled excitation generator for 3000 rpm unit', *Proc. CIGRE*, **11–13** (1984)

168 HURLEY, J. D. and BALDWIN, M. S., 'High response excitation systems on turbo-generators: a stability assessment', *LEEE T-PAS*, **101**, No. 11. 4211–4221 (November 1982)

169 HOGG, B. W. *et al*., 'The design and development of a self-tuning voltage regulator for a turbine generator', *Proc. CIGRE*, **11–08** (1988)

170 HERZOG, H. and BAUMBERGER, H., 'Digital control of generator excitation—Unitrol', *Brown-Boveri Rev*., **1**, 27 (1990)

171 PENEDER, F. and BERTSCHI, R., 'Static excitation systems with and without a compounding ancillary', *Brown-Boveri Rev*., **7**, 343–348 (1985)

172 DINELEY, J. L., 'Tutorial on Power Systems stability, Part 1' *IEE Power Eng. J*., **5** (January 1991); Part 2 (July 1991), Part 3 (probably January 1992)

173 HURLEY, T. and KEAY, W. F., 'Power System Stabilisation via excitation control', *81EHO 175–0-PWR*, Chap. 2, 'Overview of Power System Stability Concepts'; IEE Tutorial Course (1981)

Useful sources of references are:

(1) Electrical and Electronic Abstracts, published monthly by the IEE as part of its Inspection Service.
(2) Cumulative Index of *IEEE Transactions on Power Apparatus and Systems*, 1975–1984, and for 1985.
(3) Combined Index for *IEEE Transactions on Power Delivery, Power Systems, Energy Conversion*, published annually from 1986 onwards.

29

Batteries

M A Laughton BASc, PhD, DSc(Eng), FREng, CEng, FIEE
Formerly of Queen Mary & Westfield College, University of London
(Section 29.5)

M Barak MSc, DPhil, FRSC, CChem, CEng, FIEE, FRSA
Formerly Consulting Chemist and Engineer
(Sections 29.1 to 29.4, 29.6 to 29.8)

D Inman PhD, DSc, DHon Causa (Grenoble), DIC, ARCS, MIMM, FRSC, CChem, CEng
Department of Materials, Imperial College, London

Contents

29.1 Introduction 29/3

29.2 Cells and batteries 29/3
  29.2.1 Definitions 29/3
  29.2.2 Principles 29/3
  29.2.3 Redox process 29/3

29.3 Primary cells 29/4
  29.3.1 Leclanché (zinc–carbon) cell 29/4
  29.3.2 Standard cells 29/5
  29.3.3 Alkaline cells 29/5
  29.3.4 Water activated cells 29/6
  29.3.5 Acid cells 29/7

29.4 Secondary cells and batteries 29/7
  29.4.1 Lead/acid cells 29/7
  29.4.2 Nickel/cadmium and nickel/iron alkaline cells 29/9
  29.4.3 Silver/zinc alkaline cells 29/11
  29.4.4 Secondary battery technology 29/11
  29.4.5 Lithium cells 29/13
  29.4.6 Sodium/sulphur cells 29/13

29.5 Battery applications 29/13
  29.5.1 Stationary or standby power batteries 29/13
  29.5.2 Self-contained power supplies 29/14
  29.5.3 Traction batteries 29/14

29.6 Anodising 29/14
  29.6.1 Process 29/14
  29.6.2 Vats 29/16
  29.6.3 Workpieces 29/16

29.7 Electrodeposition 29/16
  29.7.1 Electroplating 29/16

29.8 Hydrogen and oxygen electrolysis 29/17
  29.8.1 Process 29/17
  29.8.2 Electrolysers 29/18
  29.8.3 Gas purity 29/18
  29.8.4 Plant arrangement 29/18

29.1 Introduction

Electrochemical science is an interdisciplinary subject, which is unique both in its concepts (the chemical interactions between matter and electrons particularly across interfaces) and general in its applications. Electrochemical reactors use electrical energy to isolate elements from their compounds, as in the electro winning of metals (copper, silver, aluminium) or chlorine, or the production of hydrogen and/or oxygen by the electrolysis of water. On the other hand electrochemical power sources, batteries and fuel cells store chemical energy, which can be instantly converted into d.c. electrical energy. Corrosion, an electrochemical process, wastes both energy and materials.

29.2 Cells and batteries

29.2.1 Definitions

The term 'battery' means an assembly of voltaic primary or secondary cells. Batteries of secondary cells are known also as storage batteries or accumulators.

- A primary cell is used once until it is discharged and then discarded.
- A secondary cell needs to be charged after it is made before use. Once discharged the cell can be recharged and used again.

29.2.1.1 Terminology

Open-circuit voltage is the voltage at the terminal of the cell when no current flows, i.e. without an external loading circuit connected. It depends on the operating history of the cell, but cells are usually designed to have a relatively flat discharge curve until the cell is almost totally discharged after which the voltage drops quickly.

Cell capacity is the amount of energy, usually stated in ampere-hours that the cell can provide without the terminal voltage falling below a specified value. This is also stated as its 'C' rate which is the rate at which a fully charged cell would be discharged in one hour, i.e. a 5 Ah cell would have a C rating of 5 A.

Depth of discharge is the percentage of the cell's capacity which has been discharged.

Shelf life is a measure of the ability of the cell to retain its charge under the storage conditions encountered.

Cycle life is the number of charge/discharge cycles a secondary cell can tolerate before performance failure.

Charge acceptance is the proportion of the charge input which the cell can give out again without the voltage falling below a specified value, i.e. is a measure of the ability of a cell to accept charge.

Charge voltage is the voltage developed across a secondary cell when it is under charge. This voltage can be up to 50% higher than the rated discharge voltage of the cell. It increases with the charge rate and with low temperatures.

29.2.2 Principles

In both primary and secondary batteries the individual cells consist of a positive and negative electrode immersed in an ion-conducting medium called the electrolyte and generally separated by a porous non-conducting diaphragm, called the separator. The electrodes, which must be electrically conducting, may consist of a single rod or plate, or a number of these welded or bolted together in parallel. In some cells (for example, the conventional primary 'dry' cell) the outer metal container may constitute one of the electrodes. The two electrodes have different electrical potentials when immersed in the common electrolyte and the difference between these potentials represents the e.m.f., or open-circuit voltage, of the cell.

In both primary and secondary cells the electrical energy released during discharge is derived from the chemical energy liberated as a result of the chemical reactions taking place in the cell. These reactions involve charged particles in the electrolyte, known as 'ions'. Ions, if positively charged, have a deficiency of electrons, and if negatively charged, carry an excess of electrons. As indicated below, certain ions tend to react with the electrode in their vicinity, causing a transfer of electrons from ions to electrode, or vice versa. If this reaction is allowed to proceed (for example, by closing the external circuit to which the battery is connected), the transfer of electrons from one electrode to the other gives rise to an electric current flowing in the external circuit conventionally from the positive to the negative electrode: the flow of electrons is in the opposite direction. Thus, an *anodic* reaction involving the release of electrons occurs at the *negative* electrode, and a *cathodic* reaction involving the capture of electrons at the *positive* terminal of the battery. In electrochemical reactors the reverse is true, i.e. *anodic* reaction at the *positive* and *cathodic* at the *negative* electrode.

29.2.3 Redox process

The chemical reactions at the electrodes are either 'reduction' or 'oxidation', i.e. 'redox' processes. The basic feature of such reactions is the gain or loss, respectively, of one or more electrons, e.g.

reduction/cathodic reaction $2H^+ + 2e^{-\leftarrow} \rightarrow H_2$
$Cu^{2+} + 2e^{-\leftarrow} \rightarrow Cu$
oxidation/anodic reaction $Zn \rightarrow Zn^{2+} + 2e^{-\leftarrow}$
$Pb \rightarrow Pb^{2+} + 2e^{-\leftarrow}$

Here *e* represents the electron. H, Cu, Zn and Pb represent atoms, or (when charged) ions of hydrogen, copper, zinc and lead. The sign + indicates a deficiency of one electron; 2 + indicates a deficiency of two electrons.

The gaseous hydrogen formed by the first reduction process, if allowed to accumulate, would rapidly polarise the electrode, and the electrochemical reaction would virtually cease. To overcome this, the positive electrodes of many kinds of batteries are selected from substances which readily undergo a depolarising reaction with hydrogen. Typical examples are manganese dioxide (MnO_2), used in the primary dry battery, and lead dioxide (PbO_2), used in the lead/acid storage battery. Negative electrodes must be readily oxidisable, and for these active metals such as zinc, lead and cadmium are generally chosen.

Electrode potentials and redox reactions are not confined to metals, but include such elements as hydrogen, oxygen, chlorine or fluorine, either in the gaseous form or in combination with some metal in the form of an inorganic salt.

Electrode potentials are generally expressed with respect to hydrogen, which for standard conditions is assigned a potential of zero. It is an advantage to have at the two electrodes redox reactions widely spaced on the potential scale to give the highest cell e.m.f. But the choice is restricted by other factors—in particular, the type of electrolyte. Formerly electrolytes were aqueous solutions of salts, acids or bases. Organic electrolytes (e.g. propylene carbonate containing ionic conductors) and molten salts (e.g. LiCl–KCl eutectic, which melts at 352°C) are now being used, so that more active 'anodes' such as lithium or its alloys, which react with aqueous solutions, can be incorporated.

29.3 Primary cells

Primary cells differ from secondary cells in that the electrochemical reactions are not reversible, or, if so, only to a very limited extent. This may be primarily due to physical factors such as a loss of electrical contact by the chemical products of the discharge. In the case of the primary conventional 'dry' cell, for example, failure often occurs through excessive corrosion of the zinc can which forms one of the electrodes. In secondary cells or accumulators the reactions are readily reversible. The original chemical compounds can be re-formed by passing a direct current through the cell in the reverse direction, and accumulators can generally be submitted to many cycles of discharge and charge in this way.

The simplest primary cell, and one of the earliest (1836), is the *Daniell* cell, named after its inventor. Since many of the principles involved are common to other systems, this cell will be described in some detail. The electrodes are copper and zinc, and the electrolyte is a solution containing sulphate ions. The copper electrode is immersed in a solution of copper sulphate held in an inner porous pot, and the zinc electrode is held in a solution of zinc sulphate or dilute sulphuric acid in the outer glass vessel. The cell is shown diagrammatically in *Figure 29.1.*

The chemical reactions on discharge can be represented as follows:

positive electrode: $Cu^{2+} + 2e^- \rightarrow Cu$
potential $E_+ = 0.34\,V$
negative electrode: $Zn \rightarrow Zn^{2+} = 2e^-$
potential $E_- = -0.76\,V$
open-circuit e.m.f. $E = E_+ - E_- = 0.34 - (-0.76)$
$= 1.10\,V$

Early forms of this cell were made with both electrodes immersed in dilute sulphuric acid, and had somewhat different reactions: the main reaction at the negative electrode was (as above) the formation of positively charged Zn ions, but in the presence of the acid (H_2SO_4) hydrogen was also produced, by a corrosion reaction. At the positive electrode the reaction was $2H^+ + 2e^- \rightarrow H_2$: and as the discharge reaction proceeded, hydrogen gas accumulated on the electrode surface, polarising the cell and depressing the terminal voltage. Other factors in the voltage 'fall-off' are (a) overvoltages arising from the kinetic limitations of the electrode reactions themselves, (b) the slow diffusive mass transfer of the ions up to the electrode surfaces, and (c) the ohmic resistance of the electrolyte. Substituting sulphuric acid by copper sulphate ensured that copper and not hydrogen ions were discharged at the positive electrode, eliminating the counter-e.m.f. By placing the zinc electrode in its own zinc sulphate electrolyte, the evolution of hydrogen was eliminated; but this made it necessary to enclose the copper electrode and its electrolyte in a porous pot, increasing the ohmic resistance of the cell.

29.3.1 Leclanché (zinc–carbon) cell

In its 'dry' form, the Leclanché cell is the most common. A zinc plate or container forms the negative, a carbon plate or rod the positive electrode. The electrolyte is aqueous ammonium chloride. Through ionisation of the water, hydrogen and hydroxyl ions are produced ($H^2O \rightarrow H^+ + (OH)^-$). During the discharge the zinc electrode is oxidised in the reaction

$$Zn + 2(OH)^- \rightarrow Zn(OH)_2 + 2e^-$$

and hydrogen migrates to the carbon. Polarisation is avoided by surrounding the carbon by manganese dioxide, which participates in the redox reaction

$$2MnO_2 = 2H^+ + 2e^- \rightarrow Mn_2O_3 + H_2O$$

and the complete cell reaction, in which the electrolyte is also involved is

$$2MnO_2 + 2NH_4Cl + Zn \rightarrow 2MnOOH + Zn(NH_3)_2Cl_2$$

corresponding to an open-circuit e.m.f. of about 1.5 V. The presence of zinc and ammonium chlorides in the electrolyte keeps the acidity at the right level and helps to reduce polarisation of the zinc electrode by the flocculent jelly-like zinc hydroxide that would otherwise coat the surface. In practice, the cells are not strictly 'dry', as the electrolyte is a thick paste. The cell (*Figure 29.2*) comprises: (a) a central rod forming the positive terminal; (b) a depolariser of manganese dioxide (mixed with graphite or a highly active form of carbon black known as acetylene black to improve its

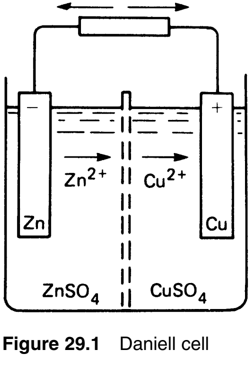


Figure 29.1 Daniell cell

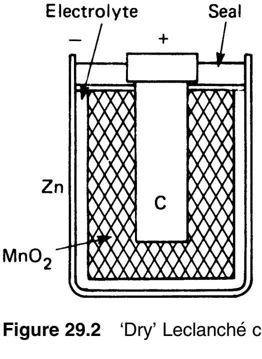


Figure 29.2 'Dry' Leclanché cell

conductivity) that is liberally moistened by the electrolyte; (c) a thin layer of jellified electrolyte; (d) a hard-drawn zinc outer can forming the negative terminal; and (e) a seal, such as a card disc covered by a layer of bitumastic compound. Recent improvements, particularly in high rate performance, have been achieved by the replacement of the bulk of the ammonium chloride in the electrolyte with zinc chloride.

29.3.1.1 Performance

In many applications the use is intermittent. Polarisation develops as the current flows; during rest periods the cell recovers. *Figure 29.3* shows typical curves for a U2 cell discharging continuously through a lamp, the resistance of which varied from 1.32 to 1.03 Ω during the discharge. The cell dimensions are 32 mm diameter, 57 mm height and 0.1 kg mass. Cell behaviour in store and in transit ('shelf life') is lengthened by adding a soluble mercury salt to the electrolyte, which lightly amalgamates the active surface of the zinc container. Shelf life is reduced with high, and extended with low, ambient temperatures.

29.3.1.2 Flat and layer cells

Flat and layer cells have been developed for transistorised electronic equipments. In a duplex type the positive and the negative electrodes are placed on opposite sides of an electrically conducting diaphragm impermeable to the electrolyte. In the layer-built battery the zinc plate serves as the conducting diaphragm as well as the negative plate. On the opposite side this is first coated with a thin adherent layer of carbon mixed with a plastic resin, which functions much as the carbon rod in the cylindrical cell. Against this carbon layer is placed a moulded cake of the usual black depolariser mix and next to it a layer of absorbent material, such as filter paper, impregnated with the electrolyte. Another duplex electrode is placed on top of the first, and in this way a multicell battery of any desired voltage can be built up.

Sealing presents a particular difficulty with layer-built batteries. Obviously, there must be no electrical contact between the electrolytes of neighbouring cells, as this would permit leakage current. Also, some provision must be made for the release of adventitious gas from each cell. One means of sealing is to separate the adjoining duplex electrodes with annular spacing pieces and to cover the edges of the whole battery in wax or a suitable cold-setting plastic resin. Another way is to encapsulate the unit in a tightly stretched plastic stocking which is shrunk into position by gentle heat after application. Yet another method is to enclose the edges of each duplex diaphragm in an annular envelope of rubber or a plastic material, shrunk into position by heat. To reduce the risk of leakage, such envelopes may be cemented to the diaphragm and the assembled unit sealed together or firmly wrapped with plastic tape. Thin strips of rubber or suitable plastic materials in this form are sufficiently permeable to hydrogen to allow the release of gas, and the construction allows for sufficient expansion to release any internal pressures that may develop, without permitting significant leakage of electrolyte.

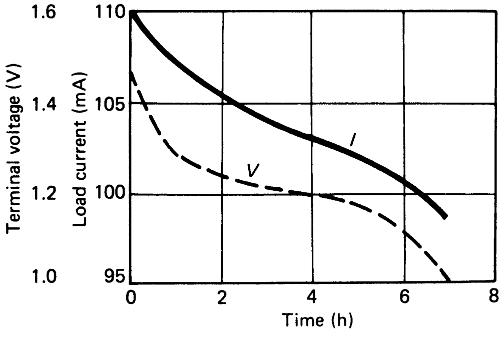


Figure 29.3 Discharge of a U2 cell

29.3.2 Standard cells

The Weston standard cell (*Figure 29.4*) has a cadmium/cadmium sulphate negative electrode, an electrolyte of cadmium sulphate and a mercury/mercurous sulphate positive electrode. To give added stability, the cadmium is amalgamated with mercury and the cell assembled in an H-shaped glass vessel with platinum leads to the terminals. The cadmium and mercury sulphates are usually prepared as thick pastes by digesting fine crystals of the salts with cadmium sulphate solution. Small quantities of sulphuric acid are added to reduce any tendency of the salts to hydrolyse.

When a saturated solution is used for the electrolyte, the cell is termed 'normal' and has an e.m.f. of 1.01862 V at 20°C (for e.m.f.s at other temperatures see reference 2).

29.3.3 Alkaline cells

Three alkaline cells use zinc for the negative electrode.

29.3.3.1 Copper oxide/zinc

The positive electrode or depolariser is CuO, the electrolyte a solution of caustic soda of relative density about 1.2. The cell was formerly common in railway signalling, being cheap to make and able to deliver intermittent currents over considerable periods. The open-circuit e.m.f. is about 1.0 V, and on load 0.5–0.7 V. The reaction at the positive electrode involves the reduction of the copper oxide, while at the negative the zinc is oxidised:

positive: $\mathrm{CuO} \rightarrow \mathrm{Cu(OH)_2} \rightarrow \mathrm{Cu^{2+}} + 2(\mathrm{OH})^{-\leftarrow}$
and $\mathrm{Cu^{2+}} + 2e^{-\leftarrow} \rightarrow \mathrm{Cu}$
negative: $\mathrm{Zn} + 2(\mathrm{OH})^{-} \rightarrow \mathrm{Zn(OH)_2} + 2e^{-\leftarrow}$

29.3.3.2 Silver oxide/zinc

The silver oxide/zinc cell has military and other special fields of application. With appropriate modifications this couple will also function as a reversible cell, at any rate for a limited number of cycles. Batteries of cells have outputs four to five times those of any other system for the same weight and

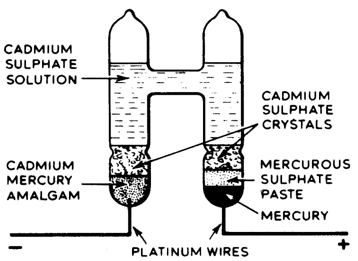


Figure 29.4 Weston standard cell

volume, and they are particularly in demand where very high currents and a low weight and volume are required, as in torpedo batteries, guided missiles and space satellites. In the primary form the batteries may be stored dry for long periods and brought into action by priming with electrolyte either by some pressure device or simply under the influence of gravity. To reduce polarisation, the electrodes are generally made porous. In one method of manufacture the positive electrode is made by sintering silver powder, the porous compact then being anodised to convert it to silver peroxide (Ag_2O_2). The zinc electrode is made from a paste of zinc oxide and caustic potash solution which is pressed into a screen of a suitable metal, such as silver or silver-plated copper. This is then reduced electrolytically to spongy zinc in dilute alkali solution. The electrolyte is caustic potash solution of about 30% concentration. A thin separator made of cellophane or paper may be used to prevent the plates from coming into contact and causing a short-circuit.

The discharge reactions at the electrodes are analogous to those already described for the copper oxide/zinc cell, but the silver peroxide passes through an intermediate stage to silver oxide (Ag_2O) before being reduced completely to metallic silver. The two separate reduction stages each have their own characteristic potential, so the discharge voltage curve of the cell with freshly prepared electrodes has two plateaux, a short one at about 1.80 V and a much longer one at about 1.50 V. The second portion of the discharge at about 1.50 V per cell or less, depending on the rate, remains remarkably steady until the active materials are exhausted, when the cell voltage falls sharply.

29.3.3.3 Mercury oxide/zinc

The electrolyte is caustic potash, but the elements are merely moistened. Known as the Ruben–Mallory cell, the mercury oxide/zinc cell has very low standing loss and gives steady voltages for long periods on low discharge rates. For such applications as hearing aids it is made in the form of a small metal button; as batteries of 6–8 V the applications are to portable transistorised radio receivers and 'walkie-talkie' sets. Mercuric oxide (HgO), mixed with 5–10% graphite to increase conductivity, acts as the positive depolariser. The mixture is compressed into a pellet, and in the button form this is placed in contact with the steel cup that makes one half of the button. A pellet of compressed zinc powder with a small amount of mercury is in contact with the other metal half of the button, which may be of copper. The separator may be a layer of absorbent paper saturated with the caustic potash solution. The outer surfaces of the two metals are insulated from each other by a plastic grommet, and form the terminals. Such sealed buttons have a capacity of a few milliampere-hours. In other forms the zinc electrode may be a coiled strip or perforated sheet, and the electrolyte is a jelly with gelling agents such as carboxymethylcellulose. The reactions are similar to those of other alkaline couples:

positive: $\mathrm{HgO} + \mathrm{H_2O} + 2e^- \rightarrow \mathrm{Hg} + 2(\mathrm{OH})^{-\leftarrow}$

negative: $\mathrm{Zn} + 2(\mathrm{OH})^- \rightarrow \mathrm{Zn(OH)_2} + 2e^{-\leftarrow}$

The open-circuit cell e.m.f. is 1.35 V, and the terminal voltage on load is 1.25–1.0 V. Prolonged low-rate discharges are obtainable. Standing losses are very low.

29.3.3.4 Air depolarised

These systems are essentially hybrids between true batteries which contain both electrode materials within themselves and fuel cells (see Section 29.5) which contain neither. In principle, they exhibit very high energy densities. In practice, however, various problems have been encountered, such as polarisation of the air cathode.

Zinc–air batteries (see below) have been the most successful, but aluminium–air, iron–air and lithium–air systems all have their protagonists. Lithium must be employed in strongly alkaline solutions to avoid reactions with the aqueous cell electrolyte. Aluminium–air systems are particularly attractive, at least in principle. In this case it is envisaged that the aluminate solutions which are produced as the batteries are discharged would be recycled to the aluminium smelters and 'fuel' (e.g. aluminium foil) would, if the batteries were used for traction, be obtained at suitable aluminium filling stations(!) during a journey to replace that used in the cells. The aluminate solutions could be tapped off at the same time.

Analogous to the carbon/zinc Leclanché cell, the air depolarised cell uses atmospheric oxygen as depolariser for the porous carbon electrode, which is made by roasting finely divided carbon and charcoal with a binder and then wet-proofing (e.g. by paraffin wax) to render the interior hydrophobic. The carbon electrode is only partially immersed in the electrolyte, and oxygen is readily adsorbed from the air. The negative electrode, of large plates of zinc lightly amalgamated with mercury, is completely immersed. A cell generally consists of two zinc plates flanking one carbon electrode, with capacities up to several hundred ampere-hours.

One range of cells has as electrolyte a solution of caustic soda. The essential reaction at the positive (oxygen) electrode is the reduction of oxygen with the production of $(\mathrm{OH})^{-\leftarrow}$ions, and at the negative (zinc) electrode these contribute to the usual oxidation reaction, as in the copper oxide/zinc cell. The open-circuit e.m.f. is about 1.45 V; the terminal voltage on load for small currents is about 1.3 V, reducing to about 1.1 V at the end of the discharge.

The Le Carbone AD cell has as electrolyte a solution of ammonium chloride, which reduces polarisation at the zinc electrode by the action of zinc hydroxide. Large batteries of this type have been used in European railway signalling.

29.3.3.5 Alkaline manganese batteries

Another important recent development, providing up to 50% more power than Leclanché cells of the same volume uses dilute KOH electrolyte in a manganese dioxide/zinc cell with a reverse type of electrode assembly. The zinc anode paste is held in a porous tubular separator surrounding the centrally placed current-collecting nail anode. The cathode mixture of MnO_2 and graphite is highly compressed in the form of annular tablets and these are packed concentrically around the zinc anode. The whole assembly is contained in a thin-walled steel can, with the usual plastic disc-bitumen closure. The extra capacity over standard cells is due to the greater amount of MnO_2 (40–70%) the absence of solid NH_4Cl and the higher conductivity of the KOH electrolyte.

29.3.4 Water activated cells

Water activated cells are stored dry for long periods, then activated by filling with (or immersion in) salt or fresh water; in the latter case, sodium or potassium chloride is included in the cell to increase its conductivity.

29.3.4.1 Silver chloride/magnesium

Silver chloride/magnesium cells are costly, but can deliver very large currents for short periods, typically for electric torpedoes. When flooded with sea water, they become immediately active. The positive electrode (silver chloride) is a thin plate obtained by rolling a slab cast from the molten salt or by chloridising thin sheets of silver. The negative electrode is formed from magnesium strip or sheet. The electrodes are separated by rubber bands, absorbent paper, ebonite forks, or glass beads cemented to the electrode surface. Cylindrical cells with coiled plates have been used in batteries for meterological pilot balloons. During discharge, silver chloride is reduced to metallic silver and the chloride ions Cl^- migrate through the aqueous electrolyte to the negative (magnesium) electrode, which is oxidised to magnesium chloride ($MgCl_2{\cdot}6H_2O$), with the transfer of electric charge. The formation of silver raises the conductivity during discharge: this counteracts polarisation and tends to stabilise the terminal voltage. The e.m.f. is about 1.7 V, and the terminal voltage during discharge falls from 1.5 V to 1.0 V.

29.3.4.2 Copper chloride/magnesium

The chemical reactions are similar to those in the foregoing; the cell is cheaper, but the voltage is lower. Batteries of this type have been used in radar-sonde and other meteorological equipment. For higher voltage at low discharge currents a bipolar construction with duplex electrodes has been used. Positive plates are made by pressing the pelleted salt into copper screens or by dipping into molten chloride. Duplex electrodes are made by pressing or welding positive and negative electrodes to opposite sides of a thin copper foil.

29.3.4.3 Lead oxide/magnesium

The lead oxide/magnesium cell is simple, and comprises a fully formed lead dioxide positive plate (as in a lead/acid secondary cell), flanked by a U-shaped magnesium negative plate, with absorbent paper separators. The paper is impregnated with potassium chloride solution, then dried before assembly. The base of the element is left open and the cell is activated by dipping into fresh water for about 30 s. The electrode reactions are the reduction of lead dioxide to lead dioxide at the positive, and oxidation of magnesium to magnesium hydroxide at the negative, electrode. A two-cell battery for a pilot balloon lamp gives typically 0.3 A at 3.0 V for 30 min at temperatures down to 0°C. As the discharge reactions are exothermic, batteries can still operate in ambient temperatures of – 40°C.

29.3.5 Acid cells

Primary cells with acid electrolytes have been developed for special military and meteorological requirements.

29.3.5.1 Lead oxide/zinc (or cadmium)/sulphuric acid

These cells can be stored 'dry' for long periods and activated as required, an advantage in radio-sonde equipment, meteorological balloons, telemetering in experimental guided missiles and similar special applications. Batteries are referred to as 'short-duration reserve'. Lead dioxide electrodes (like those in lead/acid secondary cells) are welded in parallel according to the capacity required, interleaved with sheets of zinc or cadmium, and separated by thick strips of absorbent paper. The sulphuric acid electrolyte, of relative density 1.270 at 15.5°C, can be introduced some hours before use. The chemical reactions involve the reduction of lead dioxide to lead sulphate and oxidation of the negative electrode to zinc (or cadmium) sulphate. The cell e.m.f. is 2.5 V (with zinc) or 2.2 V (with cadmium). Corrosion of the zinc is reduced by amalgamation obtained by adding about 1% of mercuric sulphate to the electrolyte. The zinc couple gives higher voltage and discharge rate but at temperatures below 10°C, the zinc is polarised by the build up of a reaction product; and at higher temperatures local action is acute, standing loss is high and accelerated activity on discharge can lead to gas polarisation. The cadmium negative is less temperature sensitive, does not require amalgamation and gives satisfactory performance over the range 0–60°C. Typical voltages for a three-cell battery at 3 A discharge for 3 min are 6.8–6.5 V (zinc) and 6.3–6.0 V (cadmium).

29.3.5.2 Lead dioxide/lead/perchloric acid

Excellent discharge reactions are obtainable with perchloric (or fluoboric or fluosilicic) acid producing soluble lead compounds. Polarisation by the build up of reaction products is considerably reduced, so that it is not necessary to increase the surface of the electrodes by making them porous, as in a *secondary* cell. Nearly all of the active electrode material is usable, giving a high output/mass ratio. But for the same reason the cells must be used soon after priming, to avoid high standing loss. The cells are useful where they can be stored dry and primed immediately before use, and where low-mass batteries are required for large currents for relatively short periods. The positive plate is made by electrodepositing a thin non-porous film of lead dioxide on a sheet of nickel, iron or copper from a bath of lead perchlorate, lead nitrate or sodium plumbate. The negative plate may be made by electrodepositing thin layers of lead on a similar metal conductor, or directly from lead sheet. The plates are connected in parallel packs, the interleaved negatives being separated from the positives by spacers allowing rapid ingress of the electrolyte. The latter is held in compartments above the cell and allowed to enter the cell when the cell is primed. During discharge, lead dioxide is reduced to lead monoxide which is at once converted to lead perchlorate; the lead negative is oxidised to the same final product. Batteries have been constructed on the bipolar principle, the lead and lead dioxide being deposited on opposite sides of a nickel sheet. The e.m.f.s with electrolytes of 40–60% concentration range from 2.1 V (perchloric acid) to 1.9 V (fluoboric and fluosilicio acids). Perchloric acid in high concentration can become explosive in the presence of organic materials such as paper, sawdust, etc. Cells with the alternative electrolytes present no such risk, but have lower outputs.

29.4 Secondary cells and batteries

A secondary (or storage or accumulator) cell consists essentially of two electrodes held apart by separators and immersed in an electrolyte, the assembly being fitted into a suitable container. In the lead/acid cell the positive electrode is lead dioxide and the negative is pure lead in spongy form. In the alkaline cell the positive is nickel hydroxide and the negative is iron or cadmium. The electrolyte for the former cell is dilute sulphuric acid and for the latter, dilute potassium hydroxide.

29.4.1 Lead/acid cells

The chemical reactions follow the redox pattern (see Section 29.2.3). The current in the external circuit flows conventionally from the positive to the negative electrode. The two reactions can be written as follows:

positive: $PbO_2 + 4H^+ + SO_4^{2-} + 2e^- \rightarrow PbSO_4 + 2H_2O$

negative: $Pb + SO_4^{2-} \rightarrow PbSO_4 + 2e^{-\leftarrow}$

The reversible potentials are $E_+ = +1.685\,V$ and $E_- = -0.356\,V$, giving an open-circuit e.m.f. $E = 2.041\,V$.

29.4.1.1 Charge and discharge reactions

A secondary cell must be connected to a d.c. supply for charging, positive to positive and negative to negative. During *charging* the reactions are

positive electrode: $PbSO_4 + 2(OH)^- \rightarrow PbO_2 + H_2SO_4 + 2e^{-\leftarrow}$

negative electrode: $PbSO_4 + 2H^+ + 2e^- \rightarrow Pb + H_2SO_4$

The fully charged cell has

positive	electrolyte	negative
PbO_2 (lead dioxide)	dilute H_2SO_4	Pb(spongy lead)

and the overall cell reactions are

$$PbO_2 + 2H_2SO_4 + Pb \underset{\text{charge}}{\overset{\text{discharge}\ \rightarrow\leftarrow}{}} 2PbSO_4 + 2H_2O$$

During *discharge* the active parts of both electrodes are converted to lead sulphate, and the concentration of the electrolyte is reduced by both the removal of sulphate ions and the formation of water. During *charge* the lead sulphate at the negative plate is reduced to spongy lead, and at the positive electrode to lead dioxide with the release of sulphate ions at both plates and an increase in the concentration of the electrolyte. Measurement of the concentration indicates the state of the electrodes: the relative density of the electrolyte is read by a hydrometer.

Typical charge/discharge cell voltages are shown in *Figure 29.5*. Lead sulphate, the product of the discharge reaction, is practically insoluble in the electrolyte, a factor that endows the cell with its high degree of reversibility. During cycling the lead sulphate remains where it is formed, and the structure of the active materials is relatively undisturbed.

29.4.1.2 Materials

Lead dioxide and spongy lead are the active materials, but it is necessary to provide metallic frames or supports for them. The success of the modern secondary cell depends on plate design and manufacture, and on the separators that prevent internal short-circuits and help to retain the active material in position.

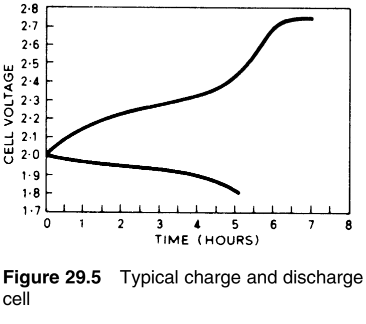


Figure 29.5 Typical charge and discharge voltages of a lead–acid cell

Plates

Planté (formed) plate The positive plate has its effective surface area increased ten-fold by forming close-pitched fins on the surface of a pure lead plate. The negative plate was commonly of a ‘box’ form.

Faure (pasted) plate The active material applied to open-mesh grids cast in antimonial lead is a paste made by mixing lead oxide with water and sulphuric acid. The plate is seasoned, dried and then electrochemically converted to lead dioxide or spongy lead by charging in dilute sulphuric acid. The grid acts both as a support for the active material and as the conductor of the current to and from the active material. The merit of the pasted plate is that the grid can be cast in precisely defined thin sections as low as 1.5 mm or less, and the ratio between active material and carrier grid is high. In addition to the standard gravity die-casting process, in modern developments, thin grids are made by stamping or punching suitable patterns from thin pure lead or lead-calcium alloy sheet. The red lead and litharge originally used for the positive and negative plates have been superseded by a grey oxide containing one-third fine metallic lead particles and two-thirds lead monoxide.

Reference has already been made to the box type of pasted negative plate used with Planté positive plates, in which the paste is held in the grid by thin sheets of perforated lead covering both surfaces. The box negative plate is however costly, both on materials and manufacture, and has now been largely superseded by the common pasted negative plate.

Separators In some cells the separators serve only to keep the plates equidistant; in others they act as diaphragms to prevent internal short-circuit or to retain the active material.

Earlier wood separators have been almost entirely superseded by artificially made forms. One of the most successful is a microporous polyvinyl chloride (PVC). It has a high degree of diffusability, a low electrical resistance in acid and great durability under normal battery conditions. It is used in all types of lead/acid batteries. Another type of separator is made by sintering fine particles of PVC, and yet another by impregnating an absorbent paper with an acid-resistant resin such as phenol formaldehyde which both stiffens the paper and protects it against attack by the acid.

The profile of the separator is of interest. One product of the chemical reaction at the positive plate is water, the other being lead sulphate, which removes SO_4 ions from the electrolyte. The acid is doubly diluted at the positive plate, whereas at the negative only lead sulphate is formed. It is, consequently, necessary to provide a greater reservoir of acid adjacent to the positive plate, and in cells where the amount of acid is minimised it is usual to make separators with ribs which rest against the plate surface.

Electrolyte Different concentrations of the sulphuric acid electrolyte may be used, depending on the application for which the particular battery has been designed. The voltage of the cell depends on the concentration of the electrolyte, being higher for higher concentrations. Also, the minimum electrical resistivity at 20°C occurs at a concentration of about 31%, equivalent to a density of 1.225; and the minimum freezing point at a relative density of 1.300. Taking these and other factors into account, stationary cells are generally filled with acid of relative density 1.200–1.215; automobile and traction batteries, 1.275–1.285, or in tropical climates 1.240–1.260.

One important operational aspect of the electrolyte concerns *impurities*. Impurities such as chloride and acetic or nitric acid attack the positive grids and are generally kept to a minimum in both the filling-in acid and the 'topping-up' water which is added to replace water lost by evaporation and by electrolysis during charging. Metals such as iron and manganese cause self-discharge of both positive and negative plates, while nickel and copper are 'plated out' on the negative and also cause self-discharge. Limits for various impurities are given in British Standards. For topping up, it is generally advisable to use only distilled water.

29.4.1.3 Construction

Battery design and construction is considerably influenced by the application.

Road vehicle starting, lighting and ignition A battery which is used for starting, lighting and ignition (SLI) must, *inter alia,* be able to deliver up to 100 A/h of electrical energy at the 20-h rate and be able to supply high currents (e.g. 400–450 A) for up to half a minute without too much voltage fall-off for engine start-up purposes. Faure plates have a special advantage where high currents are required. The larger the plate surface area, the lower the polarisation and the higher the cell voltage. Batteries normally comprise three-cell (6 V), six-cell (12 V) and 12-cell (24 V) assemblies, with plates connected in parallel groups in moulded containers.

In the past, batteries were supplied which, after filling with acid, required an extended first charge to reduce the active spongy lead negative plate. More recently, a dry-charged automobile battery has been developed to give about 75% of its nominal capacity shortly after filling, even after lengthy storage. The basic requirement in manufacture is to ensure a high percentage of lead dioxide in the positive plates and a minimum (e.g. 10%) of lead monoxide in the negative plates. This is achieved by the inclusion in the plates of an antioxidant, or by drying the plates after formation in an oven in which they cannot come into contact with atmospheric oxygen.

Traction These batteries often have larger voltages and higher capacities than the SLI type. Batteries for electric vehicles may have a flat-plate or an iron-clad tubular form. The application involves deeper cycling and causes shedding of the active material. In the Faure plate design, 'retainers' (generally of thickly matted glass-wool fibres) are placed in close contact with the surfaces of the positive plates. Batteries of this type (*Figure 29.6*) have lives of 6 years or more. With tubular plates (*Figure 29.7*) tubes of high porosity are fitted: one such plate uses non-woven fibres of a plastic inert to sulphuric acid, such as Terylene. Another type uses an inner woven stocking of glass fibre, strengthened by individual thin-walled perforated PVC tubes. By raising the permeability to acid, these designs have enabled the output/mass and output/volume ratios to be increased by 30% without loss of durability and cycling life.

29.4.2 Nickel/cadmium and nickel/iron alkaline cells

Both of these commercially available alkaline cells have the same electrolyte, dilute potassium hydroxide, and the same positive active material, nickel hydroxide. The nickel/cadmium cell has a negative plate of cadmium with a small proportion of iron. Both have an e.m.f. of about 1.2 V. Cadmium gives the nickel/cadmium cell a lower charging voltage and reduced ohmic resistance, and its characteristics resemble those of the lead/acid cell. Use of the nickel/iron cell is mainly in traction, where the higher charging voltage and internal resistance are less important.

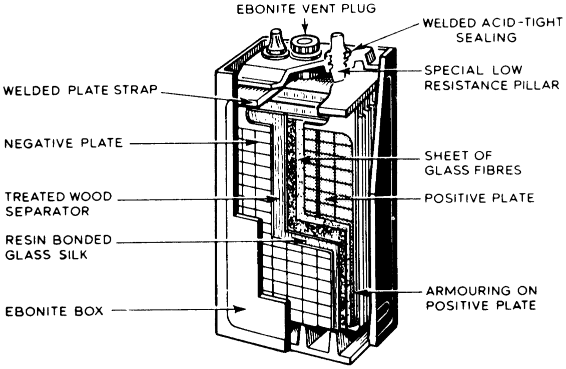


Figure 29.6 Sectional view of a traction cell

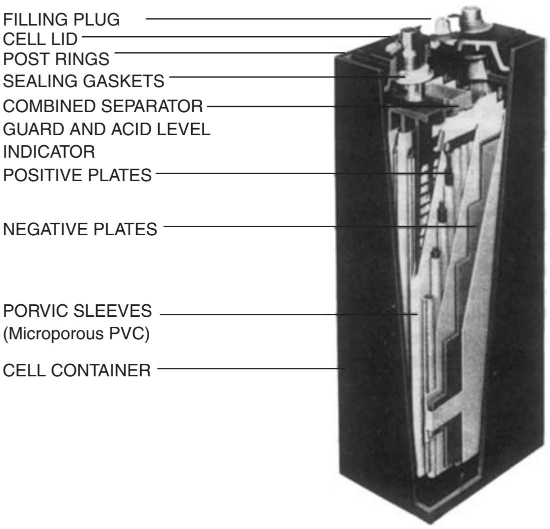


Figure 29.7 Iron clad cell with tubular positive plates

29.4.2.1 Charge and discharge reactions

The chemical reactions are complicated. The following gives a general guide:

positive: $2Ni(OH)_3 - H_2O \to 2NiOOH + 2H^+ + 2e^- \rightarrow\leftarrow 2Ni(OH)_2$

negative: $Cd + 2OH^- \to Cd(OH)_2 + 2e^{-\leftarrow}$

The reversible potentials of these reactions are respectively $E_+ = +0.49$ V and $E_- = -0.81$ V, giving an open-circuit e.m.f. per cell of $E = 1.30$ V. The overall reaction is

$$2NiOOH + 2H_2O + Cd \overset{\text{discharge}\ \rightarrow\leftarrow}{\underset{\text{charge}}{}} 2Ni(OH)_2 + Cd(OH)_2$$

These reactions apply to the nickel/cadmium cell. In the nickel/iron cell the cadmium (Cd) is replaced by iron (Fe).

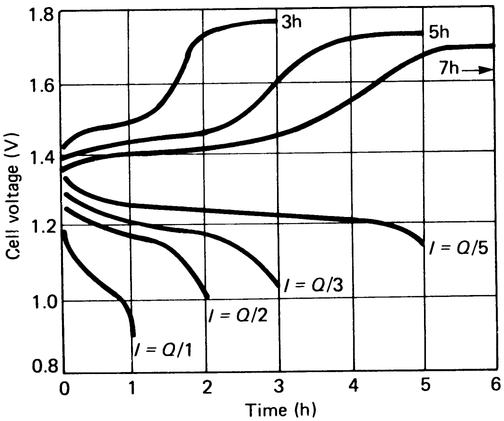


Figure 29.8 Charge and discharge characteristics of a low resistance nickel/cadmium cell

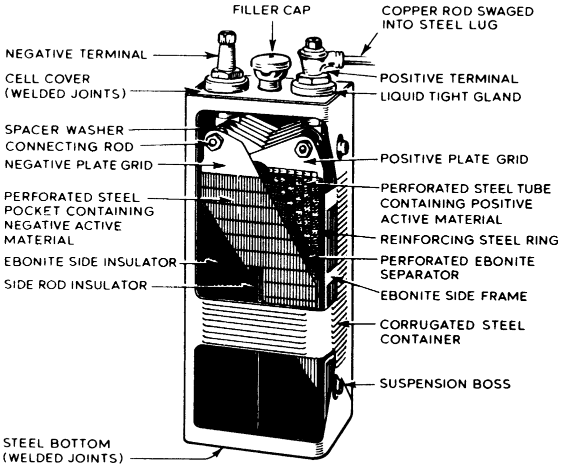

FILLER CAP
COPPER ROD SWAGED INTO STEEL LUG
NEGATIVE TERMINAL
CELL COVER (WELDED JOINTS)
POSITIVE TERMINAL
LIQUID TIGHT GLAND
SPACER WASHER
CONNECTING ROD
NEGATIVE PLATE GRID
POSITIVE PLATE GRID
PERFORATED STEEL POCKET CONTAINING NEGATIVE ACTIVE MATERIAL
PERFORATED STEEL TUBE CONTAINING POSITIVE ACTIVE MATERIAL
REINFORCING STEEL RING
EBONITE SIDE INSULATOR
PERFORATED EBONITE SEPARATOR
SIDE ROD INSULATOR
EBONITE SIDE FRAME
CORRUGATED STEEL CONTAINER
SUSPENSION BOSS
STEEL BOTTOM (WELDED JOINTS)

Figure 29.9 Nickel/iron cell

The electrolyte is a solution of pure potassium hydroxide (KOH) of relative density about 1.200, a small amount of lithium hydroxide being sometimes added. The electrolyte takes no apparent part in the reactions and its density remains substantially constant. Cells can stand indefinitely in any state of charge, provided that the plates are kept immersed. As in the lead/acid cell, water is lost during gassing on charge and is made up with distilled water. Electrolyte is added only to make good accidental spillage.

Typical charge and discharge characteristics are shown in *Figure 29.8*. The time for a full charge is 7 h, the voltage per cell rising from about 1.4 V at the start to about 1.8 V (nickel/iron) or 1.7 V (nickel/cadmium) after about 5½ h, and then remaining constant for the remainder of the charge. The input (in ampere-hours) should be 1.4–1.5 times as great as the previous discharge. Filler caps or vents should be kept closed except when topping up or taking gravity readings.

As alkaline cells are not damaged by overcharging, a full normal charge can be given irrespective of the state of the cells.

29.4.2.2 Construction

Nickel hydroxide forms the active material of positive plates, the form being 'tubular' for the nickel/cadmium and 'pocket' for the nickel/cadmium cell. Negative plates are of the pocket form for both: the active material is first pelleted, and the pellets then firmly enclosed in pockets of this nickel-plated steel, perforated with many minute holes. In tubular plates helically wound tubes of similar perforated material are constructed, and active material interspersed by layers of thin nickel flake is tamped into them. Groups of plates of the same polarity are bolted or welded to steel terminal pillars. Plates are kept apart by ebonite rod separators. The cell assembly is fitted into a welded sheet-steel container, the terminals being brought out through the lid in suitably insulated glands.

The construction (*Figure 29.9*) produces sturdy, robust cells, unaffected by vibration and shock. Batteries are built up by assembling cells in hardwood crates.

The need to increase the electrical conductivity of the nickel hydroxide positive active material of the alkaline cell was recognised at an early stage of its development. In the tubular plate this was done by introducing small pockets of extremely thin metallic nickel flake at regular intervals in the positive active material as it was tamped into the tube. In the pocket type of plate graphite powder is mixed with nickel hydroxide before it is pressed into a pellet. The same object was achieved in a *sintered* plate developed in the late 1930s. The plate grid or support was prepared in the form of a highly porous sintered nickel plaque and the active materials (nickel hydroxide for the positive plate and cadmium hydroxide for the negative) were deposited in the fine pores from suitable solutions of their salts. In this way the active materials were distributed evenly over a very large conducting surface and relatively high coefficients of use were obtained. By this method also very thin plates could be made in closely controlled thicknesses of 0.75 mm or less, and when these were interleaved with thin separators of woven or felted cloth, close-packed assemblies were produced capable of delivering very high currents, particularly at low temperatures.

Sintered plate batteries have been extensively used in aircraft and in other applications calling for high discharge currents.

It was later discovered that if the amount of electrolyte were reduced to such a point that there appeared to be no free potassium hydroxide, the cell could be submitted to overcharge in a fully sealed condition without any risk of the container bursting. The chemistry of this reaction can be explained simply by saying that any oxygen produced at the positive electrode during charge is immediately absorbed by the cadmium in the negative electrode and this is converted to cadmium oxide. The negative charge is therefore used to reduce this cadmium oxide and no gaseous hydrogen is evolved.

Sealed cells require no topping up, evolve no vapour or spray, and can be installed in equipments without hazard. But occasionally they partially dry out, so it is the custom to fit release valves. Batteries of sealed cells have found application in aircraft.

Development has resulted in small cells of capacities less than 1 A-h. They are made in the form of a large button or in cylindrical shape. The former generally have compressed pellets of the active materials; the latter have thin sintered electrodes. Both have restricted amounts of electrolyte and function as fully sealed cells. They are applied to transistor radio receivers and other electronic devices.

29.4.3 Silver/zinc alkaline cells

Silver/zinc alkaline cells have recently come to the fore in applications where very high outputs per unit of weight and volume are required. In these respects they give outputs from four to five times those of the equivalent lead/acid and nickel/cadmium alkaline cells. Thus, a 100 A-h cell may be only 10 cm × 5 cm in plan and 20 cm in height, and have a mass of 1.5 kg. The silver is costly, and in view of the relatively low reversibility of multicell batteries, silver/zinc batteries have found few commercial applications. They have been used in aircraft, in guided missiles and to supply power to the communication systems in satellites.

29.4.3.1 Charge and discharge reactions

As in the conventional alkaline system, potassium hydroxide of about 1.200 relative density is used as the electrolyte. The silver oxide used as the depolariser in the positive plate passes through two stages of oxidation (AgO and Ag_2O) each of which has a characteristic electrode potential. During discharge the following reactions take place:

positive: (a) $2AgO + H_2O + 2e^- \rightarrow Ag_2O + 2OH^{-\leftarrow}$

$E_+ = +0.57\,V$

(b) $Ag_2O + H_2O + 2e^- \rightarrow 2Ag + 2OH^{-\leftarrow}$

$E_+ = +0.34\,V$

negative: $Zn + 2OH^- \rightarrow Zn(OH)_2 + 2e^{-\leftarrow}$

$E_- = -1.2\,V$

The overall cell reactions are therefore

$$\left.\begin{matrix} 2AgO + H_2O + Zn \\ Ag_2O + H_2O + ZN \end{matrix}\right\} \begin{matrix} \text{discharge} \rightarrow\leftarrow \\ \text{charge} \end{matrix} \left\{\begin{matrix} Ag_2O + Zn(OH)_2 \\ 2Ag + Zn(OH)_2 \end{matrix}\right.$$

For the former reaction the open-circuit cell voltage is 1.81 V and for the latter it is 1.58 V.

The discharge voltage curves of silver/zinc cells therefore generally show two plateaux, a relatively short one at about 1.8 V and a relatively long one at about 1.5 V. With some pre-treatment, as, for example, a very short preliminary high-rate discharge, it is possible almost to eliminate the first plateau, and the discharge voltage then remains fairly steady at about 1.5 V per cell. A curve of this type is shown in *Figure 29.10*.

The nominal cell voltage is 1.5 V, and this is held (at normal discharge rates) substantially constant over most of the discharge period, as shown. High output currents and charge rates can be employed with little sacrifice in effective capacity. Normal ampere-hour efficiency is 90–95% and watt-hour efficiency 80–85%. Practically no gassing takes place. The electrolyte is to a major degree absorbed in the active materials and almost completely immobilised.

Owing to the slight solubility in potassium hydroxide of the higher silver oxide, overcharging of reversible silver/zinc batteries should be avoided. In general, charging should be stopped when the voltage reaches 2.1 V per cell and, for this reason, recharging is best carried out with constant-voltage control. If this is not done, short-circuits may develop through the separators. To reduce this risk, manufacturers generally enclose the zinc negative plates in several layers of the separator material, which is usually of a cellulosic base. Single cells of this type have a fair degree of reversibility, but in a multicell battery the irregularity between cells, which is difficult to eliminate, has an adverse effect on reliability. And the greater the number of cells in the battery, the shorter the cycling life becomes.

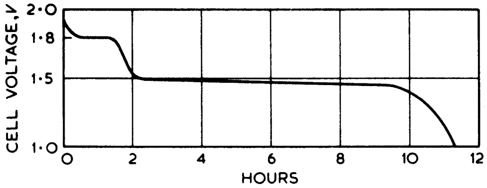


Figure 29.10 Discharge voltage of a silver/zinc alkaline cell

29.4.4 Secondary battery technology

29.4.4.1 Voltage

The open-circuit voltage of a fully charged secondary cell is the same, however large or small, but varies according to the type, i.e. lead/acid or alkaline. That of a fully charged lead/acid cell is approximately 2 V; that of an alkaline cell is about 1.2 V.

Because of the difference in cell terminal voltages (2.0 V for lead/acid and 1.2 V for Ni/Cd), five Ni/Cd cells are required to replace batteries with three lead/acid cells.

29.4.4.2 Capacity

The capacity of the lead acid battery increases with decreasing rates of discharge (currents) and increasing durations (hours). It is therefore usual to state the rate of discharge for any declared capacity, e.g. at the 10-h rate, the 5-h rate, the 1-h rate and so on. *Figure 29.11* shows families of discharge- and charge-voltage curves for a typical lead/acid cell on discharge at constant current in each case. The salient properties are:

Discharge time (h)	10	7.5	5.0	3.0	2.0	1.0	0.5
Current rate (A)	11	14	19	29	40	65	118
Mean voltage (V)	2.0	2.0	1.98	1.94	1.90	1.84	1.70
Energy (A-h)	110	105	96	88	79	65	59
Power (W-h)	220	210	190	170	150	120	100

Points to note are: (a) the O/C voltage is about 2.0 V; (b) the discharge voltage curves all start with a relatively flat plateau which ends with a sharp decline, usually referred to as the 'knee' of the curve, beyond which little further capacity is available; (c) the slope of the discharge curves becomes steeper as the rate (current) increases and the knee becomes less prominent; and (d) the 'energy' capacity is represented by the product of the current and the duration (A-h) and the 'power' capacity by the product of energy and the mean voltage (W-h).

The theoretical amounts of the active materials consumed per A-h of discharge energy are: PbO_2, 4.45 g, Pb, 3.86 g and H_2SO_4, 3.68 g. The actual amounts are considerably greater as the coefficients of use fall very much below 100%, as indicated below:

	5-h rate	*1-h rate*	*5-min rate*
Positive, PbO_2	52%	33%	16%
Negative, Pb	66%	40%	19%

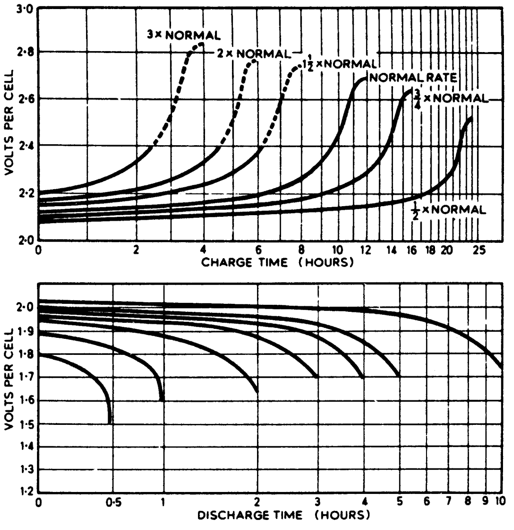


Figure 29.11 Representative curves of charge and discharge

The capacity is related to the amounts of the active materials and to their porosity and intrinsic surface area. The main product of the discharge in both positive and negative active materials is lead sulphate. This has a very high electrical resistivity. It polarises the active material and, by clogging the pores restricts diffusion of the electrolyte into the reaction zones. Planté positive grids are cast with a developed area at least 10 times the superficial area and the anodically prepared active material—PbO_2 is spread in a thin layer over a large area. In Faure pasted plates on the other hand, the porosity of the active material must be adequate to permit rapid diffusion of the sulphuric acid electrolyte, particularly during discharges at high rates.

Thus, it is found that (a) for stationary batteries of the lead/acid type, on the score of durability and efficiency it is best to use a Planté positive with a large surface area lightly coated with active material; (b) for automobile or diesel engine starter batteries it is best to use a large number of comparatively thin pasted plates; and (c) for traction battery service, where lower rates are usual, thicker plates with denser material can be used.

29.4.4.3 Charging

All secondary batteries require a supply of direct current for recharging. The method of charging is important in its effect on battery performance and service life. The three principal methods are as follows.

System control The batteries work in parallel with a generator across the load. Typical applications include motor vehicles, in which the battery forms an essential standstill reserve for ignition, lighting, signalling, etc. Most small motor cars employ a d.c. generator with control gear for cutting out the generator when its speed is too low, and for voltage regulation for operation over the wide range of active speed. Larger cars and road vehicles employ a.c. generators (alternators) with rectifiers and appropriate automatic control equipment.

Manual or semi-automatic control Here the battery is recharged from a separate source, such as a transformer/rectifier. Typical applications are to electric vehicles, in some cases with the transformer/rectifier carried on the vehicle.

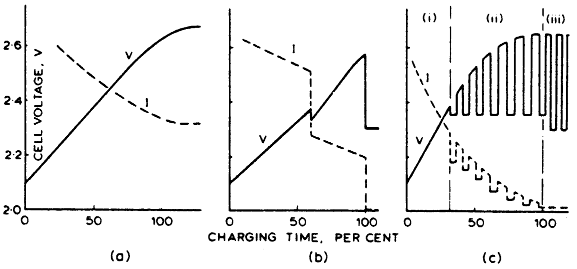


Figure 29.12 Charger waveforms

Since the performance and life of motive power batteries largely depend on the efficiency of the charging system, these will be described in more detail. Three forms of charger are in general use, the single-step taper charger, the two-step taper charger and the pulse-control charger, of which the Chloride Spegel is a typical example. The voltage-current-time relationships for these three types are shown in *Figure 29.12*. With the single step charger, the voltage is held constant just below the battery gassing point, 2.4 V per cell. When the battery voltage reaches this point, the current is allowed to decay until the battery becomes fully charged. With the two step process, the first stage follows the same line as with the single step; the current is then reduced to a pre-determined value and charging continued for a pre-determined period. The pulse type of charger is devised more to keep the battery fully charged. When the main charge is stopped, the battery receives short pulses of charge, controlled by the decay in the battery voltage during the intervening open-circuit period.

Float-and-trickle control Large stationary batteries in generating stations and (for emergency lighting) in hospitals are connected permanently to a charger, and may also be connected across the load. In normal operation the battery receives a trickle charge to keep it fully charged and ready for intermittent loading. Such a battery requires ventilation and ready access for servicing.

29.4.4.4 Maintenance-free, gas recombination technology

One of the most important developments in storage batteries during the past 20 years or so has been the production of maintenance-free (MF) systems. The primary object was to reduce or eliminate altogether the need to replace water lost through three main factors; (a) evaporation; (b) local electrochemical action at lead negative plates due to the deposition of metals, such as antimony having a lower hydrogen overvoltage than lead; and (c) electrolysis of the aqueous electrolyte towards the end of each recharge. With well-stoppered cells, (a) presented few problems. So far as (b) was concerned, from the earliest days of manufacture, antimony has been the favourite hardener for the lead grids used to support the active materials and to collect the current. But, because of their different hydrogen overvoltages, when lead and antimony come in contact in the electrolyte, hydrogen is evolved, water in the electrolyte is decomposed, some of the lead active material is oxidised causing a corresponding loss of capacity—the so-called 'standing loss'. Many metals have been tested in lead alloys for this application and the most successful has been lead–calcium alloy, containing 0.08% calcium and small amounts of tin. With grids cast in this metal, standing losses are negligible. Losses by the third factor (c) proved more difficult to contain, but following the route taken for alkaline cells, the principle of gas re-combination has now been fully established for lead–acid batteries. The main requirements are: (a) no free electrolyte (to permit rapid diffusion of oxygen to the negative plate surfaces, the elements should be 'starved' of electrolyte); (b) the bulk of the electrolyte should be retained by the porous separators; and (c) there should be excess negative capacity. Oxygen is evolved from the positive group before the negative is fully charged. It diffuses to the negative plates where it is absorbed causing discharge of the active material. The negative plates are, therefore, never fully charged and hydrogen is not evolved. This principle is now being applied to cells of various types from small button cells to batteries for SLI, traction and stationary service. Cells can be fully sealed, though it is usual to incorporate a pressure-controlled safety valve.

29.4.5 Lithium cells

In recent years there has been much activity on the part of the battery companies directed towards developing cells having lithium electrodes. These electrodes have many potential advantages in battery technology, including high cell voltages, high energy densities, high power densities and rechargeability. Thus they give rise to both primary and secondary batteries. Lithium as an electrode gives a high energy/mass ratio and an inherently high power density. The standard electrode potential is over -3.0 V and when allied with, say, a fluorine electrode ($E_0 = +2.87$ V), the resulting cell should theoretically have an O/C voltage of over 6.0 V. However, problems arise in practice because of the extreme reactivity of lithium which is stable only in certain non-aqueous media (see below) and even in these it is often the formation of surface films on the lithium which determines the performance of the battery.

There are many different types of lithium cell, having both different electrolyte solutions and different positive electrodes, but with a common lithium negative electrode. The solvents include organic materials such as propylene carbonate containing lithium salts, inorganic materials such as thionyl chloride, molten salts such as the lithium chloride/potassium chloride eutectic and solid electrolytes such as lithium iodide and organic polymers, e.g. polyethylene oxide.

Lithium/thionyl chloride battery technology is both well advanced and well established. Notable features of these batteries are their very flat discharge characteristics, long shelf-life and applicability over wide temperature ranges.

Although a wide range of primary lithium batteries are now commercially available it has not been straightforward to produce viable secondary cells. The problems mainly arise in subjecting both the negative and positive electrodes to charge/discharge cycles in these cells.

The following couples have been employed with organic electrolytes: Li/MnO_2, Li/SO_2, $Li/(CF)_n$, Li/MoS_2, Li/Ag_2CrO_4, Li/CuO, Li/FeS_2 and Li/TiS_2. Unfortunately, the electrical resistance of these organic electrolytes, even when improved by inorganic salt additives, is relatively high and cells of this type are only suitable for low rate discharges.

With molten LiCl–KCl, the alloys Li–Al and Li–Si have been used as negative electrodes, and FeS and FeS_2 as positive electrodes. Operational temperatures may be as high as 400°C.

The Li/SO_2 and $Li/SOCl_2$ couples have inorganic electrolyte and can be operated at high discharge rates at ambient temperatures. They have, in fact been tested for possible torpedo propulsion.

29.4.6 Sodium/sulphur cells

Another high-temperature contender (300/375°C) is the sodium/sulphur cell, which employs liquid sodium and sulphur as the negative and positive electrodes, respectively, with the solid electrolyte β-alumina which is a sodium ion (Na^+) conductor. It is envisaged that this system would be used for load levelling and car traction. The open circuit voltage of a single cell is 2.1 V and its theoretical energy density 756 W-h/kg. Safety is a major concern. In practice, cells have capacities of about 150 A-h with practical energy densities in the 100–150 W-h/kg range. Heat management is also very important; it is desirable to employ the batteries for applications which enable them to remain molten during on and off periods. In the UK a company (Chloride Silent Power) set up by the Electricity Council and the Chloride Group is exploiting the sodium/sulphur battery for large-scale application.

29.5 Battery applications

29.5.1 Stationary or standby power batteries

Batteries in large installations having capacities of hundreds of ampere-hours, which supply power in emergency or auxiliary situations, are referred to as stationary or standby batteries. Standby power supply is needed when the mains source of electricity fails. The choice of standby supply is usually between lead–acid or nickel–cadmium batteries or auxiliary generators, or a combination of both sources. Applications include emergency lighting in department stores, offices, factories, cinemas and other public places, power to keep burglar alarms and other communications working, emergency power for important production processes and equipment such as in hospitals, mines, airports, power stations, telephone exchanges, lighthouses and, more recently, computer operations especially in the financial service sector.

29.5.1.1 Lead–acid batteries

Various types of lead–acid secondary batteries are used for stationary or standby purposes:

- Plante cells giving a long life of 20–35 years or even longer with high reliability which are kept on continuous trickle charge. Cell capacities are rated by relating current flow to time in ampere-hours. High performance cells up to 2000 Ah are usually in transparent containers allowing acid levels to be checked.
- Flat plate cells provide a cheaper but short-lived alternative to Plante cells. Pasted flat plate cells, again in transparent containers are available in capacities of up to 500 Ah with a life expectancy of 10–12 years to meet emergency lighting regulations. Car batteries are based on this design but are totally unsuitable as an emergency power source being designed instead to give a high current for a short time.
- Tubular cells are normally used to power electric trucks on which daily recharging is needed, but are suitable for standby applications requiring frequent charge/discharge cycles. When on standby they have a life expectancy of 10–12 years compared to 5 years when powering electric trucks.

29.5.1.2 Nickel–cadmium batteries

Large nickel–cadmium secondary batteries, open and semi-open or sealed are made in a variety of designs to meet many purposes including those of stationary or standby batteries. Their uses cover the application areas mentioned above. Other applications include engine starting for motor-generating sets as required for standby power systems where bad weather, low temperature conditions might prevail. Under these circumstances engine-starting batteries must be highly reliable, have a high performance capability and be free from the effects of high stress and vibration, such as occurs in vehicles or engine rooms of ships. A similar requirement is for emergency batteries in trains for lighting at stations or in the case of a power supply failure. Nickel–cadmium cells are used also for the uninterrupted emergency power supply (UPS) to railway mainline stations and signalling systems. Marine applications include UPS systems for radar, communications, steering and navigation and for the emergency lighting on board. In air transport these batteries provide ground power supplies, the support battery systems on board aircraft and at airports all the essential lighting and communications depend on batteries. UPS systems operate in standby mode, providing power only in the event of mains failure.

29.5.2 Self-contained power supplies

Applications as self-contained power supplies include a wide range of portable devices such as transceivers, razors, portable equipment and tools, radios, torches, toys, hearing aids, cameras, calculators, computers, clocks, etc.

Sealed nickel–cadmium secondary batteries have a range of uses in these markets resulting from their favourable characteristics including:

(1) total absence of maintenance;
(2) very long operational life;
(3) ability to accept permanent overcharge;
(4) prolonged storage without deterioration;
(5) ability to be overcharged at high rates;
(6) constant discharge voltage characteristic;
(7) mechanically robust; and
(8) operation over a wide temperature range of −40°C to +50°C.

If the application requires a low initial cost, or the life of the equipment is short, however, then primary batteries are the obvious choice. Secondary batteries can prove cheaper in the long run because they can be recharged and reused, but they require a charger, so the initial cost of the battery system is high.

Of the primary cells zinc–carbon is cheap and readily available. It is particularly useful in applications having a light, intermittent duty cycle, but the battery has a low shelf life and a drooping discharge characteristic. For heavy-duty applications needing continuous operation at high current, alkaline manganese dioxide batteries are preferred to zinc–carbon. They have 50–100% more energy for the same weight and a longer shelf life, but are more expensive at low currents, e.g. below 200 mA. The mercury-oxide battery has a flat discharge characteristic and a high energy to weight ratio. It is often used in voltage reference applications. Lithium batteries are used in applications that require a long shelf life or a wide operating temperature range, or a high energy density.

29.5.3 Traction batteries

Much of the recent interest in battery development has been linked to traction purposes. There have been a number of obstacles to the widespread introduction of electric vehicles (EVs) based on batteries alone:

- Inability to store sufficient energy for long distance travel;
- Time taken to recharge the battery;
- Lack of supporting infrastructure of fore-court services and maintenance facilities;
- Replacement costs of battery pack; and
- Additional vehicle weight of battery pack.

Lead–acid and nickel–cadmium store too little energy per unit weight and volume to be viable for electric traction; therefore the development of the electric vehicle is critically dependent on designing advanced batteries which can store two to three times the amount of energy per unit weight and volume compared with these conventional rechargeable batteries. Significant advances have been made in electrochemical battery development, however, over traditional lead–acid battery performance. Advanced battery technologies such as room temperature lithium ion and high temperature sodium nickel chloride have energy densities of 2.5 to 3 times that of lead acid, resulting in electric vehicle ranges of possibly 150–250 km at 100 km/h, i.e. suburban use, with sufficient energy to power normal levels of auxiliary equipment. The high temperature sodium nickel chloride battery (sodium metal negative electrode—nickel chloride positive electrode separated by an ion conducting β-alumna ceramic) is a robust technology with demonstrated cycle life, performs well in cold weather, but would not be energy efficient in vehicles with low mileage because of the self discharge needed to keep the battery heated. The lithium ion battery has good energy performance, is safe, but has relatively low specific power and requires considerably more development; nevertheless the development of high-power systems are anticipated in the future.

The state of development of secondary batteries along with the targets set by electric passenger vehicle (EV) requirements is summarised in *Table 29.1*. The data does not account for improvements which can be made by properly designed engine-management systems. The use of ultracapacitors, for example, can supplement power supplied. A battery's life could be as much as doubled by the use of ultracapacitors for load-levelling purposes. This technology is well suited for delivering high power to EV's during acceleration and generating new energy supply during braking using power electronics for four-quadrant control.

29.6 Anodising

Anodising is the production of a film of oxide or hydrated oxide on the surface of aluminium or of its alloys, to prevent corrosion of aircraft metal and other surfaces likely to be exposed to the effects of the sea. The anodic film is usually of a light grey colour. When newly formed, it can absorb dye, a property of use for decorative purposes.

29.6.1 Process

The workpiece is made the anode in a bath of chromic or sulphuric acid, through which a direct current is passed at a voltage of about 50 V for the former acid bath and 10 V for the latter. In large installations a current of 1–2 kA may be required from a rectified a.c. supply. The action produces on the aluminium workpiece a semi-insulating film, and

Table 29.1 Battery performances and traction requirements

Cost £/kWh	*Required performance*	*Present achievements*
Lead–Acid	75–100	50–100
Nickel–Iron		100–200
Nickel–Cadmium		150–200
Nickel–Metal hydride		100–150
Sodium–Sulphur		100–200
Sodium–Nickel chloride		200–300
Lithium polymer		25–250
Specific power W/kg		
Modern car engine		400
Lead–Acid	300–400	67–183
Nickel–Iron		70–140
Nickel–Cadmium		100–200
Nickel–Metal hydride		150–200
Sodium–Sulphur		90–130
Sodium–Nickel chloride		100–150
Lithium polymer		75–100
Specific energy Wh/kg		
Modern car engine		200
Lead–Acid	100–200	18–56
Nickel–Iron		40–70
Nickel–Cadmium		33–70
Nickel–Metal hydride		54–80
Sodium–Sulphur		80–140
Sodium–Nickel chloride		75–100
Lithium Polymer		125–150
Energy density Wh/litre		
Lead–Acid	150–300	50–80
Nickel–Iron		60–120
Nickel–Cadmium		60–120
Nickel–Metal Hydride		150–200
Sodium–Sulphur		75–125
Sodium–Nickel Chloride		125–175
Lithium Polymer		100–120
Life–years		
Lead–Acid	5–10	2–3
Nickel–Iron		3–5
Nickel–Cadmium		?
Nickel–Metal Hydride		5–10
Sodium–Sulphur		?
Sodium–Nickel Chloride		3–5
Lithium Polymer		?
Number of Discharge cycles		
Lead–Acid	500–1000	500–1000
Nickel–Iron		500–2000
Nickel–Cadmium		1500–2000
Nickel–Metal Hydride		750–1000
Sodium–Sulphur		250–750
Sodium–Nickel Chloride		500–750
Lithium Polymer		250–500

the vat voltage is increased as the action proceeds. Non-aluminium parts must be screened by plastics materials. The current density is 0.5–0.7 A/cm^2 of active anode surface. In a typical process about 10 V is initially applied, the current being 1.5 kA; after 30 min the voltage is steadily raised in 5 min to 50 V, and maintained for a further 5 min. The work is then removed from the vat and washed.

The liquid carries the anodising process into every bare recess so long as there is no exclusion by gas. A tube 2 m long and of 5 mm internal diameter can be satisfactorily anodised over its entire surface. The effectiveness of the surface can be tested by applying 50 V between the workpiece and a 25 mm diameter steel ball. Anodising is best carried out at a standard temperature (e.g. 40°C).

29.6.2 Vats

In small baths the cathodes are carbon plates or rods hung from the cathode connection. Current is conveyed to the cathodes and the anode workpiece by round copper rods resting on porcelain insulators attached to the rim of the wooden or metal vat. A typical vat is 6 m × 2 m with a depth of about 2 m, sunk into a concrete floor for convenient access. The cost of filling a bath with diluted acid is considerable, making necessary scrupulous attention to cleanliness and the avoidance of contamination. Precipitates must be allowed to settle or be filtered off. Gentle agitation of the liquid by means of compressed air from a small motor-driven pump is applied to avoid slight pitting of the workpiece at the liquid surface level: this keeps foreign bodies in suspension, so that filtering may be necessary at the end of the day's work.

29.6.3 Workpieces

The pieces to be anodised are wired for connection to the anode rod by aluminium wire if they are small. Larger pieces may be connected by dural rods and clamps. The cross-section of the wires and rods must be adequate for the appropriate current, and that of the connection point to the workpiece must also be sufficient. Multiple connections can be used for this purpose. The action of the bath produces an effervescence of oxygen from the aluminium and dural, and a gas accumulation may exclude workpiece recesses from the anodising process. This can be overcome by tilting the workpiece, or may necessitate two runs with the workpiece in different orientation.

29.6.3.1 Cleaning

Before anodising, workpieces must be cleaned by a process similar to those for the preparation of work for electroplating. Items that have not been heat-treated, welded or riveted are more easily cleaned than fabricated parts or castings. If the workpiece has been previously treated in a salt bath, all trace of the salt must be removed by flushing in running water. Built-up fittings require the joining surfaces to be scratch-brushed or buffed before refitting, and immersed in boiling water immediately prior to anodising. Welded aluminium parts must be kept in contact with boiling water for 30 min to remove flux residues, which lead to pitting. In large anodising shops greasy articles are held in the smoke arising from electrically heated trichloroethylene, which cleans them thoroughly in about 2 min. Before anodising they are dipped in a hot swill to remove traces of the smoke, which would contaminate the anodising bath.

29.7 Electrodeposition

Electrodeposition is carried out in electrochemical reactors that use electrical energy to extract metals from their compounds. The processes include electrowinning, electrorefining, electroplating and electroforming.

Electrowinning A mineral, or a compound prepared therefrom, is decomposed to a metal and (usually) a gas. The prime process in this class is the refining of aluminium: the Hall–Heroult process involves the electrolytic decomposition of alumina (prepared from bauxite) dissolved in molten cryolite (Na_3AlF_6) at about 1000°C. Aluminium is produced as a liquid at the cathode. The primary anode product, oxygen, reacts with the carbon anode to give mainly carbon dioxide, with a small proportion of carbon monoxide. The process is highly energy intensive.

Electrorefining An impure anode (perhaps a cathode from an electrowinning stage) is converted to a pure cathode. Copper is refined in this way. The process is much less energy intensive than electrowinning, mainly because of the lower voltages employed.

Electroplating This is the electrodeposition of a thin protective metal coating on another metal. Protection is obtained with the minimum amount of material and with a low expenditure of energy. The process has imperfections: for example, chromium plating has high porosity and poor adherence, particularly in extreme conditions.

Electroforming The aim in this extension of electroplating is to build up the plate so that the cheap substrate can be dissolved to leave an article of the plate metal which has mechanical integrity. Nickel electroforms are made in this way, but the process is particularly advantageous for refractory metals (e.g. tantalum, niobium). Thus, an article that would be costly and wasteful when made from bulk metal can be prepared directly.

29.7.1 Electroplating

In modern mass production processes automatic plating plants are regularly used for the deposition of nickel, chromium, brass, zinc, silver, copper and cadmium, and, in addition, for depositing composite coats of two metals, such as in nickel and silver plating, brassing and bronzing. In using plants of this kind, the product is extremely consistent, as the plater is relieved from manual duties and left free to devote his attention to the control of essential operations. The plant is usually arranged so that the various vats are placed in line one behind the other in proper sequence. Two conveyors travel slowly over the vats, the articles to be plated being loaded on insulated suspenders which are arranged in two or three rows on cross rods attached to the conveyor chains. The suspended articles pass in succession through cleaning and swilling vats before passing to the vats for deposition, and finally to the swilling and drying-out apparatus after plating is completed.

As each loaded cross rod arrives at a vat or tank it is lifted by an auxiliary fast-moving transfer chain, and lowered into the next vat along the line. After passing through drying apparatus, the suspenders are unloaded from the cross rods, the latter being automatically returned to the loading position for the work to continue. The cycle of operations in the case of automatic nickel-plating plants is usually as follows: hot alkaline electrolytic cleaner used cathodically at about 300 A/m^2; cold water swill; anodic etch in sulphuric acid in the case of iron and steel articles; agitated cold water swill; nickel plating; cold water swill; hot water swill; and drying.

29.7.1.1 Barrel plating

The barrel planting method of electroplating is automatic insofar as the actual deposition of metal is concerned, and the method is extremely efficient and economical within the limits of the capacity of the barrel and the size of the articles. Any difficulties experienced are usually traceable to overloading the barrel, which causes the articles to be carried round in a mass as the barrel revolves, or to attempts to plate articles having awkward shapes which shield one another in the mass and prevent regular deposition.

Plating barrels are available for use with all types of solutions, and excellent results can be obtained with such articles as screws, cycle and car fittings, and hooks and eyes. Some of the early machines were extremely clumsy, and offered great resistance to the flow of current, since when many improvements have been made, particularly in the use of rubber lined tanks, stoneware vats and containers of other non-metallic material.

An important advantage of barrel plating is that, as the articles are receiving the deposit of metal, they rub together and become burnished, while the deposit being consolidated by the rubbing is close-grained and durable. As in so many plating processes, the speed at which a plating barrel should revolve is critical and is governed by the class of article being plated and the diameter of the barrel.

Speeds vary between 30 and 45 rev/min, although in the case of large plating barrels the speeds may be as low as 10–15 rev/min. It sometimes happens that articles all of one shape cannot conveniently be plated in a barrel by themselves, and in such cases scrap metal pieces or steel balls of suitable size can also be put into the barrel.

Although barrel plating can be carried out with some solutions at 6 V, it is usually necessary to use voltages of 10–18 V. In the case of nickel a current of about 45–60 A at 10 V can be passed. A higher current is needed for 'brassing'.

Barrel plating was developed specially for handling small articles in quantities, and in many cases this process has a number of advantages over the original suspended anode method. In one of the best-known makes the anode insulated by a vulcanite cover is situated at the base of the barrel. The barrel itself is made of welded steel internally lined with vulcanite, and is mounted on a swivel arrangement, which permits immediate removal of the work by dumping it into a suitable container for transfer to the drier. Electrical contact is made through an insulated rod passing through the centre of the shaft and directly connected to the anode, the current returning to the plating rectifier by way of the barrel and framework of the machines. An important feature of this type of barrel is that it can be relined with vulcanite at very small cost, and it is also easily and quickly rinsed.

29.7.1.2 Polishing

The speed of a polishing mop or bob is a factor affecting efficiency and economy, and speeds between 30 and 45 m/s are the rule according to the article to be polished.

Variable-speed machines are necessary, so that when a mop or bob becomes worn, say from 30 to 20 cm in diameter, the peripheral speed may be increased. Too low a speed causes a mop to drag and the articles to become heated, and a burnishing rather than a polishing effect is produced. 'Polishing bobs' is the term used for all solid leather wheels, compress wheels, made up of sections of leather, canvas or felt, and solid felt. These are dressed with emery, and the felt bob is the type in most general use. Polishing mops consist of discs of cotton cloth, varying in size from 5 to 40 cm in diameter and held together at the centre by means of washers of leather or fibre.

29.8 Hydrogen and oxygen electrolysis

There are several processes by which oxygen and hydrogen are produced as a main product or by-product. The industrial uses are considerable.

Oxygen is used to a very large extent in metal working and metallurgical industries. In conjunction with acetylene or hydrogen, it is used for welding and cutting; with butane and propane, for tempering steel. Very large (tonnage) quantities are now used in the production of steel, in the gasification of coal and to a lesser extent in the oxidation of olefines. The use for medical purposes and in super-atmospheric aviation is relatively small in quantity but of high importance.

In the chemical industry oxygen or oxygen enriched air is sometimes used in place of air in the oxidation of ammonia to produce nitric acid in highly concentrated form. The electrolysis of water can therefore conveniently be used for the production of hydrogen for the synthesis of ammonia and of oxygen for the subsequent preparation of nitric acid. The nitric acid and ammonia together produce ammonium nitrate, an important fertiliser.

Hydrogen is used in a number of industries. In the food industry the production of margarine and cooking fats from liquid oils such as groundnut, cottonseed, whale, etc., is based upon the partial catalytic hydrogenation of these oils, which converts them into solids at normal temperatures. Sorbitol, a sugar used by diabetics and in the production of synthetic resins, is produced from glucose by hydrogenation. Ammonia, a basic chemical for the fertiliser industry, is produced in very large quantities annually by the catalytic combination under high pressure of hydrogen and nitrogen.

Large quantities are used in the fuel industries, mainly in the processing of mineral oils but also in the production of synthetic liquid fuels from coal.

Many other uses exist in the chemical industry, such as in the production of synthetic solvents and the hydrofining of benzole. In the metallurgical industry hydrogen alone or mixed with nitrogen is used to provide inert atmospheres for the annealing of alloy steels, and in the lamp industry it is used for the reduction of tungsten and molybdenum ores to the metals and in the subsequent manufacturing processes for the production of lamps. Large synchronous machines are sometimes cooled by the circulation of hydrogen in a closed system. Meteorological balloons are filled with the gas.

29.8.1 Process

Water dissociation to give small (equal) concentrations of the hydronium ion (H_3O^+) and the hydroxide ion (OH^-) is in accordance with the equation

$$2H_2O \rightarrow H_3O^+ + (OH)^{-\leftarrow}$$

Thus, it is possible in principle to decompose water electrolytically in an electrochemical reactor, to give hydrogen at the cathode and oxygen at the anode according to the reactions

cathode: $2H_3O^+ + 2e^{-\leftarrow}\rightarrow H_2 + H_2O$

anode: $4(OH)^- \rightarrow 2H_2O + O_2 + 4e^{-\leftarrow}$

Practical reactions are slightly more complex, as the electrolyte added to the water to improve its conductivity plays a role in the reaction.

In simple water electrolysis the hydrogen produced per coulomb is 0.010 45 mg/C. Thus, 1 kA produces 0.42 m^3/h (at 0°C, 760 mmHg, dry). The corresponding quantity of oxygen is 0.21 m^3/h. Stray current prevents full attainment of these outputs, but the current efficiency can be as high as 98–99%. The specific energy consumption depends on the voltage. The decomposition voltage of water is 1.23 V, but that to operate a hydrogen/oxygen cell exceeds this value, because of the ohmic resistance of the electrolyte and the overpotentials for the evolution of the gases at the electrodes. Solutions of caustic soda or potash are almost invariably used to lower the resistivity, using mild steel cathodes and nickel-plated anodes, which are immune to attack by the electrolyte or the gases. Soda and potash solutions are always prepared from the purest material: the caustic must have limited amounts of chlorides, sulphates and carbonates, which raise the resistance, while the former two may cause corrosion of the electrodes. The water feed must be pure (i.e. a very high resistivity). It is produced in a water still or by treating the raw water in an ion exchange plant, followed sometimes by active carbon treatment to remove oil traces, as when the source is a steam condensate.

29.8.2 Electrolysers

Water electrolysers are distinguished by electrode arrangement and gas pressure.

29.8.2.1 Tank type

The tank or unit-cell electrolyser has unipolar electrodes in one tank and connected in parallel. Normally several tanks are connected in series, facilitating extension and enabling a faulty tank to be bypassed.

29.8.2.2 Filter-press type

The filter-press type has bipolar electrodes arranged to act as a positive on one side and a negative on the other, with terminal connections at the end of the battery. *Figure 29.13* shows a section, the construction having some similarity to a filter press. The cell frames (1) are made of an electrically insulating caustic-resistant material and into them are fitted the diaphragms (9) and main electrodes (10). In order to maintain the optimum distance between the active faces of the electrodes, the main electrodes each carry a perforated auxiliary cathode (11) and a similar auxiliary anode (12). The various ports are sealed off from one another, and the whole assembly of a number of cells is made leak proof by means of the joints (2). The action is as follows.

Hydrogen evolved at the cathode (11) rises through the gas port (5), carrying with it some of the electrolyte. The mixture passes along the channel (4) common to the pack of cells, thence to an arrangement of gas washing and cooling drums mounted on the electrolyser, from which the gas passes to process. The electrolyte, after cooling, returns by a pipe system through a filter to the electrolyte channel (7). The ports (8) then ensure an electrolyte feed to each cell to replace that taken out with the gas. A similar circulation occurs with the oxygen from the anodes (12) through the ports (6) and channel (3) returning through (7) and (8). The current connections to the electrolyser are to the electrodes at the ends of the pack, which are extended beyond the bottoms of their respective cells for cable or bus-bar.

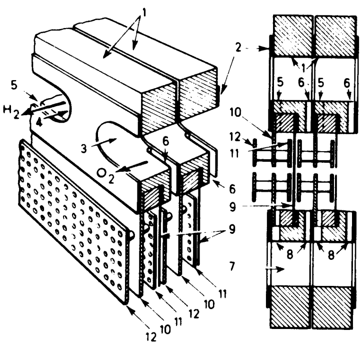


Figure 29.13 Section through a typical 'filter-press' electrolyser. 1, cell frame; 2, joint; 3, oxygen gas channel; 4, hydrogen gas channel; 5, hydrogen gas port; 6, oxygen gas port; 7, electrolyte channel; 8, electrolyte ports; 9, diaphragm; 10, main electrode; 11, auxiliary cathode; 12, auxiliary anode

It will be seen that, for a given output, a filter press electrolyser is constructed as one unit consisting of a number of cells of an appropriate size, clamped together as a pack, variations in the size and number of electrodes being made to suit the required electrical conditions. A tank type electrolyser, on the other hand, will consist of a number of unit cells connected in series externally, but here again variations in size and number can be made to suit the electrical conditions.

The filter-press electrolyser can produce in a single unit up to 500 m^3/h of hydrogen: the connections are simple; gases can be delivered at normal gas holder pressures without the need for separate boosters; the electrolyte is in an enclosed system and the constant circulation through a filter ensures thorough mixing and freedom from contamination; and the quantity of electrolyte is small compared with the output.

29.8.2.3 Pressure type

The pressure type electrolyser is a development of the filter-press electrolyser in which electrolysis is carried out under pressure. Lower cell voltages are obtained because of decreased blanketing effect of the smaller gas bubbles. The optimum pressure is about 30 atm, giving a 14% power advantage over that achieved at normal atmospheric pressure. Further, the cost of post-compression of the gases is less. The basic design is similar to that of the filter-press type, but adapted to work under pressure, having cells of circular section. Gas purity is a little less than for the unpressurised equipment, but is adequate for most industrial purposes.

29.8.3 Gas purity

The makers of atmospheric pressure filter-press electrolysers guarantee a purity of 99.7% for oxygen and 99.9% for hydrogen. A check is normally kept on the gas purities by drawing off samples into measuring burettes. In the case of oxygen the sample is then passed over copper, which combines with the oxygen and leaves the hydrogen impurity as residue; while in the case of hydrogen, a platinum spiral electrically heated to redness in the gas causes the oxygen impurity to be removed by a combination with twice its volume of hydrogen to form water, leaving the main bulk of hydrogen as residue.

Automatic gas analysers can also be fitted which give a continuous check on gas purities.

29.8.4 Plant arrangement

The electrolytic plant will usually comprise a source of direct current, the electrolyser, a tank for the storage and make-up of electrolyte and a transfer pump, a source of purified water, gas holders and compressors for hydrogen and oxygen, and pressure storage tanks. In the case of pressure electrolysers buffer tanks are provided in place of low-pressure gas holders. The direct connection of electrolysers working at low pressure to compressors is not recommended.

Depending on the required specification of the gases, dryers are installed before the gases are delivered to process. The capacities of gas holders, compressors and high-pressure storage vessels depend upon the user process and its demand cycle. The compressors are normally started and stopped automatically by pressure switches on the high-pressure storage vessels and capacity switches on the gas holders. Complete automatic control of electrolytic plants including variation of output to suit demand is possible but so far has not been attempted.

Because of the explosion hazard, electrolysers are housed in rooms isolated from the electrical conversion equipment by a gasproof wall, intercommunication between the two being by outside doors. Lighting of the electrolyser room can be either by pressurised fittings or by lamps placed outside the windows.

The electrical energy consumption for the production of given quantities of hydrogen and oxygen depends on the type of electrolyser and on the output per unit cell, because the cell voltage increases with the current density at the electrodes. For atmospheric types the operating voltage per cell is 1.9–2.1, corresponding to specific d.c. energy consumptions of 4.5–5.0 kW-h/m^3 of hydrogen. The selection of an electrolyser for a given duty is based on the capital cost of the installation and the cost of energy. Thus, where the energy cost is low, it is economic to work at a high current density (and therefore a high specific energy consumption).

References

1 PLETCHER, D. and WALSH, F. C., *Industrial Electrochemistry*, 2nd edition, Chapman and Hall, London (1990)

2 KAYE and LABY, *Tables of Physical and Chemical Constants*, 15th edition, Longman, London (1986)

3 CROMPTON, T. R., *Battery Reference Book*, 2nd edition, Butterworth-Heinemann (1995)

4 '*Towards zero emissions for road transport*', House of Lords Paper 117-1, London: The Stationery Office (1995)

5 MAZDA, F. F. (Ed.), '*Electronics Engineers Reference Book*', 6th edition, Butterworth-Heinemann (1989)

Section G

Transmission and Distribution

30

Overhead Lines

G Orawski EurIng, BSc(Eng) Hons, CEng, FIEE
Formerly of Balfour Beatty Power Construction Ltd

Contents

30.1 General 30/3

30.2 Conductors and earth wires 30/3
  30.2.1 Materials 30/3
  30.2.2 Nomenclature 30/3
  30.2.3 Mechanical characteristics 30/3
  30.2.4 Sag and tension 30/5

30.3 Conductor fittings 30/7

30.4 Electrical characteristics 30/8
  30.4.1 Bundled conductors 30/8
  30.4.2 Electrical parameters 30/8
  30.4.3 Voltage-gradient effects 30/9
  30.4.4 Distribution lines 30/9

30.5 Insulators 30/10
  30.5.1 Types 30/10
  30.5.2 Selection 30/12
  30.5.3 Pollution 30/12
  30.5.4 Voltage distribution over insulator strings 30/12

30.6 Supports 30/13
  30.6.1 Materials 30/13
  30.6.2 Configurations 30/13
  30.6.3 Tower geometry 30/15
  30.6.4 Foundations 30/15

30.7 Lightning 30/16
  30.7.1 Mechanisms of insulation flashover 30/16
  30.7.2 Lightning performance 30/16

30.8 Loadings 30/16

30.1 General

The overhead line is the cheapest form of transmission and distribution of electrical energy. Line design and construction involve several engineering disciplines (electrical, civil, mechanical, structural, etc.), and must conform to national and international specifications, regulations and standards. These refer to conductor size and tension, minimum clearance to ground, stresses in supports and foundations, insulation levels, etc. Lines must operate in conditions of large temperature change and in still air and gales, and (in non-tropical climates) may have ice formation on conductors and supports.

The structural design of a d.c. line does not differ essentially from that of an a.c. line, but electrical features related to frequency (such as inductive, capacitive and skin effects) do not apply in the d.c. case under normal operating conditions.

30.2 Conductors and earth wires

30.2.1 Materials

Many years of operating experience, and the costs as affected by metal market trends, have combined to favour aluminium based conductors[1] (see also Chapter 5, Section 5.1.2). Copper and cadmium–copper are seldom used, even for distribution lines.

The intractability of large solid conductors has led to the almost exclusive use of stranded conductors (in spite of their greater cost), which have a larger diameter than equivalent solid circular conductors. With strands of diameter *d*, the outside diameter of uniformly stranded conductors is:

No. of strands	3	7	19	37	61
Overall diameter	2.15d	3d	5d	7d	9d

Aluminium based conductors in normal use are categorised as

AAC	All-Aluminium Conductors (BS 215, IEC 207)
ACSR	Aluminium Conductors Steel Reinforced (BS 215, IEC 209)
AAAC	All-Aluminium Alloy Conductors (BS 3242, IEC 208)
AACSR	All-Aluminium Alloy Conductors Steel Reinforced (IEC 210)
ACAR	Aluminium Conductors Alloy Reinforced

Figure 30.1(*a*) illustrates typical strandings of ACSR. The conductor with an outer layer of segmented strands has a smooth surface and a slightly reduced diameter for the same electrical area.

Any of the above types can be used as an earth wire, although generally preference is given to steel. In areas where high short-circuit currents are anticipated, one layer (or more) of aluminium or aluminium alloy strands is added.

For distribution lines, any of the above conductor types can be used, but in addition mention must be made of Aerial Bundled Conductors (ABC), commercially available for systems up to 30 kV with greatly reduced environmental impact.[2] They consist essentially of four PVC covered cables bunched together and strung overhead rather than underground.

Recently, there has been an increased tendency to favour earth wires with incorporated optical fibres (*Figure 30.1*(*b*)) which can be used for communication purposes. These are often referred to as optical ground wires (OPGW) (Chapter 27, Section 27.5.2). Also, developed for telecommunication purposes, all-dielectric self-supporting (ADSS) aerial optic cables with a high strength non-metallic component as support are occasionally installed. These can in fact be added at a later date (subject to a check on the mechanical strength of the structures). In such cases, a knowledge of the electric field conditions along the overhead line is required to locate the ADSS cable in a zone of minimum field strength.

30.2.2 Nomenclature

There is as yet no international agreement on nomenclature. In the UK conductors were referred to the approximate area in (inches)^2 of a copper conductor having the same conductance. Nowadays, aluminium based conductors are referred to their nominal aluminium area. Thus, ACSR with 54 aluminium strands surrounding seven steel strands, all strands of diameter $d = 3.18$ mm (designated: 54/7/3.18; alu. area = 428.9 mm^2, steel area = 55.6 mm^2) is described as 400 mm^2 nominal aluminium area. In France the total area (485 mm^2) is quoted; in Germany the aluminium and steel areas are quoted (429/56); whereas in Canada and the USA the area is stated in thousands of circular mils (1000 circular mils = 0.507 mm^2).

Code names using animal, bird or flower names, etc., are also used. Thus, the 400 mm^2 nominal aluminium area ACSR is known as ‘Zebra’.

30.2.3 Mechanical characteristics

The choice of a conductor from the mechanical viewpoint depends on external loading conditions (such as wind speed, ice loading and ambient temperature), and on internal characteristics (such as stranding, modulus of elasticity, thermal expansion and creep). For lines of 33 kV and over, hard-drawn AAC is not likely to be adopted; economics and creep behaviour lead to the choice of ACSR or AAAC. As can be seen from *Figure 30.1*(*a*), a wide range of breaking strength/weight ratio can be achieved by modifying the aluminium and steel content, and by using aluminium alloy strands where conditions demand. *Table 30.1* gives typical properties of ACSR conductors.

As a result of many investigations into the economics of the complete overhead line (system economics) and the improved ability of manufacturers to produce alloy at competitive prices, the tendency to use AAAC on the British grid has increased. The overhead line engineer should now familiarise him/herself with names such as Totara, Rubus, Araucaria for AAAC and Collybia for ACAR.[3] These are certainly favoured for refurbishment work, since they quite often permit increased transmission capacity for the same wayleaves. Probabilistic techniques applied to existing lines are one of the many factors enabling relevant decisions to be taken.

Modulus of elasticity For a complete ACSR the modulus of elasticity can be estimated from

$$E = [(70r + 200)/(r + 1)] \times 10^3 (\mathrm{MN/m^2})$$

where r is the ratio aluminium area/steel area.

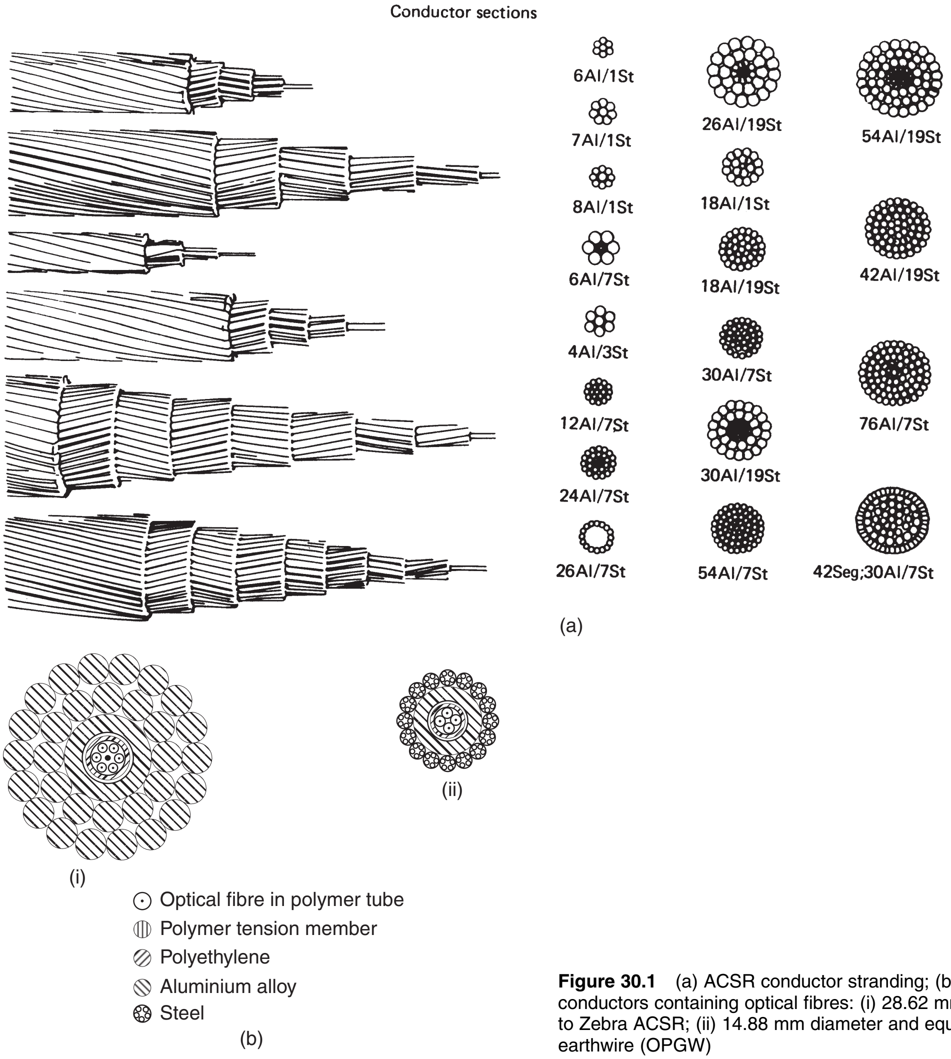

Conductor sections
6Al/1St
7Al/1St
8Al/1St
6Al/7St
4Al/3St
12Al/7St
24Al/7St
26Al/7St
26Al/19St
18Al/1St
18Al/19St
30Al/7St
30Al/19St
54Al/7St
54Al/19St
42Al/19St
76Al/7St
42Seg;30Al/7St
(a)
(i)
(ii)
⊙ Optical fibre in polymer tube
◍ Polymer tension member
▨ Polyethylene
▧ Aluminium alloy
❂ Steel
(b)

Figure 30.1 (a) ACSR conductor stranding; (b) typical cross-section of conductors containing optical fibres: (i) 28.62 mm diameter and equivalent to Zebra ACSR; (ii) 14.88 mm diameter and equivalent to half-inch-steel earthwire (OPGW)

Linear expansion The coefficient of linear expansion of an ACSR conductor per degree Celsius is given by

$$\alpha = 10^{-6}[(23r + 32.6)/(r + 2.83)]$$

Some useful reference data are given in *Table 30.2*.

Oscillation phenomena Exposed to wind and ice loading, overhead line conductors are subject to oscillations that affect the design and application of the conductor fittings.

Aeolian vibrations These are typically in the range 8–40 Hz, occasionally higher, and are generated by wind speeds of 2–40 km/h. As a laminar air flow is involved, certain regions of terrain are prone to develop the aeolian phenomenon. The consequent vibration causes strand breakdown by fatigue.

As a result of international enquiries covering thousands of kilometres-years of lines, the concept of EDS (everyday stress defined as the general order of stress which exists over the larger proportion of the life of a conductor, and therefore,

Table 30.1 Typical properties of ACSR conductors

Code name	*Stranding*	*Aluminium area* (mm^2)	*Steel area* (mm^2)	*Diameter* (mm)	*Mass* (kg/km)	*Breaking load* (kN)	*Resistance at* 25 °C (Ω/km)
Horse	12/7/2.79	73.4	42.8	13.95	538	61.2	0.3936
Lynx	30/7/2.79	183.4	42.8	19.53	842	79.8	0.1441
Zebra	54/7/3.18	428.9	55.6	28.62	1621	131.9	0.0674
Dove	26/3.72 + 7/2.89	282	45.9	23.55	1137	99.88	0.1024

Table 30.2 Characteristics of conductor materials at 20°C

Property	*Units*	*Annealed copper*	*Hard-drawn copper*	*Cadmium–copper*	*Hard-drawn aluminium*	*Aluminium alloy (BS 3242)*	*Galvanised steel*
Relative conductivity	%	100	97 (avg.)	79.2 (min.)	61 (min.)	53.5	—
Volumetric resistivity	Ω-mm^2/m	0.017 24 (std.)	0.017 71 (avg.)	0.021 77 (max.)	0.028 26 (max.)	0.032 2 (std.)	—
Mass resistivity	Ω-kg/km	0.153 28	0.157 41	0.194 72	0.076 40	0.086 94	—
Resistance at 20°C	Ω-mm^2/km	17.241	17.71	21.77	28.26	32.2	—
Density	kg/m^3	8890	8890	8945	2703	2700	7780
Mass	kg/mm^2/km	8.89	8.89	8.945	2.703	2.70	7.78
Resistance temperature coefficient at 20°C	per °C	0.003 93	0.003 81	0.003 10	0.004 03	0.0036	—
Coefficient of linear expansion	per °C	17×10^{-6}	17×10^{-6}	17×10^{-6}	23×10^{-6}	23×10^{-6}	11.5×10^{-6}
Ultimate tensile stress (approx.)	MN/m^2	255	420	635	165	300	1350
Modulus of elasticity	MN/m^2	100 000	125 000	125 000	70 000	70 000	200 000

at or around mean temperature with little or no superimposed load) was introduced as a means of ensuring a longer conductor life before strand damage. Typical values as proportion of breaking loads are: for ACSR: 18 to 24%; for Aluminium: 17% for Aluminium alloy: 18%. The subject is now well documented and much more sophisticated techniques can be used[4,5,6] to refine these values. The concept of EDS should not be underestimated as it constitutes one of the bases for 'sag and tension' calculations.

Subspan oscillations These affect only bundled conductors (Section 30.4.1). They occur at a frequency of 0.5–5 Hz,[7] and wind speeds of 15–65 km/h. Their intensity is considerably reduced if the ratio between subconductor spacing and conductor diameter exceeds 16.

Galloping This involves complete spans in a fundamental mode. Amplitudes reaching the value of the sag have been recorded. Bundled conductors may be more prone to galloping than single conductors, but the latter have been known to gallop.[8] It is now generally accepted that near-freezing temperatures are required for galloping to occur with conductor diameters up to 35 mm. Large conductors can gallop under specific conditions due to aerodynamic instability, even in the absence of ice.

Creep Aluminium based conductors are subject to a permanent non-elastic elongation ('creep'), which must be predicted with adequate precision in order that ground clearance regulations may be satisfied. Creep changes the conductor length, tension and sag, and this in turn affects the creep, as does the conductor temperature.[9,10]

30.2.4 Sag and tension

Almost all the elements of an overhead line are based on the mechanical loading of the conductors. The loading includes the effect of stringing tension, wind pressure, temperature and ice formation.

The configuration of a conductor between supports approximates a catenary curve, but for most purposes it may be taken as a parabola. The catenary has a constant mass per unit length *along the conductor*; the parabolic curve assumes a constant mass along the *straight horizontal line* between the points of support. The following nomenclature is employed.

Quantity	*Symbol*	*Unit* *SI*	*Imperial*
Span length	L	m	ft
Half span length	l	m	ft
Conductor diameter	d	m	ft
Conductor section	a	m^2	ft^2
Thickness of ice	t	m	ft
Tension			
Total	T	N	lb-f
Horizontal	T_h	N	lb-f
Stress	f	N/m^2	lb-f/ft^2
Maximum sag (at midspan)	s	m	ft
Weight			
Conductor	w_c	N/m	lb-f/ft
Ice	w_i	N/m	lb-f/ft
Wind pressure	p	N/m^2	lb-f/ft^2
Wind loading	w_w	N/m	lb-f/ft
Total loading	w	N/m	lb-f/ft

lb-f = pound force.

The stress is $f = T/a$, the wind loading is $w_w = p_2(d + 2t)$ and the 'total' loading is $w = \sqrt{[(w_c + w_i)^2 + w_w^2]}$ as the conductor and ice loadings are directed vertically downward, whereas the wind loading is directed horizontally and acts on the projected area of the conductor and its ice loading (if applicable).

The catenary equation for a conductor between supports at the same height is expressed in terms of the distance x from the centre of the span (point of maximum sag) and of y the conductor height above this point (*Figure 30.2*(*a*)). Then

$$y = (T_h/w)[\cosh(wx/T_h) - 1] \Leftarrow$$

is the catenary equation. Putting $x = L/2$ gives y at either support and, therefore, the sag s. The hyperbolic cosine term can be expanded into a series. For typical low-sag cases all terms but the first are negligible. Furthermore, with the assumption of a uniform conductor tension throughout, the equation reduces to $y = wx^2/2T_h$, representing a parabola. Therefore, for a span L the sag is

$$s = wL^2/8T_h$$

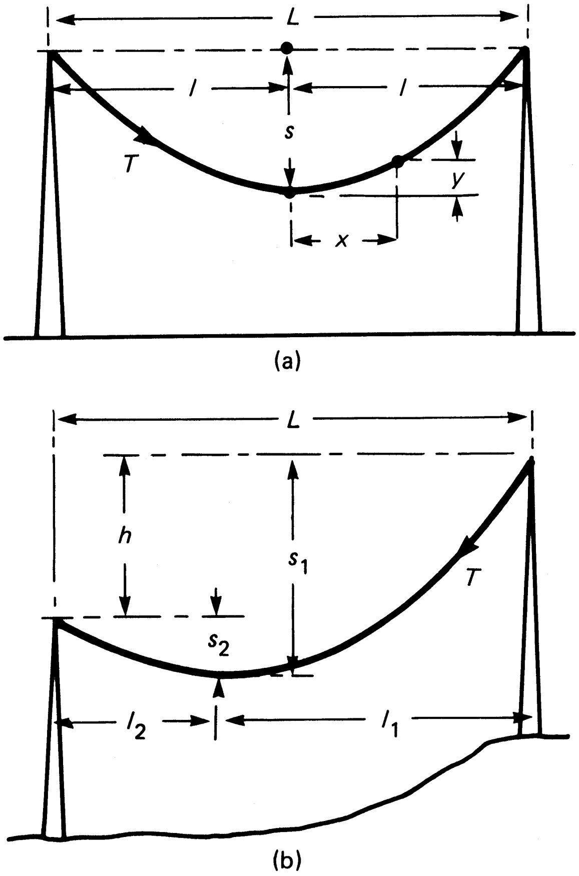


Figure 30.2 Sag and tension

and the length of conductor in a half-span is

$$l_c = l + w^2l^3/6T_h^2 = l + 2s^2/3l$$

For supports at different heights (*Figure 30.2(b)*) that differ by *h,* the distances of the supports from the point of maximum sag are

$$l_1, l_2 = \frac{1}{2}[L \pm 2hT/wL]$$

with the parabolic assumption. For large values of *h* the value of l_2 may be negative, indicating that the lowest point of the conductor is outside the span, on the left of the lowest support. Such a case would involve an upward component of pull on the lower support, not admissible with suspension insulators.[38]

30.2.4.1 Change of state

If after erection the conductor temperature rises (because of I^2R loss or of a rise in ambient temperature), the conductor expands, increasing the sag; but at the same time reduction of the tension allows the conductor to contract elastically. Furthermore, if the loading increases (owing, for example, to wind pressure and/or ice), the tension rises and the conductor stretches. Analysis of these opposing tendencies leads to a cubic equation relating tension, temperature, loading and elasticity. For two sets of conditions (subscripts 1 and 2)

$$\frac{w_2^2L^3}{24f_2^2a^2} = \frac{w_1^2L^3}{24f_1^2a^2} + \frac{f_2 - f_1}{E} - (\theta_1 - \theta_2)\alpha$$

Here E is the modulus of elasticity, θ is the temperature (°C) and α is the coefficient of linear thermal expansion per degree Celsius. The above equation illustrates the relationships between the various factors affecting the conductor behaviour under 'change of state' conditions. With the availability and extensive use of computers, much more complex and more accurate catenary equations (hyperbolic functions) can be used. Surprisingly programming is easier with those equations.

The maximum permissible stress f_2 (or tension f_2a) that will occur under the most onerous conditions and at a low temperature θ_2 is usually accepted and known. However in tropical countries the EDS could be a starting point for the calculations. The stress f_1 at which the line must be strung can then be calculated. 'Stringing charts' for a range of spans and temperatures can then be prepared for use by linesmen in the field. Such charts are readily prepared by computers. For convenience, for spans up to 300 m the parabolic assumption can still be used, but for longer spans, the catenary equations should be preferred.

30.2.4.2 Equivalent span

As an actual overhead line comprises a series of spans of not necessarily equal length, supported by suspension insulator sets and with tension insulators at the ends of a section, it is generally assumed that the behaviour of a section is that given by a series of equal equivalent spans, each given by

$$L_{eq} = \sqrt{[\Sigma L_i^3/\Sigma L_i]}$$

where L_i is an individual span length. It is possible to improve on this estimate by use of a complex computer program.

30.2.4.3 Creep

As a result of international discussions, two distinct methods of evaluation have been adopted. The first one[9,10] is based on results of laboratory tests on complete conductors and the empirical expression

$$\epsilon = K \cdot f^{\beta} \cdot \exp(\phi\theta) \cdot t^{\gamma/f^{\delta}}$$

for the creep extension ϵ (in mm/km) in terms of conductor stress f (in kg/mm^2) and temperature θ (in °C), time t (in h), a constant K and creep indices β, ϕ, γ and δ. Typical values of K and the creep indices are given in *Table 30.3* for ACSR conductors. Evaluation of the expression for the range of operating conditions requires a sophisticated computer program. *Figure 30.3* shows a predicted creep–time curve, a series of measured values and a 'best fit' for them, in the case of a Zebra 54/7/3.18 conductor.

The second method[10] calls for laboratory tests on single wires destranded from the conductors, or even on the wires which will be used to form the conductors. In this case, the mathematical model must allow separately for the metallurgical creep and for the geometric settlement, both contributing to the inelastic elongation.

Table 30.3 Creep coefficients for ACSR conductors

No. of strands		*Aluminium/steel area ratio*	*Process**	*Coefficients*				
Al	*Steel*			$\kappa\varsigma$	$\phi\varsigma$	$\beta\varsigma$	$\gamma\varsigma$	$\delta\varsigma$
54	7	7.71	HR	1.1	0.0175	2.155	0.342	0.2127
			EP	1.6	0.0171	1.418	0.377	0.1876
48	7	11.4	HR	3.0	0.0100	1.887	0.165	0.1116
30	7	4.28	EP	2.2	0.0107	1.375	0.183	0.0365
26	7	6.16	HR	1.9	0.0235	1.830	0.229	0.08021
24	7	7.74	HR	1.6	0.0235	1.882	0.186	0.00771
18	1	18	EP	1.2	0.0230	1.503	0.332	0.1331
12	7	1.71	HR	0.66	0.0115	1.884	0.273	0.1474

* Industrial processing of aluminium rod: HR, hot-rolled; EP, extruded or Properzi.

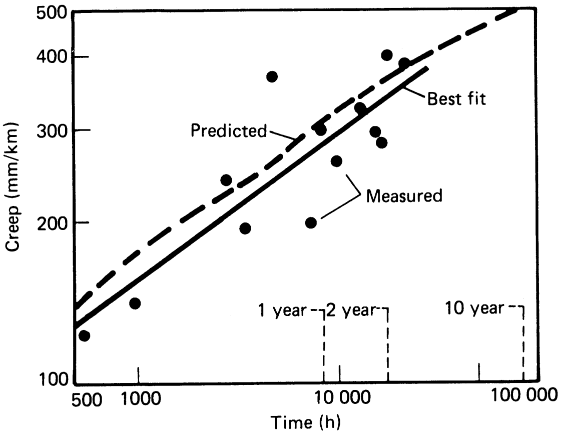


Figure 30.3 Typical creep/time relations

Comparison of results for a particular conductor has shown acceptable agreement.

30.3 Conductor fittings

Suspension clamp *Figure 30.4* illustrates a typical UK design, light and suitably profiled to limit the effects of combined static tension, compression and bending, and stresses due to dynamic bending. Positioning the axis of rotation on the axis of the conductor is considered desirable.

Tension and mid-span joints A bolted clamp or a cone-type grip is usually adequate for monometallic conductors of small size. For composite conductors (especially when greased) compression clamps must be used for joints and conductor ends.

Vibration dampers These absorb vibration energy. The Stockbridge damper (*Figure 30.5*) is well known. Other

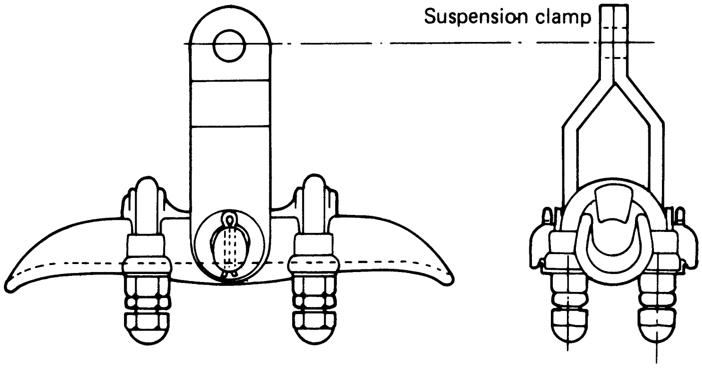


Figure 30.4 Suspension clamp

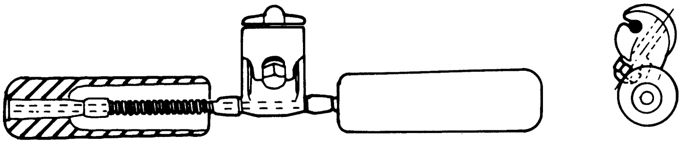


Figure 30.5 Stockbridge vibration damper

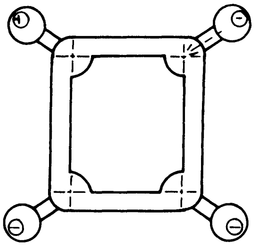


Figure 30.6 Schematised spacer-damper

types, such as ‘bretelles’, festoons, Elgra, etc., have also proved successful.

Spacers and spacer-dampers These maintain bundled conductors in proper configuration. Experience has shown the desirability of incorporating a damping element in the fitting which then becomes a spacer-damper. *Figure 30.6* illustrates such a device for a quad bundle. The specifications for such a fitting are quite complex since the damping element should desirably be efficient to damp aeolian vibrations and subspan oscillations.

Anti-galloping devices When fitted to bundled conductors, significant reduction in gallop amplitude is achieved, reducing outages due to interphase flashover.

30.4 Electrical characteristics

Electrical characteristics are concerned with voltage regulation and current carrying capability. The electric field strength (or voltage gradient) at the conductor surfaces affects corona and radio interference phenomena.

Voltage regulation This must be maintained within specified limits (normally 5–12%). The voltage drop depends on the electrical line parameters of resistance, inductance and capacitance, the two latter being influenced by the geometry of the support structures and the frequency of the system. For power lines of length up to about 80 km the capacitance is usually ignored in electrical calculations.

Current rating Aluminium is subject to annealing at temperatures exceeding 75°C. Hence 75°C is usually accepted as an upper limit for normal conditions. Higher temperatures can only be tolerated for short periods.

For design purposes, current rating is assessed from a heat balance equation: i.e. I^2R loss + heat absorbed from solar radiation = heat loss by radiation + heat loss by convection. As the current rating (which has no unique value) depends on local meteorological conditions, knowledge of these conditions is paramount. Recently, probabilistic techniques[11] have been introduced for the definition of current ratings, recognising that wind speeds, temperatures and solar radiation could be described by random functions. During a period of testing which preceded the calibration of the mathematical models, in the UK it was interesting to discover that the heat gained from solar absorption contributed little (if at all) to the heat input simply because higher ambient temperatures led to increased air convection currents which cancelled the effect of the heat absorption. On this basis, system operators in several countries are introducing special techniques to adapt the thermal rating of the conductors to the changing ambient conditions.

Radio interference Considerations of amenity require the radio noise from overhead lines to be kept within acceptable limits of signal/noise ratio in the vicinity of the line.

30.4.1 Bundled conductors

At operating frequency the inductive reactance of an overhead line with an equivalent spacing s' and a conductor system of geometric mean radius r' has the form $X = K \ln(s'/r')$. An increase of r' reduces the inductance (and therefore increases the capacitance) of the line; it also lowers the voltage gradient for a given working voltage. High-voltage lines can utilise bundles of two or more spaced subconductors to give an effective increase in r'. For a bundle of n subconductors each of radius r, arranged symmetrically around the circumference of a circle of radius R, the geometric mean radius of the assembly is

$$r_i = [nrR^{n-1}]^{1/n}$$

Bundled conductors have the following advantages:

(1) reduced inductance and increased capacitance, improving ‘surge impedance loading’, i.e. raising the power transmission capability of a long line;
(2) greater thermal rating because of the greater cooling surface area compared with that of a single conductor with the same total cross-sectional area; and
(3) lower surface voltage gradient, so reducing corona loss and radio interference.

The main disadvantage is the sometimes unsatisfactory aerodynamic performance: interbundle oscillations can occur between subconductor spacers. The effect can be controlled by judicious arrangement of spacers and spacer-dampers.[7]

30.4.2 Electrical parameters

Resistance Because of the lay in strands, the current flow in a conductor is helical, developing an axial magnetic field. If the conductor has a steel core, the magnetic loss therein may increase the effective resistance by 4–5%. The proximity effect is usually negligible, but the skin effect may produce a small increase in the effective resistance.

Inductance For a three-phase line with asymmetrical phase spacing and transposition, each phase comprising (a) a single conductor of radius r and geometric mean radius r' ($\simeq 0.78r$), or (b) a bundle of geometric mean radius $r' = r_i$, the line-to-neutral inductance is

$$L = [0.20 \ln(s'/r') + K] \text{(mH/km)}$$

where $s' = (s_{ab} \cdot s_{bc} \cdot s_{ca})^{1/3}$ is the geometric mean spacing between phases ab, bc and ca. K is a correction factor for the steel core and is equal to zero when there is no steel.[12]

Capacitance An isolated three-phase asymmetric transposed line with single conductors per phase of overall radius r has a line-to-neutral capacitance

$$C = 1/[18 \ln(s'/r)] (\mu\text{F/km})$$

As the earth is a conductor, it influences the line capacitance. An adjusted value of C is obtained by modifying the logarithmic term to $[\ln(s'/r) - \ln(s'/s_i)]$, where $s_i = (s_{aa}' \cdot s_{bb}' \cdot s_{cc}')^{1/3}$ is the geometric mean of the respective distances between conductors abc and their images a′b′c′ across the earth plane. A similar (but much more complex) expression is required for a bundle-conductor assembly.[13]

30.4.3 Voltage-gradient effects

For a single conductor of radius r (in centimetres) operating at a phase voltage V_{ph} (in kilovolts), the surface electric field strength is

$$E = 1.8 V_{ph} C/r \,(\text{kV/cm})$$

where C is the capacitance of the conductor (in pF/cm). In a bundle conductor the voltage gradient is not the same at all points on a subconductor surface.[13] For n subconductors each of radius r (in centimetres) spaced a distance a apart, the maximum voltage gradient (in r.m.s. kV/cm) is given by

$$E = \frac{1.8 V_{ph} C}{nr}\left[1 + \frac{2(n-1)\sin(\pi/n)}{a/r}\right]$$

Corona The corona onset voltage gradient[12] can be estimated from the Peek formula

$$E_c = 32\delta\left[1 + 0.308/\sqrt{(\delta r)}\right](\text{kV/cm})$$

where δ is the relative air density $3.92b/(\theta + 273)$, b is the barometric pressure (in cmHg), θ is the temperature (in degrees Celsius) and r is the conductor radius (in centimetres). The number 32 is an average value of the peak breakdown electric strength of air, corresponding to about 22 r.m.s. kV/cm. Thus, a surface voltage gradient of 18–19 kV/cm r.m.s. represents an acceptable upper limit.

Radio noise Electromagnetic fields, including interference fields, are generally expressed either in microvolts per metre (μV/m), in millivolts per metre (mV/m) or in decibels, as follows:

Field in decibels = 20 log (field in μV/m)

Thus, an interference level expressed as 46 dB corresponds to an electric field of 200 μV/m (i.e. 46 = 20 log (200/1)). Consequently, the radio noise from overhead lines is expressed in decibels above one microvolt per metre (dB/1 μV/m). The main source of radio interference (RI) is conductor corona which depends on voltage gradient. Other factors come into play. There are many methods for calculating RI, both analytical and empirical,[39] but the following empirical formula is often accepted in which RI is the radio noise to be calculated (in dB/1 μV/m), E_a is the average surface gradient, r is the radius of the subconductor, n is the number of subconductors in the bundle, h is the altitude of the line, D is the distance of the line at which the noise is measured (typically 15–30 m) and f is the frequency at which noise is calculated (typically 0.5 or 1.0 MHz in the medium-wave band).

$$\text{RI} = (\text{RI})_0 + 3.8(E_a - E_0) + 40\log\frac{r}{r_0} + 10\log\frac{n}{n_0} + 30\log\frac{D_0}{D} + 20\log\frac{1+f_0^2}{1+f^2} + \frac{h-h_0}{300}$$

The suffix ‘0’ refers to the same parameters as obtained from a known line with fairly close characteristics.[14] For example, a 500 kV line is designed so that $E_a = 16$ kV/cm, $r = 1.6$ cm, $n = 4$, $D = 30$ m, $f = 1.0$ MHz, and $h = 600$ m.

Analysis of data for a comparable construction yields $(\text{RI})_0 = 57$, $E_0 = 17$ kV/cm, $r_0 = 2.14$, $n_0 = 2$, $D_0 = 15$, $f_0 = 0.5$ MHz, and $h_0 = 300$ m.

Thus,

$$\begin{aligned}\text{RI} &= 57 + 3.8(16-17) + 40\log\frac{1.6}{2.14} + 10\log\frac{4}{2} \\ &\quad + 30\log\frac{15}{30} + 20\log\frac{1+0.5^2}{1+1^2} + \frac{600-300}{300} \\ &= 57 - 3.8 - 5.05 + 3 - 9.03 - 4.1 + 1 \\ &= 39.02 \text{ dB above } 1\ \mu\text{V/m}.\end{aligned}$$

Before the design of the line is accepted, it is necessary to obtain the strength of the radio signal received at the position of the line. The decision will be governed by the desired quality of reception, on the basis of the following table:

Class of reception	*Quality of reception*	*Approximate signal/noise ratio*
A	Background undetectable	32
B	Background detectable	27
C	Background evident	22
D	Background evident but speech still understood	16

If, for example, a class of reception B is desired, in the case of the line mentioned above, the signal strength received at 30 m from the axis of the line should be 39.02 + 27 = 66.02 dB/1 μV/m under fair weather conditions. If this is not the case, another look at the design or at the route of the line would be justified.

Under foul weather conditions, the radio noise from the line would be some 10 dB/1 μV/m greater.

30.4.4 Distribution lines

Voltage regulation For distribution lines of short length, the capacitance is usually ignored and the voltage drop is given by:

$$\Delta V = I \cdot l \cdot (R\cos\phi + X\sin\phi)$$

and the regulation is

$$(\Delta V/V)100\%$$

where I is the current in the line, l is the length of the line, R is the resistance per unit length, X is the inductive reactance per unit length and V is the nominal voltage. Unfortunately, for conductors greater than approximately 100 mm^2 the reductions in resistance and reactance as a function of increased conductor sizes are small.[2] *Table 30.4* and *Figure 30.7* illustrate this point. Hence, using conductors with a cross-section above 2×10^2 mm^2 is unlikely to have a serious impact on regulation.

Table 30.4 The inductive reactances (Ω/km) all-aluminium alloy conductors (AAAC)

	Equivalent aluminium area (mm^2)					
	25	*75*	*100*	*150*	*200*	*250*
Stranding (No./mm)	7/2.34	7/4.04	7/4.65	19/3.48	37/2.87	37/3.23
Resistance at 20 °C	1.093	0.3669	0.2769	0.1831	0.1385	0.1093
Resistance at 75 °C	1.295	0.4347	0.3280	0.2169	0.1641	0.1295
0.3 m ES (415 V)	0.300	0.265	0.256	0.240	0.230	0.223
1.4 m ES (11 kV)	0.396	0.362	0.353	0.337	0.327	0.319
3.0 m ES (66 kV)	0.444	0.410	0.401	0.385	0.375	0.367
4.9 m ES (132 kV)	0.475	0.441	0.432	0.415	0.406	0.398

ES, equivalent spacing.

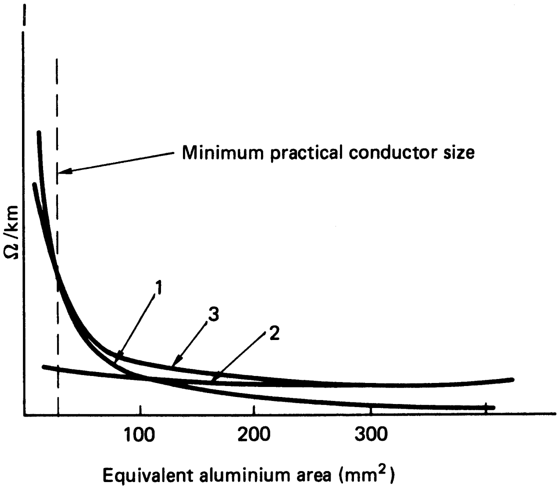


Figure 30.7 Plots of (1) the resistance, (2) the reactance and (3) $R \cos \phi + X \sin \phi$ for the data given in *Table 30.4*. Drawn for 11 kV line and $\cos \phi = 0.8$ (typical only)

Quick method of design When the voltage regulation is likely to be the controlling factor in the selection of conductors, then the technique of kVA · km is useful.[2] It can be shown that, for 10% voltage drop:

$$\text{kVA} \times l = 100\text{V}^2/(R \cos \phi + X \sin \phi)$$

i.e. if the regulation is not to exceed 10%, then the kVA loading of the line multiplied by its length (kVA · km) must be equal to the right-hand side of the above equation. *Table 30.5* illustrates some answers.

30.5 Insulators

Air is ubiquitous and is an excellent insulator. The performance of air gaps is well known as regards breakdown as a function of electrode geometry, so that clearances between phase conductors and from conductors to ground can be determined. However, at line supports it is necessary to provide man-made insulators to maintain adequate air gaps and to provide mechanical support.

The electric stresses applied to an overhead line are: (1) internal stresses resulting from (a) normal power frequency voltage, (b) power frequency overvoltage caused by faults or abnormal conditions, and (c) impulse overvoltage (switching); and (2) externally applied stresses due to lightning. Insulation requirements strongly affect the support tower geometry.

The insulation of an overhead line cannot be determined in isolation. It has to be co-ordinated with the system of which it is a part[15,16] and meet specific requirements.[17,18] (See also Chapter 7, Section 7.2.3.)

30.5.1 Types

The most common materials are porcelain (ceramic) and glass. Porcelain insulators are surface glazed, usually in brown but occasionally in other colours (e.g. blue) for reasons of amenity. With glass, the toughening process develops two zones, an outer layer in compression, an inner layer in tension. Damage to the outer skin usually results in an 'explosion' of the unit. This feature is sometimes

Table 30.5 The kVA · km product for a 10% voltage drop and cos ϕ–0.9 for all-aluminium alloy conductors (AAAC)

	Equivalent aluminium area (mm^2)				
	25	*75*	*100*	*200*	*250*
Current (A)					
Temperate	158	311	370	573	664
Tropical	131	253	299	453	521
415 V	13.3	34.0	42.3	69.5	80.6
11 kV	9042	22 038	26 935	41 706	47 310
66 kV	320 515	764 321	926 600	1 400 638	1 574 662
132 kV	—*	—*	3 603 404	5 370 515	6 006 871

*Voltage gradient exceeds 18 kV/Cm. Conductor not recommended for this voltage.

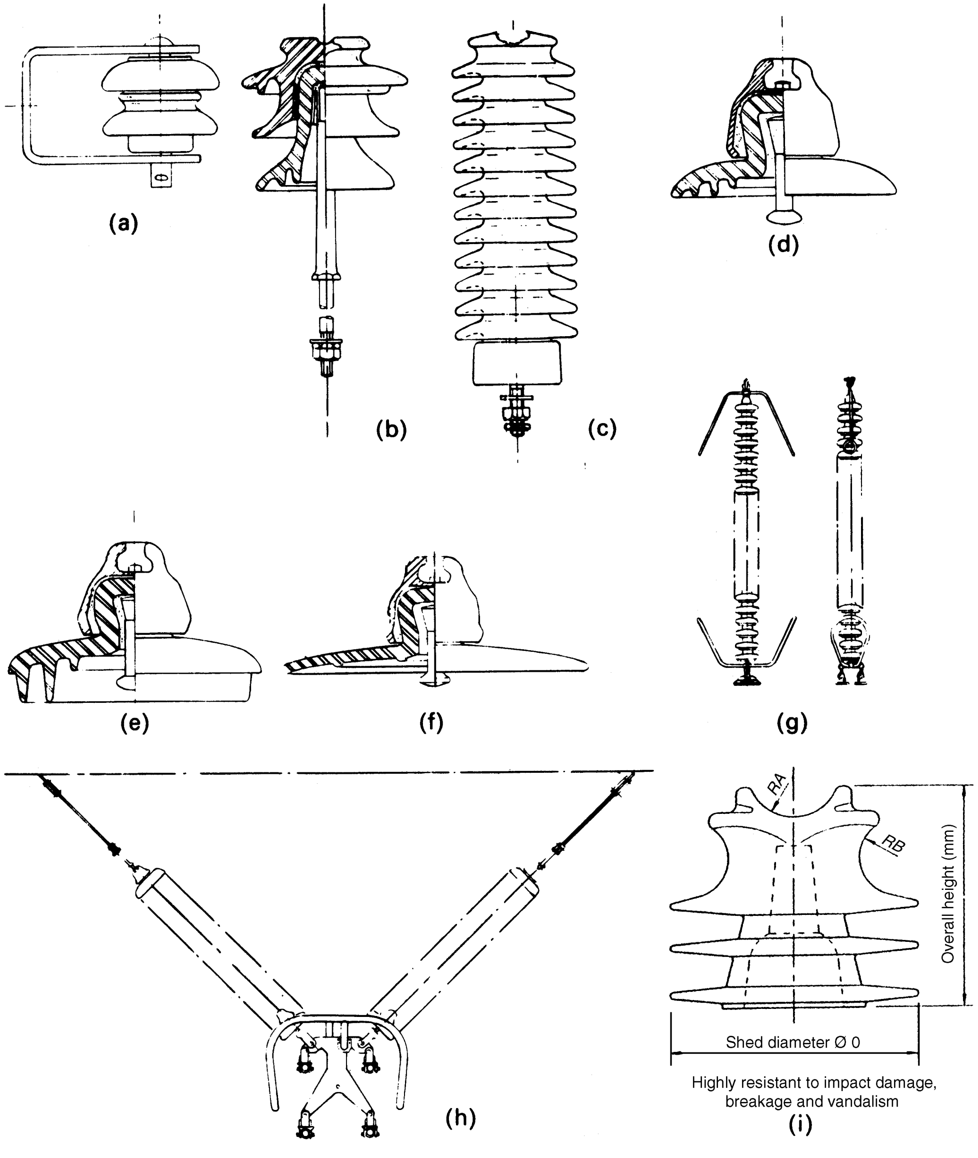
regarded as a maintenance advantage, as a damaged unit is readily observed by foot or helicopter patrols.

Although the above materials have been considered suitable for insulation purposes as from the early days of overhead line engineering, designers, laboratories, manufacturers, etc. have been investigating not only new shapes, but also new materials to improve the electrical and mechanical performances of insulators. Thus, in the last few decades, composite insulators, consisting of a high mechanical strength core of parallel glass fibres, within a body of synthetic rubber or cycloaliphatic resin, with improved insulating properties, have been actively investigated with a high degree of success. In fact the combination of such materials offers the possibility of almost endless permutations of strength with a large degree of freedom in the shapes.[42]

Composite insulators have also the added advantage of being more resistant to vandalism than the classical types. They are now used at all voltages from distribution to EHV. It is not always appreciated that on an angle tower at 745 kV the weight of classical insulators could reach 30 tonnes, whereas a single man could carry each tension set if it consists of composite insulators.

Figure 30.8 shows some types of insulator units and sets:

(a) Shackle: for low-voltage distribution lines.
(b) Pin: used for voltages up to 33 kV (occasionally 66 kV).
(c) Post: used up to 66 kV, mainly for the support of air insulated bus-bars.
(d) Cap and pin: as elements in insulator strings.



Highly resistant to impact damage, breakage and vandalism

Figure 30.8 Typical insulator units and insulator sets. (a) Shackle insulator; (b) pin insulator; (c) post insulator; (d) normal cap and pin unit; (e) anti-fog cap and pin unit; (f) aerodynamic cap and pin unit; (g) insulator set (single string); (h) 500 kV vee insulator set; (i) composite pin insulator on 11 kV line (courtesy of EA Technology Ltd, Capenhurst)

(e) Anti-fog cap and pin: for regions of high pollution level, especially if humid.
(f) Aerodynamic cap and pin: for desert regions.
(g) Insulator set.
(h) Vee insulator set.
(i) Composite insulator on 11 kV line (by courtesy of EA Technology Ltd, Capenhurst)

The elements (d), (e) and (f) can be assembled into one, two or more strings in parallel. Vee types (h) are often favoured for extra high voltage (EHV) lines, as they show economic advantage.

30.5.2 Selection

As insulators must meet specified criteria, the choice of type, configuration, etc., is complex. Units are characterised by lightning impulse voltage; power frequency voltage when wet; electromechanical failing load, mechanical load and puncture voltage (when applicable); and dimensions, of which creepage path is the most important parameter for selection.

In addition, strings or sets are characterised by: lightning withstand voltage (dry); switching impulse withstand voltage (wet); and power frequency voltage (wet). The wave shapes for switching surge and lightning impulse tests have been standardised in terms of time to crest and time to half-value on the tail. These, in microseconds, are:

Switching surge: (250 ±400)/(2500 ±400)
Lightning impulse: 1.2/50

The performance of an air gap depends on the electrode shapes and on the applied voltage wave shape. *Figure 30.9* shows typical values for a rod/plane gap. Comparable information (or correction factors) is needed for other configurations.[19] The types of gaps are as follows:

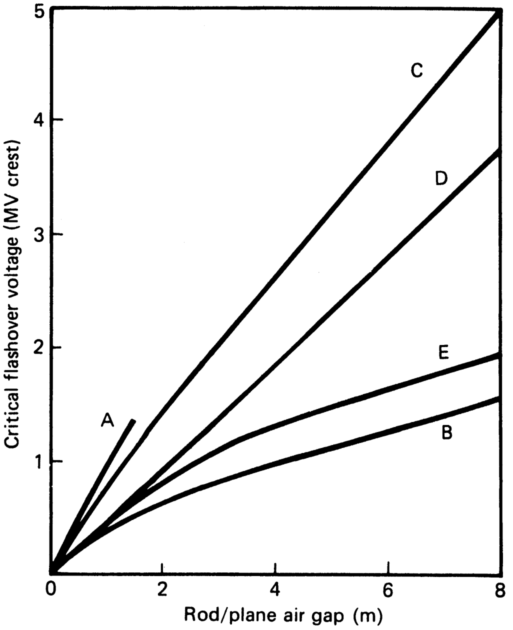


Figure 30.9 Insulation strength of rod/plane air gaps. A, 200/3000 μs, negative, switching surge; B, 200/3000 μs, positive, switching surge; C, 1.2/50 μs, negative, impulse; D, 1.2/50 μs, positive, impulse; E, 50 Hz, power frequency

Rod/plane: live ends of insulators and fittings to structure and ground.
Conductor/structure: inside or outside a tower window, outer phases to any part of the tower.
Conductor/plane: clearance at mid-span to ground, or conductor jumper to cross-arm.
Conductor/conductor: flashover between phases or between conductor and earth wire.

For insulation co-ordination, the electric strength of the air gaps must be about 10% greater than that of the insulators.[20]

Up to about 400 kV, line insulation is governed by the requirements of creepage distance. For higher voltages, switching-surge considerations predominate.

30.5.3 Pollution

A first approximation to the number of units in a string is by an assessment of the intensity of pollution in the area, adopting a creepage distance of 18 mm/kV of line voltage for clean air and up to 30 mm/kV or more for severe pollution.[3,21]

Insulator surfaces may be contaminated by dust and dirt, whose composition varies according to the area where the line is built. If frequent rains occur, these layers may be washed, but occasionally insoluble layers are formed in the presence of mist or dew; and in the absence of heavy downpours, they reduce the insulation resistance of the insulators. Under certain conditions, flashover can occur which ionises a path in air, providing a channel for follow-currents at power frequency.

To assess the insulation performance of insulators under pollution, two series of tests have been devised.[22]

(1) *Salt-fog method*: generally representative of coastal areas or of zones where rain can wash the insulators. In this test, insulators are submitted to electrical stresses in an artificial fog of variable salinity.
(2) *Solid-layer method*: generally representative of industrial zones or desert areas where a solid crust of insoluble deposits can be formed. In this test, the insulators are covered with a special mixture attempting to reproduce natural pollution.

If measurements are made of leakage currents under natural pollution (e.g. at an open-air station), these can be reproduced in the laboratory by varying either the degree of salinity or the composition of the solid layer. Thus, a correlation can be found and a rational assessment can be made as to the ultimate behaviour of the insulators.

Work has been done on palliative measures such as coatings of hydrocarbon or silicon grease, or live-line working. These measures are normally restricted to specific areas.

30.5.4 Voltage distribution over insulator strings

When assessing radio noise from an insulator string, it is desirable to know how the voltage along the string is distributed. This can be found by test: (1) a small spark-gap of known flashover voltage is connected across the cap and pin of one insulator element; and (2) the voltage across the string is raised until flashover at the spark gap, giving the voltage across the element as a fraction of the test voltage. *Table 30.6* gives the results of such a test for a string of

Table 30.6 Voltage distribution over insulator strings without grading fitting

No. of units	*Unit number from live end*	*Voltage across unit (%)*
4	1, 2, 3, 4	32, 24, 21, 23
6	1, 2, 3, 4, 5, 6	25, 18, 14, 13, 14, 16
8	1, 2, 3, 4, 5, 6, 7, 8	24, 15, 13, 11, 9, 8, 9, 11
12	1, 2, 3, 4, 5, 6,	23, 13, 10, 8, 7, 6,
	7, 8, 9, 10, 11, 12	5.5, 5, 5, 5.5, 6, 7
16	1, 2, 3, 4, 5...	20, 13, 9, 7.5, 5.5...

porcelain elements of diameter 254 mm, the string having no protective devices. It will be noted that, for strings with eight or more units, the live-end unit carries 0.20–0.24 of the string voltage, and the second 0.13–0.15.

The application of grading rings and horn fittings modifies the voltage distribution by introducing capacitance in shunt with the line-end units to reduce the voltage across them.

30.6 Supports

The supports of an overhead line have the greatest impact on visual amenity. They cannot be hidden, but their appearance may be made less objectionable.

30.6.1 Materials

Common structural materials are wood, concrete, tubular or rolled-section steel and aluminium alloys. Choice is governed by economics, availability, resistance to deterioration (e.g. by termites in wood, and corrosion in steel), and sometimes by problems of transportation and erection. When helicopters are used in regions of difficult access, aluminium alloy structures may show a lower cost of installation that counterbalances their greater prime cost.[3]

Wood poles are usually preferred for low voltage distribution lines; they have been used for 345 kV lines in the USA and 132 kV in the UK, but normally wood poles are unlikely for voltages over 66 kV.

Concrete (reinforced, prestressed or spun) has for poles the drawbacks of weight and fragility during transport. Concrete poles have been employed for lines up to about 220 kV. The material is unlikely to be competitive unless local production facilities are available or there are constraints on imported material.

Steel, the material most preferred, can be fabricated in convenient lengths for transport after rolling into angles, beams and rounds. Parts are readily bolted together, facilitating unit-type construction. Steel for overhead lines is normally galvanised by the hot-dip process (BS 729) as a protection against corrosion. In clean air, galvanising can ensure 30–50 years of trouble-free life. Where the atmosphere is potentially corrosive, it may be necessary to paint steel structures after 10 years or less. Two grades of steel are commonly applied to line towers, *viz*. grade 43A and grade 50B (high tensile) to BS 4360, the latter grade showing advantage in weight reduction.

Aluminium alloy is advantageous in special cases (sites difficulty of access or with corrosive atmospheres). A structure of this material is 0.6–0.75 of the weight of an equivalent steel tower, but its cost may be three times as much.

30.6.2 Configurations

There is a considerable variety in configuration: the more usual forms are shown diagrammatically in *Figure 30.10* for wood poles and *Figure 30.11* for steel towers.

Wood poles These are more economical than lattice-steel supports for lines having spans less than about 200 m and are widely used for distribution in rural areas at voltages up to and incuding 33 kV. Where ample supplies of wood are available, e.g. in Sweden and parts of America, they have been used for lines up to 220 kV; at these high voltages the portal type of construction is essential to keep the length of pole to a minimum. For lower voltage lines the single pole is generally used, although its transverse strength can be increased three or four times by using two poles arranged in A or H formation. The cross-arms in both the single and the portal types may be of wood or steel, the latter being more usual. A zinc or aluminium cap covers the top of the poles to protect the end grain. The life of a red fir pole, if properly creosoted, is 25–30 years, although British fir or larch has a shorter life.

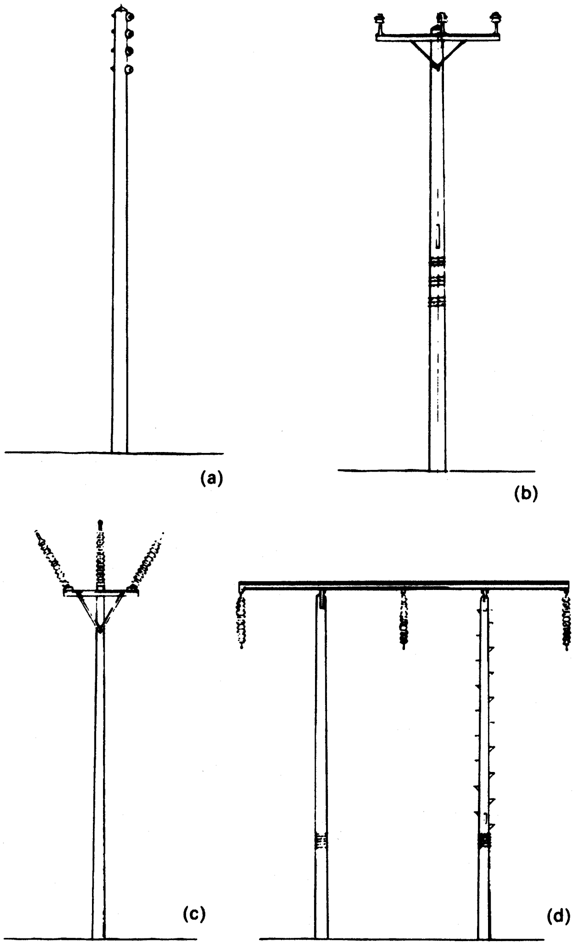


Figure 30.10 Typical wood pole constructions for 415 V to 132 kV: (a) low voltage single member pole; (b) 11 kV single member pole; (c) 132 kV support with composite (polymeric) insulators; (d) 132 kV portal (H) structure with standard insulator sets

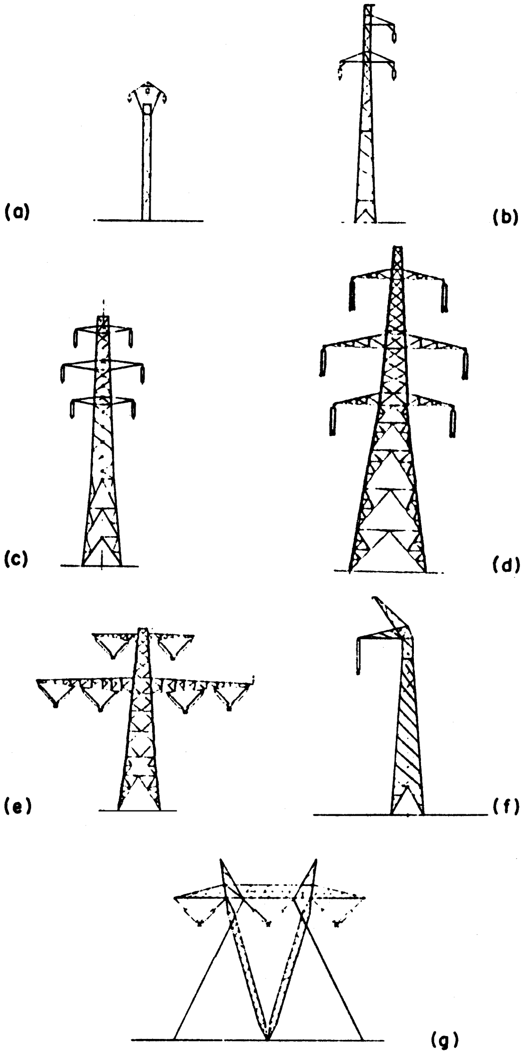


Figure 30.11 Typical steel towers

Wood is an insulator, so that the pole adds to the insulation strength between the conductors and earth and renders flashover due to lightning less likely. On the other hand, if a pole is struck, it may be shattered, causing complete failure of the line. The probability of shattering can be reduced by earthing all metal supports for the insulators either individually at each pole or by connection to a continuous earth wire.

Tubular steel poles These are being increasingly used in the UK, because of the reduction in timber supplies. They are usually formed from three tapered, hollow sections, which rest inside one another during transportation and storage, and which are assembled on the site by means of a special winch, or are driven together by a sledgehammer or wooden maul.

Steel towers These are employed where long spans and high supports are needed. Normal spans of 200–500 m are usual, with special cases (such as river crossings) requiring spans up to 2 km. Lines of 66 kV and over usually have lattice steel supports, especially if they are double-circuit lines.

The lattice-steel support may be of the *broad base* (*Figure 30.11*(*b*)) or rigid type, in which each leg of the support has a separate foundation, or of the *narrow base* or flexible type, with only a single foundation (*a*). The latter is, of course, cheaper, but is less resistant to the twisting moment caused by a broken conductor; on the other hand, it gives some flexibility in the direction of the line which tends to relieve the forces to some extent. The narrow base may also simplify wayleave problems in cultivated areas on account of the smaller area occupied.

Double-circuit lines (*c*) are more economical than two single-circuit lines, although for a given voltage the tower height is greater. There is the possibility of both circuits being affected simultaneously by lightning, but wayleave problems are eased.

The cat's head horizontal configuration (*g*) minimises height and reduces the possibility of conductor clashing, but line erection is complicated by having to thread one phase through the steelwork.

The tower (*f*) for a single-circuit d.c. line is simpler, smaller and 20–30% cheaper than a three-phase single-circuit tower to carry the same electrical load.

The 400 kV double-circuit tower (*d*), widely used in Britain, is 50 m high and may interfere with amenity. To avoid this the lower type (*e*) may be used but is 10–15% more expensive.

Compacted lines Overhead Line Engineers have always been conscious of the fact that their structures could be an eyesore on the countryside, and one way to deal with this aspect would be to reduce their visual impact.

There is no internationally agreed definition of 'compact lines'. Essentially, it is a subjective assessment of the visual impact of the lines on the countryside. It may be that some countries (e.g. Germany, Japan, etc.) would consider as compact lines those that carry an increased number of circuits, thereby avoiding the need to negotiate new wayleaves, but at the expense of wider existing wayleaves for the 'compacted lines'. Other countries would just re-arrange the positions of the circuits on the supports to reduce height at the expense of increased width.

An illustration of the later concept is given by *Figure 30.11*(*d*) and (*e*) for towers in the UK drawn to the same scale and designed to transmit exactly the same electrical power where (*e*) illustrates the compacted design.

When the phase conductors are supported by vertical suspension insulators (e.g. *Figure 30.11*(*d*)), considerations of electrical clearances can lead to fairly high structures (see next paragraphs on tower top geometries). Almost invariably, because of the presence of earthed steelwork, the horizontal separations are greater than would be justified by straightforward electrical or dynamic (wind induced motions) considerations.

Cap and pin insulators (*Figure 30.8*(*f*), (*d*), (*e*)) can only be used in tension and cannot accept compression loads. However, if an insulator set is made up of long rods, or composite insulators, some degree of compression can be accepted, a feature which can be used to achieve an assembly which could control the horizontal swing of insulators at the towers. A reasonable horizontal spacing can also be achieved by using 'V' sets, provided that the inside angle of the 'V' is properly selected as function of the horizontal wind loads and of the vertical conductor weight span loads. Such a solution is shown on *Figure 30.11*(*e*) where the phase spacing between the two outer phases is much smaller than on the adjacent tower with vertical sets.

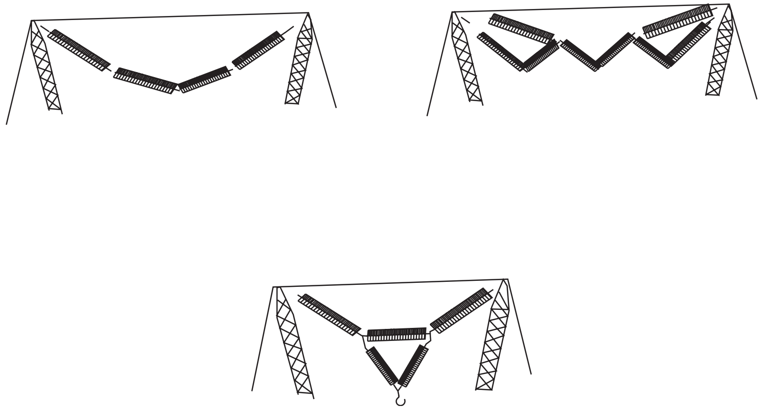


Figure 30.12 Possible tower top geometries for 'compact lines', with no steelwork between conductors

Clearly then the choice of the insulators has a profound impact on the aesthetics of towers. The structure on *Figure 30.10(c)* can be considered as an element of a compact line.

Figure 30.12 illustrates some possible tower top geometries with no steelwork between the conductors. The horizontal spacings are only governed by electrical and dynamic considerations. It is thus possible to reduce the visual impact due to height, but at the expense of an increased volume impact of a conglomeration of insulator sets. Slim composite insulators would clearly reduce such an impact. Unfortunately installation costs may be increased as conductor stringing may prove more complex.

30.6.3 Tower geometry

Considerations of system design will decide whether a line should carry a single or a double circuit. The choice is influenced by the problem of wayleave. In the UK most grid lines carry a double circuit; elsewhere a multiple-circuit construction to accommodate circuits of different voltages is sometimes used.

The geometry of the tower top is controlled essentially by electrical considerations.

(1) *Height to bottom cross-arm*: the sum of the clearance to ground, the sag and the insulator-set length.
(2) *Vertical spacing*: the sum of the electrical clearance plus the insulator length.
(3) *Horizontal spacing*: (a) in still air, a clearance at least equal to the impulse strength of the insulation; (b) in wind conditions (on the assumption that a fault is unlikely to occur during maximum wind speed), a clearance corresponding at least to the power frequency withstand voltage of the air gap; the deflected position of the insulator string determines the cross-arm length.

It is generally found that horizontally placed conductors have sufficient clearance when they are separated by earthed steelwork. For exceptionally long spans, or when two conductors at the same level are on the same side of the structure, their minimum separation should be

$$D = K_1\sqrt{(s+l)} + K_2 V$$

where s is the maximum sag of the conductor in the span, l is the length of a suspension-insulator set, V is the line voltage (kV), $K_1 = 0.6 \times K_3$ (usually $0.7 < K_1 < 0.9$), $K_3 = [\sqrt{(W^2 + p^2)}]/W$ where W is unit weight of conductor (kgf/m) and p is 50% of wind design pressure (kgf/m) per unit length; and $K_2 = 0.0025 \times \sqrt{3}$. The expression is empirically based on the performance of long spans.

The phase geometry of single-circuit lines is either triangular or horizontal. Double-circuit lines are more likely to have a vertical formation. From the system viewpoint, a triangular arrangement results in minor imbalance of the phase impedances (see also Section 30.7).

30.6.4 Foundations

Foundations[3] must be adapted to the loadings that they have to withstand and the properties of the soil, i.e. cohesion C, angle of internal friction ϕ, and density γ.

For single poles and narrow-based towers, the foundations are designed on overturning; for broad-based towers, on uplift/compression. Compression effects are well documented, but uplift has been the subject of lengthy investigation.[23]

Overturning (monobloc) foundations Resistance to overturning is of the form $M = kbd^3$ for a width b and a depth d. The coefficient k depends on the shape of the bloc and the properties of the soil.

Uplift foundations In the UK, where 'pad and chimney' foundations are favoured, it is customary to assume that the resistance to uplift is due to (a) the dead weight of the foundation and (b) the weight of the earth contained in a volume of a frustum defined by the base of the foundation and an angle to the horizontal reflecting the soil properties. In addition, the 'chimney' is designed in bending. A general expression for the uplift resistance[24] is

$$F_t = A[Ck_1 + d\gamma(k_2 + k_3) + qk_4] + p_c + p_s$$

where A is the lateral area rising from the base of the foundation to the surface of the soil, d is the depth, p_c is the weight of concrete and p_s is the weight of soil above the foundation, and q is the overburden pressure for foundations set below the critical depth. The critical depth is a function of the geometry of the foundation and of the soil parameters. The constants k depend on the type of soil. In a pad and chimney foundation the pad can be shaped by formers or cast on site into an excavated hole. When access is difficult and concrete expensive (because of lack of available water or aggregates), grillage pads may be considered, as individual bars can be transported to site. In bad soils pile or raft foundations may be used; in good rock the foundation may employ rock anchors.

30.7 Lightning

Lightning results from a phenomenon in which the clouds in the ambient atmosphere acquire substantial electric charges, with corresponding potentials which may be several megavolts. When the limit breakdown value of the air is reached, a lightning discharge takes place between clouds, or between clouds and earth. Because of their height, overhead transmission lines 'attract' lightning discharge. Data are available for the percentage of days when thunder is heard[20] and the number of flashes to earth per annum per unit area. The 'keraunic level' is the number of days thunder is heard in a year.

30.7.1 Mechanisms of insulation flashover

30.7.1.1 Induced voltage

A lightning stroke to ground near a line discharges the energy in the cloud very rapidly, initiating by travelling wave action an induced voltage rise across the insulators. The problem arises only at line voltages of 66 kV and less, for which lines it would be uneconomic to provide a complete guard against this hazard.

30.7.1.2 Shielding failure

Shielding failure occurs when the leader stroke of the lightning discharge bypasses the earth wire (which is usually at ground potential) and strikes a conductor. Some protection against this hazard is provided by positioning the earth wire(s) above the conductors. Analytical models[25] have been evolved for the calculation of the appropriate position(s). A shielding angle of 30° is generally adequate: the angle is that between the vertical through the earth wire and the line joining the earth wire and the protected conductor. However, in areas of high lightning intensity, zero (or even negative) angles have been used.

30.7.1.3 Back-flashover

Figure 30.13 shows that about 10% of lightning strikes involve a current i exceeding 50 kA of very steep wavefront. When lightning strikes (a) a tower or (b) an earth wire, their potential may be raised to a value well in excess of the insulation level, possibly high enough to cause flashover to the line conductors.

Tower If the tower resistance (mainly that of the footing) is R and the inductance as a vertical circuit element is L, then the potential of the tower top will be raised from zero to $v = Ri + L(\mathrm{d}i/\mathrm{d}t)$. This simple approach illustrates the effect of the basic parameters: a more critical study would be required to take account of capacitive and inductive couplings between conductors, the effects of multiple reflections, etc.

Earth wire Given that the surge impedance of the earth wire is Z_0, a lightning strike raises the earth wire potential from zero to $v = \frac{1}{2}Z_0 i$. Two travelling waves are propagated in opposite directions from the point of strike, until each reaches a tower top and is partially discharged through the tower to ground, as in A in *Figure 30.13*.

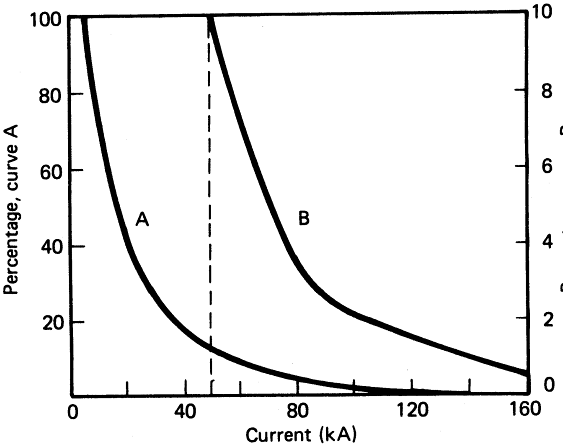


Figure 30.13 Lightning strike magnitude probability curves. Abscissa: lightning strike peak current (kA). Ordinate (left side): percentage of strokes exceeding abscissa. A, Ordinate (right side): percentage of strokes exceeding abscissa. B, for lightning currents exceeding 50 kA

30.7.2 Lightning performance

Various methods (ERA,[20] IEEE[26] and EPRI[27]) have been evolved, and refined in the light of experience, for the estimation of the lightning performance of overhead lines. The basic step is the calculation of the potential rise of the tower top or earth wire as a function of the lightning current discharged to ground. It is then necessary to develop an analytical model to include the statistical distribution of current wave shapes, grounding parameters, leader approach angles, ground flash density, back-flashover and shielding failure.[28,29] The ERA method employs a step-by-step procedure clearly demonstrating the factors affecting line performance; the IEEE method, easier to apply, is based on generalised graphs. The EPRI method is very explicit.

In overhead-line design, some parameters will depend on lightning as they affect the number and positioning of earth wires, the length of insulator strings and tower-footing resistance. Further considerations relate to tower height, phase configurations and span lengths. It is now accepted that an earth wire with 10% smaller sag than that of the associated conductor minimises the likelihood of mid-span flashover between them. The main purpose of an earth wire is to shield phase conductors from direct strokes, which would almost always result in insulator flashover.

Increased tower height attracts more lightning strokes. Thus, the lightning performance of a line is improved by shorter towers (and consequently shorter spans). Short towers with wide bodies have lower surge impedance. But in many cases it is economic considerations that settle the final choice.

30.8 Loadings

The mechanical loadings on an overhead line are due to conductor tension, wind, icing and temperature variations. The design regulations or codes of practice in most countries were based on the concepts of external working loads multiplied by different factors (often termed factors of safety, overload factors, etc.) to evaluate ultimate (or limit) loads which were compared with the strength of components. In the UK, the earlier Regulations[30] gave some loadings which were to be compared with a proportion of the strength of components. For the first time, the expression 'factor of safety' was avoided, but the concept is still used in many international regulations. In the UK, the utilities (e.g. National Grid Corporation) in fact use higher loads than those specified in these Regulations. In many cases, the origins of the loads and the factors used are never very clear.

Actually, in 1988, a new set of Regulations was published in the UK[31] which concentrates mainly on electrical safety, but is exceedingly concise regarding the loadings to be applied to an overhead line. In fact, they only state that the '. . . works shall be sufficient for the purposes for, and the circumstances in, which they are used . . .'. The onus for the adequate performance lies with the operator and the designer. As a result of international discussions a probabilistic approach is now gaining acceptance.[40,41]

Accepting (*Figure 30.14*) that the load effects on an overhead line are random[32,33] and can be expressed by a probability density function of extreme values (Gumbel type 1) $f_Q(X)$, that the strength of the components can be expressed by a normal probability density function (Gaussian) $f_R(X)$, new probabilistic approaches have been evolved[34] which are likely to be used by specification writers.[3] (In *Figure 30.14* the strength function is shown as a cumulative distribution, $F_R(X)$.) These two functions lend themselves to mathematical treatment and are defined when a mean value and a standard deviation can be evaluated from measurement data. The area under the two overlapping curves is a measure of the risk (in this zone there is a probability that the load is greater than the strength), and they can be located with respect to each other in such a way that a small value of risk can be obtained. The smaller the risk (probability of failure) the greater the reliability. It will be seen that mathematically 100% reliability cannot be achieved. This feature is not due to the approach, it is a hard real fact, except that in the conventional deterministic treatment there is no way to quantify the reliability of a construction.

Some documents[34] have been revised and incorporate comments[35] which have already been made. Several application examples have also been published.[36,37]

Three important concepts have been introduced:

(1) *Reliability* (structural) is the probability that a system performs a given *task* under a set of *conditions* during a specified *time*. Reliability is thus a measure of the success of a system in accomplishing its task. Reliability, being a probabilistic concept, should always be quantified.
(2) *Security* (structural) is the ability of a system to be protected from a major collapse (cascading effect) if a failure is triggered in a given component. Security is a deterministic concept.
(3) *Safety* (structural) is the ability of a system not to cause human injuries or loss of lives. In these documents, safety relates mainly to protection of workers during construction and maintenance operations. The safety of the public should be covered by National Regulations.

A useful concept in deciding loading effects is that of the 'return period' of a given loading (e.g. wind speed and related pressure). Due to Gumbel, the return period is the mean time between the occurrences of a phenomenon equal to (or greater than) a specified value. As return period T is related to probability p by the expression $Tp = 1$ it follows that a loading event with a return period $T = 50$ years has a probability p, of 2% of being exceeded in any one year.

In order to rationalise the concept of strength, it is suggested that the value corresponding to 10% exclusion limit (hence with a 90% probability of exceeding this value) would be used as reference. In *Figure 30.14*, the load equal to a return period of T (years) is compared with (or made equal to) the strength at 10% exclusion limit. The area under the two overlapping curves is a measure of the annual risk P_f. The annual reliability P_s of this arrangement is:

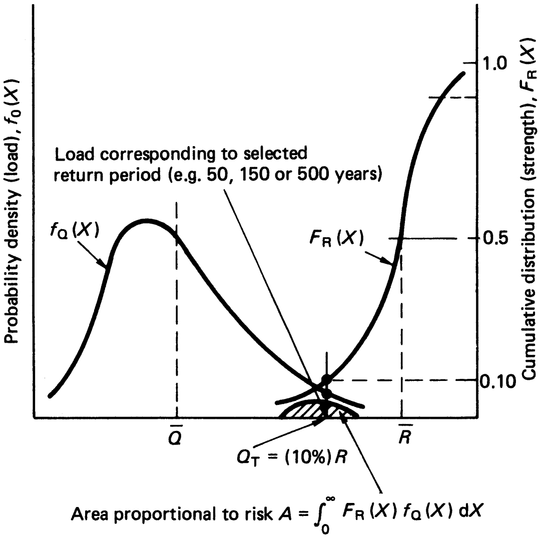


Figure 30.14 Principles of risk assessment[35]

$$P_s = 1 - P_f$$

If σ_Q is the standard deviation of the load, σ_R that of the strength, and $\bar{Q}$ and $\bar{R}$ are their mean values then the coefficients of variation are $V_Q = \sigma Q/\bar{Q}$ for the load and $V_R = \sigma R/\bar{R}$ for the strength.

As an order of magnitude, when $V_Q = 0.3$ and $V_R = 0.10$ then for a return period of load of 50 years, $P_s = 0.98$ and for a return period of load of 150 years $P_s = 0.993$.

One assumption made in the probabilistic approach is that one component would be selected to be weaker than the others, hence the reliability of the line would be that of the weakest component. All the other components would have to be designed in such a way that there is a given probability that they are stronger. *Figure 30.15* expresses this concept.[35]

Thus if the area (A1) is a measure of the risk that the load may exceed the strength of the component, the area (A2) is a measure of the probability that component $R2$ may be weaker than component $R1$.

Some typical values of the strength coordination factor ϕ_s are given below, for 90% probability that the desired probability of failure will be achieved ($R1$ to fail before $R2$):

V_{R2}	V_{R1}		
	0.05	0.10	0.20
0.05	0.91	0.81	0.63
0.1	0.92	0.83	0.66
0.2	0.93	0.86	0.69

where V_{R1} is the coefficient of variation of component $R1$ and V_{R2} is the coefficient of variation of component $R2$.

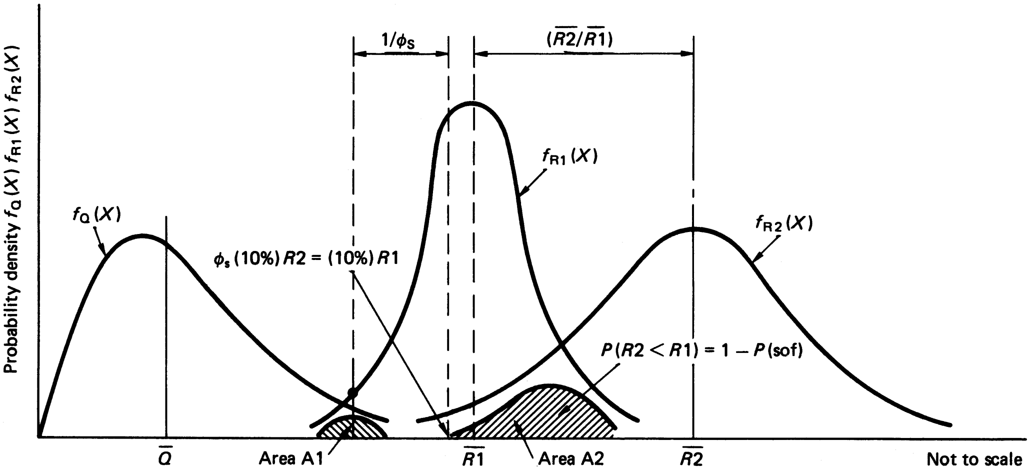


Figure 30.15 Definition of the strength coordination factor ϕ_s. $P(R2 < R1) = 1 - P$(sof) is the probability that strength $R2$ is smaller than strength $R1$. It is equal to one minus the required probability for the sequence of failure. Not to scale

The basic mathematical model used for the probabilistic design of the line is given by the following general design equation:

$$\gamma_u Q_T < \phi_R R_C$$

or load effect < strength.

All documents dealing with reliability accept this design approach. Some may be more specific as to whether Q_T is a wind load, an ice load or wind on ice load. If Q_T is considered as any load due to meteorological influences, with a return period of T years, the above relationship is always valid.

In this relation, and for IEC documents:

γ_u	is the use factor coefficient applied to load;
Q_T	is the load corresponding to a return period T. This is the limit (or ultimate) load for design purposes (50, 150, and 500 years are accepted as references);
ϕ_R	is the global strength factor;
ϕ_R	$= \phi_c \phi_N \phi_q \phi_s$;
R_c	is the characteristic or nominal strength;
ϕ_c	is the correction factor referred to the exclusion limit of material strength;
ϕ_N	is the factor related to the number N of components subjected to the critical load;
ϕ_q	is the factor related to the quality level of a component during fabrication and/or construction; and
ϕ_s	is the strength coordination factor.

Values for these factors are given in tables[34,35] together with the necessary justifications, and the equations required for their evaluation. A reasonable tool is thus available to the overhead line engineer for the mechanical design of the line. No doubt a fair amount of calibration will still be needed. However comparisons of alternative designs become possible not only with regard to economics, but also with regards to relative reliabilities.

References

1 ASH, D. O., DEY, P., GAYLARD, B. and GIBBON, R. R., 'Conductor systems for overhead lines: some considerations in their selection', *Proc. IEE*, **126**(4), (1979)
2 ORAWSKI, G., BRADBURY, J. and VANNER, M. J., 'Overhead distribution lines—some reflections on design', *IEE Proceedings, Part C*, **133**(7), (November 1986)
3 ADAM, J. F., BRADBURY, J., CHARMAN, W. R., ORAWSKI, G. and VANNER, M. J., 'Overhead lines—some aspects of design and construction', *IEE Proceedings, Part C*, **131**(5), (September 1984)
4 Study Committee 22 Report, 'Aeolian vibration on overhead lines', *CIGRE Paper 22–11*, Paris (1970)
5 IEEE COMMITTEE REPORT, 'Standardisation of conductor vibration measurements', *IEEE Transactions on Power Apparatus and Systems*, **Pas-85**(1), (1966)
6 CIGRE, STUDY COMMITTEE 22, WORKING GROUP 04, *Endurance Capability of Conductors, Final Report* (July 1988)
7 PRICE, S. J., ALLNUTT, J. G. and TUNSTALL, M. J., 'Subspan oscillations of bundled conductors', *IEE Conf. Progress in Cables and Overhead Lines for 220 kV and Above* (September 1979)
8 WATES, R. H., JACKSON, G. B., DAVIS, D. A., ERSKINE, A., BROWN, R. C. and ORAWSKI, G., 'Major high voltage long span river crossings in Great Britain', *CIGRE Paper 226*, Paris (1964)
9 BRADBURY, J., DEY, P., ORAWSKI, G. and PICKUP, K. H., 'Long term creep assessment for overhead line conductors', *Proc. IEE*, **122**(10), (1975)
10 CIGRE SC22-WG05, 'Permanent elongation of conductors. Predictor equations and evaluation methods', *ELECTRA*, **75** (March 1981)
11 PRICE, C. F. and GIBBON, R. R., 'Statistical approach to thermal rating of overhead lines for power transmission and distribution', *IEE Proceedings, Part C*, **130**(5), (September 1983)
12 BUTTERWORTH, S., *Electrical Characteristics of Overhead Lines, ERA Report O/T4* (1954)
13 *Specification for CISPR Radio Interference Apparatus for the Frequency Range 0.15 MHz to 30 MHz, CISPR Publication No. 1*, IEC
14 ANON., 'Survey of extra high voltage transmission line radio noise', *ELECTRA* (January 1972)
15 *Insulation Co-ordination*, IEC Standard—Publications 71–1, 71–2 and 71–3

16 *BS 5622: Guide for Insulation Co-ordination* Part I, Terms Definitions, principles and rules. Part II, Application guide, Milton Keynes

17 *Tests on Insulators of Ceramic Material or Glass for Overhead Lines with a Nominal Voltage Greater than 1000 V*, IEC Standard—Publication 383

18 *BS 137: Specification for Insulators of Ceramic Material or Glass for Overhead Line with a Nominal Voltage Greater than 1000 V*. Part 1, Tests, Part 2, Requirements

19 PARIS, L. and CORTINA, R., 'Switching and lightning impulse discharge characteristics of large air gaps and long insulator strings', *IEEE Trans.*, **Pas-87**(4) (1968)

20 MORRIS, T. A. and OAKESHOTT, D. F., *Choice of Insulation and Surge Protection of Overhead Transmission Lines of 33 kV and Above*, ERA Report O/T14

21 IEC, 'Guide for the selection of insulators in respect of polluted conditions', *IEC Publication 815* (1986)

22 *Artificial Pollution Tests on High Voltage Insulators to be Used on a.c. Systems*, IEC Report, Publication 507

23 MARTIN, D., 'Design of uplift foundations', *ELECTRA*, No. 38, January 1979—prepared within scope of WGO7 of CIGRE SC22

24 VANNER, M. J., 'Foundations and the effect of the change in ground conditions over the seasons', *2nd Int. Conf. Progress in Cables and Overhead Lines for 220 kV and Above*

25 WHITEHEAD, E. R., 'Lightning protection of transmission lines', Chap. 22, *Lightning* (ed. R. H. Golde), Academic Press, London

26 CLAYTON, J. H. and YOUNG, F. S., 'Estimating lightning performance of transmission lines', *IEEE Trans* **Pas-83** (Paper No. 64–138)

27 EPRI, *Transmission Line Reference Book: 345 kV and Above*, 2nd edition (1982)

28 GILMAN, D. W. and WHITEHEAD, E. R., 'The mechanism of lightning flashover on HV and EHV transmission lines' *ELECTRA*, No. 27

29 'CIGRE, survey of the lightning performance of EHV transmission lines', prepared by E. R. Whitehead, *ELECTRA*, No. 33, March (1974)

30 Statutory instruments, 1970 No. 1355—Electricity. *The Electricity (Overhead Lines) Regulations*, HMSO (1970)

31 'Electricity—the electricity supply regulations', *Statutory Instruments No. 1057* HMSO, London (1988)

32 MANUZIO, G. and PARIS, L., 'Statistical determination of wind loading effects on overhead line conductors', *CIGRE Paper 231*, Paris (1964)

33 ARMITT, J., COJAN, M., MANUZIO, C. and NICOLINI, P., 'Calculation of wind loadings in components of overhead lines', *Proc. IEE* **122**(11), (1975)

34 IEC, *Technical Report 826:1991*, second edition, replaces the first edition which was published in four separate parts between 1985 and 1987

35 CIGRE WG22-06, 'Loading and strength of overhead transmission lines. (A commentary on IEC documents in the 826 series, first edition). *ELECTRA*, 129 (March 1990) pp. 65–97. Reprinted in full as erratum to this paper, *ELECTRA*, 137, pp. 130–169. *CIGRE*, 3–5 Rue de Metz, 75010, Paris, France (August 1991)

36 ORAWSKI, G., 'Calculation of overhead lines tower loadings—wind loading only', *Convener CIGRE WG22-06* (see *ELECTRA,* **129**, 35 (March 1990) for availability

37 'Calculations of wind and ice loadings—An example of the application of IEC Reports 826-2 and 826-4', *CIGRE WG22-06, ELECTRA*, **132**, 127–147 (October 1990)

38 BRADBURY, J., KUSKA, G. F. and TARR, D. J., 'Sag and tension calculations in mountainous terrain', *IEE Conf. Progress in Cables and Overhead Lines for 220 kV and Above* (September 1979)

39 CIGRE, *Interferences Produced by Corona Effect of Electric Systems—Description of Phenomena, Practical Guide for Calculation* (document established by Working Group 36–01 (Interferences)), (1974)

40 WOOD, A. B., FAIR, I. R. and LIPTROTT, F. J., 'Transmission line design: the ultimate load concept applied to Java 500 kV transmission lines', *Power Engineering Journal*, IEE London (March 1988)

41 ORAWSKI, G., 'Overhead lines—loading and strength: the probabilistic approach viewed internationally', *Power Engineering Journal*, IEE London (September 1991)

42 LOOMS, J. S. T., 'Insulators for high voltages'. Published in 1988 by Peter Peregrinus Ltd on behalf of the Institution of Electrical Engineers

31

Cables

D W Pryer CEng, MIEE
Formerly of BICC Cables Ltd
(Sections 31.1–31.7)

P F Gale BTech, PhD, CEng, MIEE
Hathaway Instruments Ltd
(Section 31.8)

Contents

31.1 Introduction 31/3
  31.1.1 Standards 31/3

31.2 Cable components 31/6
  31.2.1 Conductors 31/6
  31.2.2 Insulation 31/7
  31.2.3 Armour 31/10
  31.2.4 Oversheaths and protective finishes 31/10

31.3 General wiring cables and flexible cords 31/11
  31.3.1 Wiring system cables 31/11
  31.3.2 Flexible cords 31/11
  31.3.3 Control and instrumentation cables 31/12
  31.3.4 Cables for electronic applications 31/12
  31.3.5 Arc welding cables 31/12
  31.3.6 Offshore and ship cables 31/13
  31.3.7 Aircraft cables 31/13
  31.3.8 Cables for railways 31/13
  31.3.9 Cables for mines and quarries 31/14
  31.3.10 Mineral insulated metal sheathed cables 31/14
  31.3.11 Cables in fire hazard 31/15

31.4 Supply distribution cables 31/15
  31.4.1 Paper insulated cables 31/16
  31.4.2 CNE cables for PME systems 31/18
  31.4.3 Service cable 31/19
  31.4.4 PVC insulated power cables 31/19
  31.4.5 XLPE insulated cables up to 3.3 kV 31/20
  31.4.6 PE and XLPE cables for 11 kV to 45 kV 31/20
  31.4.7 Cable tests 31/22

31.5 Transmission cables 31/23
  31.5.1 Historical development sequences for a.c. transmission 31/23
  31.5.2 Types of cable 31/24
  31.5.3 Submarine power cables 31/27
  31.5.4 D.c. transmission 31/27
  31.5.5 Cable ratings and forced cooling 31/28
  31.5.6 Future development 31/29

31.6 Current-carrying capacity 31/30
  31.6.1 Availability of continuous ratings 31/31
  31.6.2 Factors in cable ratings 31/31
  31.6.3 Sustained ratings 31/32
  31.6.4 Short-time and cyclic ratings 31/32
  31.6.5 Short-circuit ratings 31/32
  31.6.6 Voltage drop 31/33
  31.6.7 Protection against overload current 31/33

31.7 Jointing and accessories 31/33
  31.7.1 Aluminium conductor jointing 31/33
  31.7.2 Joints for distribution cables 31/34
  31.7.3 Joints for transmission cables 31/35

31.8 Cable fault location 31/35
  31.8.1 Diagnosis 31/35
  31.8.2 Preconditioning 31/36
  31.8.3 Prelocation methods 31/36
  31.8.4 Cable fault characteristics 31/37
  31.8.5 Pinpointing 31/39

31.1 Introduction

The essential components of a cable are a metallic conductor of low resistivity to carry the current and insulation to provide a dielectric medium for isolating conductors from one another and from their surroundings. The conductor may consist of solid metal or of wires or segments stranded together. A single-core wiring cable for installation in conduit represents this basic construction. For other applications, two or more single-core units may be assembled together with overall protective coverings to prevent moisture ingress, and provide resistance to mechanical damage and to other external influences such as corrosion and fire.

In general, the voltage range extends from automobile cables at 6–12 V to the highest transmission voltages, which now are reaching towards 760 kV. In order to specify suitable insulation and construction for the required service performance, the design voltages are quoted in the form of U_0/U, i.e. (voltage to earth)/(voltage between phases). Cables are not manufactured, however, for every individual voltage requirement—e.g. although the most common supply voltage in the UK is 240/415 V, the cables actually used are designated as 0.6/1 kV. This is largely related to the fact that the minimum thickness of insulation which can be economically applied meets the higher voltage.

It has been traditional practice to describe cables in categories of low-voltage (l.v.), medium-voltage (m.v.), high-voltage (h.v.), supertension (s.t.), extra-high-voltage (e.h.v.) and ultra-high-voltage (u.h.v.). However, the exact demarcations have never been very precise, because they vary between different countries, among different groups of engineers and with the passing of time. Not so long ago, the conventional supply cables operating at 240/415 V were known as m.v., but now they are l.v.; 11 kV cables have also changed from h.v. to m.v. Thus, there can be confusion especially at international meetings, and it is better to refer only to the actual voltage designation.

There is also some overlap when cables are classified into the three major groups of usage: (1) wiring and general, (2) power distribution, and (3) transmission. When heavy power cables were essentially of the paper-insulated type, there was little problem, because they tended to be made in different factories from wiring cables and the type of insulation governed the usage group. When paper insulation operated with internal or external fluid or gas pressure, as required for voltages above 33 kV, the cable came into the transmission grouping. Nowadays, the basic insulation materials and constructions used for wiring cables can, with appropriate insulation thickness, be used across the three major groupings, and it is becoming more complex to define specific cable categories. For public supply there is no problem, but for ship cables, offshore supplies and large factory distribution there can be considerable overlap between the wiring and the distribution categories. At the top end of the voltage range we now have 132 kV cables for distribution rather than transmission, and polyethylene or cross-linked polyethylene is finding acceptance across the whole spectrum from 1 to 275 kV.

In this chapter, the section on power distribution cables is aimed at public supply networks together with the larger power cables used in factories, etc., from 1 to 33 kV. Where the broad wiring cable category encroaches on this field, reference is made in the text. Cables for voltages above 33 kV are covered in the section on transmission cables.

31.1.1 Standards

Table 31.1 lists most of the British (BS) and International (IEC) Standards applicable to cables and cable systems, including tests. While most national standards are in accordance with IEC requirements, the IEC Standards represent a consensus of national opinion and take many years to prepare. There is, consequently, a time lag in taking account of new developments, but British Standards have been more comprehensive in requirements and in updating revisions.

However, a fundamental change affecting the British Standards arises from the activities of the European Committee for Electrotechnical Standardisation (CENELEC), of which the membership consists of the standards organisations of the European Common Market and the EFTA countries. The aim of CENELEC is to remove technical barriers to trade among the member countries, and to this end it is engaged in the harmonisation of their national standards.

The mechanism is the preparation of Harmonisation Documents and, after these are issued, the member countries are required to bring the technical requirements of their national standards into conformity with them. There must be no extra requirements or deviations, except in special circumstances and subject to general agreement, and then only on a temporary basis.

To date, the two most important harmonisation documents for cables which have come into effect are:

HD 21	Polyvinyl chloride (PVC) insulated cables and flexible cords of rated voltage up to and including 450/750 V
HD 22	Rubber insulated cables and flexible cords of rated voltage up to and including 450/750 V

and it is intended to extend harmonisation to all types of mains cables.

Among the features arising from this structure are the following:

1. The broad policy is to use IEC standards, if suitable and available, as a basis for harmonisation.
2. When CENELEC begins work on a particular type of cable (or a subject having an important bearing on cable standards), a 'standstill' or 'status quo' arrangement comes into effect. Changes in relevant national standards cannot then be made until harmonisation has been agreed or permission obtained from CENELEC, to avoid prejudicing the harmonisation.
3. While the number of cable designs and types will be kept to a minimum, it will still be possible to have a national standard for a type of cable not of interest to other member countries of CENELEC.
4. Individual customers can still obtain cables made to their own specification, but it is hoped that this will be kept to a minimum. Because they have to be produced as 'specials' they may suffer from a delivery and/or a price penalty.
5. Implementation for flexible cables and some wiring cables has been effected and work on the harmonisation of mains cables is at an advanced stage. PVC cables to BS 6346 will probably be the first mains cables involved, but the changes are likely to be relatively minor.

Another important facet of CENELEC activity relates to certification and associated marks. These are issued by national approval organisations (NAOs) to indicate independent assurance of manufacture to specification. A common marking scheme, utilising the 'HAR' mark, has been devised. There is a reciprocal agreement accepted by most of the CENELEC countries' national approval organisations, under which for cables to harmonised standards, each NAO will recognise the common marking as superseding its own national mark. The procedures for granting

Table 31.1 British and IEC Standards

Cables and Flexible Cords	
BS 638	Arc welding plant (includes cables)
BS 4553	PVC-insulated split concentric cables with copper conductors for electricity supply
BS 5055	PVC-insulated and elastomer-insulated cables for electric signs and h.v. luminous discharge tube installations
BS 5308	Instrumentation cables intended for intrinsically safe systems
	Part 1 Polyethylene insulated cables
	Part 2 PVC-insulated cables
BS 5467	Cables with thermosetting insulation for electricity supply
BS 5593	Impregnated paper insulated cables with aluminium sheath/neutral conductor and three shaped solid aluminium phase conductors (Consac) for electricity supply
BS 6004	PVC-insulated cables (non-armoured) for electric power and lighting
BS 6007	Rubber-insulated cables for electric power and lighting
BS 6116	Elastomer-insulated flexible trailing cables for quarries and miscellaneous mines
BS 6141	Insulated cables and flexible cords for use in high temperature zones
BS 6195	Insulated flexible cables and cords for coil leads
BS 6207	Mineral insulated copper sheathed cables with copper conductors
BS 6231	PVC-insulated cables for switchgear and controlgear wiring
BS 6346	PVC-insulated cables for electricity supply
BS 6480	Impregnated paper-insulated lead or lead alloy sheathed electric cables for working voltages up to 33 kV
BS 6500	Insulated flexible cords and cables
BS 6622	Cables with extruded cross-linked polyethylene or ethylene propylene rubber insulation for rated voltages from 6.6 kV to 33 kV
BS 6708	Trailing cables for mining purposes
BS 6724	Armoured cables having thermosetting insulation with low emission of smoke and corrosive gases when affected by fire
BS 6726	Festoon and temporary lighting cables and cords
BS 6862	Cables for vehicles
	Part 1 Cables with copper conductors
BS 6883	Elastomer-insulated cables for fixed wiring in ships
BS 6977	Insulated flexible cables for lifts and for other flexible connections
BS 7211	Thermosetting cables (non-armoured) for electric power and lighting with low emission of smoke and corrosive gases when affected by fire
B.S. Aerospace Series	
	G210 (PTFE)
	G212 (General requirements)
	G222 (Efglas)
	G230 (General requirements)
	G231 (Conductors for general purposes)
	G232 (Cables for 135°C use, wrapped insulation)
	G233 (Cables for interconnect use, at 135°C, extruded insulation)
	G235 (Cables for general or interconnect use, 150°C, wrapped insulation, silver plated conductors)
	G236 (Cables for general or interconnect use, 200°C, nickel plated conductors)
	G237 (Cables for general or interconnect use, 200°C, extruded insulation, nickel plated conductors)
	G238 (Cables for general or interconnect use, 260°C, wrapped insulation, nickel plated conductors)
	G241 (Fireproof cables for engine fire zone and airframe use)
	G243 (Igniter cables for engine use (4 kV, d.c.)
IEC 55	Paper insulated metal sheathed cables for rated voltages up to 18/30 kV (with copper or aluminium conductors and excluding gas pressure and oil-filled cables)
	55.1: Part 1 Tests
	55.2: Part 2 Construction
IEC 92	Electrical installations in ships
	92-352 Choice and installation of cables for low-voltage power systems
IEC 245	Rubber-insulated flexible cables and cords with circular conductors and a rated voltage not exceeding 750 V
IEC 502	Extruded solid dielectric insulated power cables for rated voltages from 1 kV to 30 kV
IEC 541	Comparative information on IEC and North American flexible cord types
IEC 702-1	Mineral insulated cables
Conductors	
BS 2627	Wrought aluminium for electrical purposes—Wire
BS 3988	Wrought aluminium for electrical purposes—Solid conductors
BS 4109	Copper for electrical purposes: wires for general electrical purposes and insulated cables and flexible cords
BS 5714	Method of measurement of resistivity of metallic materials
BS 6360	Conductors in insulated cables and cords
IEC 228	Conductors of insulated cables
	228A (Supplement). Guide to the dimensional limits of circular conductors

Cont'd

Table 31.1 (*continued*)

Insulation and Sheathing (*Non-metallic*)	
BS 6234	Polyethylene insulation and sheath of electric cables
BS 6746	PVC insulation and sheath of electric cables
BS 6899	Rubber insulation and sheath of electric cables
IEC 173	Colours of the cores of flexible cables and cords
IEC 304	Standard colours for insulation for low frequency cables and wires
IEC 391	Marking of insulated conductors
IEC 446	Identification of conductors by colours or numerals
Tests on Cables and Materials	
BS 903	Physical testing of rubber (This specification is issued in parts)
BS 4066	Tests on electric cables under fire conditions
	Part 1 Method of test on a single vertical insulated wire or cable
	Part 2 Method of test on a single small vertical insulated wire or cable
	Part 3 Method for classification of flame propagation characteristics of bunched cables
BS 5099	Spark testing of electric cables
BS 6469	Methods of test for insulation and sheaths of electric cables
IEC 55	See above
	55.1: Part 1 Tests
IEC 60	High voltage test techniques:
	60-1: Part 1 General definitions and test requirements
	60-2: Part 2 Test procedures
	60-3: Part 3 Measuring devices
	60-4: Part 4 Application guide for measuring devices
IEC 141	Tests on oil-filled and gas pressure cables and their accessories
	141-1: Part 1 Oil-filled, paper-insulated, metal-sheathed cables and accessories for alternating voltages up to and including 400 kV
	141-2: Part 2 Internal gas-pressure cables and accessories for alternating voltages up to and including 275 kV
	141-3: Part 3 External gas-pressure (gas compression) cables and accessories for alternating voltages up to 275 kV
	141-4: Part 4 Oil-impregnated paper-insulated high pressure, oil-filled, pipe-type cables and accessories for a.c. voltages up to 400 kv
IEC 229	Tests on anti-corrosion protective coverings for metallic cable sheaths
IEC 230	Impulse tests on cables and their accessories
IEC 270	Partial discharge measurements
IEC 332	Tests on electric cables under fire conditions
	332-1: Part 1 Test on a single vertical wire or cable
	332-2: Part 2 Test on a single small vertical insulated copper wire or cable
	332-3: Part 3 Tests on bunched wires or cables
IEC 754	Test on gases evolved during combustion of electric cables
IEC 811	Common test methods for insulating and sheathing materials of electric cables
IEC 840	Tests for power cables with extruded insulation for rated voltages above 30 kV up to 150 kV
IEC 885	Electrical test methods for electric cables
Jointing and Accessories	
BS 4579	Performance of mechanical and compression joints in electric cable and wire connectors
	Part 1 Compression joints in copper conductors
	Part 2 Compression joints in nickel, iron and plated copper conductors
	Part 3 Mechanical and compression joints in aluminium conductors
BS 5372	Cable terminations for electrical equipment
BS 6081	Specification for terminations for mineral insulated cables
BS 6121	Mechanical cable glands for elastomer and plastics insulated cables
IEC 702	702-2: Part 2 Mineral insulated cables—Terminations
Miscellaneous	
BS 801	Lead and lead alloy sheaths of electric cables
BS 1441	Galvanised steel wire for armouring submarine cables
BS 1442	Galvanised mild steel wire for armouring cables
BS 5345	Code of practice for selection, installation and maintenance of electrical apparatus for use in potentially explosive atmospheres (other than mining applications or explosive processing and manufacture) (in 8 parts)
BS 6387	Performance requirements for cables required to maintain circuit integrity under fire conditions
IEC 38	IEC Standard voltages
IEC 183	Guide to the selection of high-voltage cables

Cont'd

Table 31.1 (*continued*)

IEC 287	Calculation of the continuous current rating of cables (100% load factor)
IEC 331	Fire resisting characteristics of electric cables
IEC 364-5-523	Electrical installations of buildings—Current carrying capacities of wiring systems
IEC 724	Guide to the short-circuit temperature limits of electric cables with a rated voltage not exceeding 0.6/1 kV
IEC 853	Calculation of the cyclic and emergency current rating of cables
	853-1: Part 1 Cyclic rating factor for cables up to and including 30 kV
	853-2: Part 2 Cyclic rating of cables greater than 30 kV and emergency ratings for cables of all voltages
IEC 949	Calculation of thermally permissible short-circuit effects, taking into account non-adiabatic heating effects
IEC 986	Guide to the short-circuit temperature limits of electric cables with a rated voltage from 3 kV to 30 kV

a licence to a manufacturer to use the mark are identical in all the participating countries, based on initial approval, from inspection of manufacturing and testing facilities and testing of samples, and subsequent surveillance in the form of periodic testing of samples. In the UK the approval organisation is British Approvals Service for Cables (BASEC). Under the Low Voltage Directive, of which the requirements have to be incorporated in the national laws of the Common Market countries, it is required that electrical equipment for voltages up to 1000 V should be accepted in each country as meeting the safety requirements of the directive if it conforms with a harmonised standard. Moreover, it should be presumed to conform with the relevant harmonised standard if the manufacturer qualifies for the use of the common mark.

31.2 Cable components

31.2.1 Conductors

31.2.1.1 Conductor materials

Materials and the form in which they are used comprise normally: (a) copper in solid form up to 2.5 mm^2 for wiring cables and power cables, and 150 mm^2 for mineral insulated cables, and in stranded form up to 2000 mm^2; (b) tinned copper similarly but in a narrower range for wiring cables; (c) solid aluminium up to 300 mm^2 and stranded aluminium up to 2000 mm^2; (d) an aluminium sheath as a concentric neutral conductor; and (e) lead and aluminium sheaths and steel wire or strip as an earth conductor.

In the USA some use was made of sodium by filling it into an insulating polyethylene tube. Although technically satisfactory, handling difficulties were found to outweigh the economic advantage. Another, more novel, application, which has been proved technically but not yet brought to commercial fruition, is the use of niobium alloys with superconducting properties at very low temperatures. Further reference to this subject is made in Section 31.5.

Some typical physical and electrical properties of the metals used in cables are given in *Table 31.2*.

Copper Because of its excellent conductivity, reasonable price and ease of working into rod and wire, copper has always been the basic material of the cable industry. Until the 1950s it was virtually without any challenger. Except when tensile strength is important, notably for self-supporting overhead line cables, it is always used in the annealed condition, partly to obtain flexibility, but more because conductivity decreases significantly with degree of working. Impurities also affect conductivity and are kept to a maximum of 0.01%.

Tinned copper wires have been used for wiring and flexible cables, partly to improve solderability but mainly to prevent interaction between copper and the sulphur present to produce vulcanised rubber insulation. With the substitution of synthetic insulation for natural rubber the use of tinned conductors has diminished.

Aluminium Although having only 61% of the conductivity of copper and, hence, for equal conductance requiring a conductor area 1.6 times that of copper, the low density of aluminium results in the actual weight of a comparable

Table 31.2 Physical properties of metals used in cables (20°C)

Property	*Copper*	*Aluminium*	*Lead*
Density (kg/m^3)	8890	2703	11 370
Resistivity (μΩ-m)	0.01724	0.02826	0.214
Res.-temperature coefficient (per °C)	0.0039	0.0040	0.0040
Thermal expansion coefficient (per °C)	17×10^{-6}	23×10^{-6}	29×10^{-6}
Melting point (°C)	1083	659	327
Thermal conductivity (W/m-K)	380	240	34
Ultimate tensile strength			
Soft temper (MN/m^2)	225	70–90	—
$\frac{3}{4}$H to H (MN/m^2)	—	125–205	—
Elastic modulus (GN/m^2)	260	140	—
Hardness			
Soft (DPHN)	50	20–25	5
$\frac{3}{4}$H to H (DPHN)	—	30–40	—
Stress fatigue endurance limit (MN/m^2)	±65	±40	±2.8

$\frac{3}{4}$H, Three-quarters hard; H, hard.

conductor being only half that required with copper; i.e. the current-carrying capacity of an aluminium conductor is 78% of that of a copper conductor of equal area, and 1 t of aluminium does the work of 2 t of copper. However, as the size is larger, the amounts of all the other materials in the cable are increased. Any economic advantage, therefore, varies with the relative metal prices and with the type of cable. When the conductor metal is a small fraction of overall volume, the use of aluminium is uneconomical.

Another difference from copper is that, whereas solid copper conductors become difficult to handle above about 16 or 25 mm^2, solid aluminium conductors can be handled easily up to 240 or 300 mm^2, which further keeps dimensions to a minimum. Aluminium in a soft temper is quite suitable for these solid conductors, but lacks strength in wire for stranded conductors. However, as aluminium does not suffer the same penalty as copper in loss of conductivity with work-hardening, it is satisfactory to use a broad $\frac{3}{4}$-hard temper for aluminium wire.

Certain impurities lower the conductivity of aluminium, but the effect is not as great as with copper. Subject to control, the basic grade of 99.5% purity produced by electrolytic refining is satisfactory and appropriately defined in cable standards.

Apart from the economic factor and its reduced weight, aluminium has no real advantage over copper for cables and it also suffers from a positive disadvantage. This relates to the protective oxide skin which is always present and which requires somewhat greater care to be taken when making soldered joints. To a large extent, compression joints have replaced soldering and have been designed to deal satisfactorily with this aspect.

31.2.1.2 Constructions

Metric conductor sizes, standard in the UK since 1969, are used in all countries other than the USA. *Table 31.3* shows some comparisons.

Stranded conductors are available in circular form up to 2000 mm^2 and, for the lower voltage ranges, in sector-shaped contour up to 630 mm^2 in specific cases. The minimum number of wires is defined in IEC 228. Shaped conductors are normally pre-spiralled so that the cores fit easily together without applied twist in the laying-up operation. To provide a smooth surface and reduce the dimensions, it is now the practice to compact stranded conductors by a rolling process.

In the UK, aluminium conductors for cables up to 1.8/3 kV have largely been solid and of sector shape for multicore cables. Sector corner radii are fairly sharp to produce a compact construction. In Europe larger radii have been used and this form is adopted for high-voltage applications. Four 90° sector conductors laid-up together are used for 380–960 mm^2 solid-sectoral circular cables to obtain increased flexibility.

The Milliken construction is frequently used for circular conductors in sizes of 960 mm^2 and above to reduce skin effect and improve flexibility. Four or six individual sectors are used with a layer of insulation tape over alternate sectors.

The comments above refer largely to wiring and distribution cables. Different conditions apply to conductors for transmission cables, particularly of the oil filled type. Single-conductor sizes extend to 2500 mm^2 and an oil duct is required in the centre. The duct may be formed by laying the conductor strands round an open metal helix or, more usually, by creating a self-supporting centre by the use of curved segmental sections. Succeeding layers may consist of wires or flat strips, the latter improving both compactness and flexibility. For 18/30 kV cables, shaped conductors are oval rather than sectoral; circular conductors only are used at all higher voltages.

Table 31.3 Conductor data

Standard metric size (mm^2)	*Equivalent Imperial size* (in.^{2})	*Maximum d.c. resistance at* 20°C (Ω/km)	
		Aluminium	*Copper*
1.5	0.0023	—	12.1
2.5	0.0038	—	7.41
4	0.0061	7.41	4.61
6	0.0092	4.61	3.08
10	0.016	3.08	1.83
16	0.025	1.91	1.15
25	0.038	1.20	0.727
35	0.053	0.868	0.524
50	0.072	0.641	0.378
70	0.104	0.443	0.268
95	0.144	0.320	0.193
120	0.182	0.253	0.153
150	0.224	0.206	0.124
185	0.281	0.164	0.0991
240	0.369	0.125	0.0754
300	0.463	0.100	0.0601
400	0.592	0.0778	0.0470
500	0.746	0.0605	0.0366
630	0.963	0.0469	0.0283
800	1.23	0.0367	0.0221
1000	1.55	0.0291	0.0176

31.2.2 Insulation

31.2.2.1 Thermoplastic and elastomeric materials for wiring cables

In the general wiring-type cable field and for many power cables the major insulants in general use are either thermoplastic or elastomeric materials. In making the choice, several factors have to be considered. No insulant is ideal: a compromise is sought between processability, performance and economics.

An elastomeric material is one which returns rapidly to approximately its initial dimensions and shape after deformation at room temperature by a weak stress. Under such conditions a thermoplastic material shows permanent deformation. Conventional elastomeric compounds need to be cross-linked by vulcanisation, generally by chemical methods, to provide them with the characteristics which were typified by rubber compounds.

Examples of elastomeric materials are natural rubber (NR), ethylene propylene rubber (EPR), cross-linked polyethylene (XLPE), polychloroprene (PCP), chlorosulphonated polyethylene (CSP) and silicone rubber (SR). Of these, EPR and XLPE are the most common because they combine the flexibility and electrical properties of natural rubber with a higher operating temperature limit and easier strippability.

Examples of thermoplastic materials are polyvinyl chloride (PVC), polyethylene (PE) and polypropylene (PP). Until recently, PVC was the most usual insulant because of its uncomplex processability, good general-purpose performance and economic advantage. By adjustment of formulation, PVC compounds can meet a variety of requirements. Their robustness, relative chemical inertness, good ageing and attractive appearance in a range of colours have led to wide use not only as an insulant, but also as bedding for armour wires and for sheathing. PVC hardens at temperatures below 0°C but will recover its flexibility on returning to normal ambient temperatures. General-purpose PVC compounds are limited to a maximum conductor operating temperature of 70°C.

Where electrical properties are paramount, e.g. radiofrequency cables, polyethylene is the preferred insulant. It is also a more effective water barrier where a water-resistant property is important.

Elastomeric compounds are of advantage for long-term operation at temperatures higher than those that PVC can tolerate. EPR and XLPE can operate up to 90°C continuously, and silicone rubber at 150°C continuously. Elastomers are also the first choice where flexibility combined with mechanical ruggedness is required. Applications for this type of material range between flexible cords for domestic flat-irons, fixed wiring and power cables, and flexible trailing cables for mines and quarries.

Where light weight and high operating temperatures are of paramount importance, fluorocarbon tapes or extrusions are adopted, particularly in the aircraft industry. Among the materials used in such high-performance cables are polytetrafluoroethylene (PTFE), fluorinated ethylene propylene (FEP), ethylene tetrafluoroethylene (ETFE) and polyimide/FEP tapes. They are characterised by low coefficient of friction, excellent electrical properties, resistance to chemical attack and stability at elevated temperatures. Besides use in aircraft applications, these materials have also been used in specialised radiofrequency cables and equipment wires.

Glass fibre in the form of lappings and braids is the insulation in a range of cables and cords for use, for example, in luminaires.

Tables 31.4 and *31.5* provide data on physical and electrical properties for a range of thermoplastic and elastomeric insulating materials.

31.2.2.2 Thermoplastic and elastomeric materials for power distribution cables

In the distribution field, PVC is being displaced by the thermosetting material XLPE as the most widely used insulation and, as flexibility is not important, XLPE is more favoured than the possible alternative, EPR. (It is more usual to refer to cross-linking rather than curing or vulcanising; and to thermosetting rather than to elastomeric materials.)

Table 31.4 Physical properties of polymeric materials

Material	*Type*	*Tensile strength (min.)*(N/mm^2)	*Elongation at break (min.)(%)*	*Limiting temperature** (°C)	
				Rating	*Installation*
Thermoplastic †					
Polyvinyl chloride	TI 1	12.5	125	70	0
Polyvinyl chloride	TI 2	18.5	125	70	0
Polyvinyl chloride	TI 2	10	150	70	−10
Polyvinyl chloride	TI 4	7.5	125–150	85	0
Polyvinyl chloride	TI 5	12.5	125	85	0
Polyethylene LD	PE 03	7	300	70	−60
Polyethylene LD	PE 2	7	300	70	−60
Polyethylene HD		37	500	80	−40
Polypropylene		37	400	80	−10
LSF‡⇐	BS 6724	10	100	70	−10 to 0
Elastomeric§					
General-purpose GP rubber	EI 1	5.0	250	60	−45
Heat resisting GP rubber	GP 1	4.2	200	85	−45
Heat resisting GP rubber	GP 2	4.2	200	85	−45
Heat resisting MEPR rubber	GP 4	6.5	200	90	−45
Flame-retardant rubber	FR 1	5.5	200	85	−30
Flame-retardant rubber	FR 2	5.5	200	85	−30
OFR rubber	OR 1	7.0	200	85	−30
Silicon rubber	EI 2	5.0	150	150	−55
Ethylene vinyl acetate	EI 3	6.5	200	105	−25
Hard ethylene propylene rubber		8.5	200	90	−40
Cross-linked polyethylene		12.5	200	90	−40
Fluorocarbons					
Polytetrafluoroethylene		24	300	260	−75

* Maximum temperature for sustained operation and minimum temperature for installation.

† BS 6746 for PVC types and BS 6234 for polyethylene types.

‡‘LSF’ (low smoke and fume) is a generic term and the characteristics given are typical of the material which is still under active development and improvement to meet stringent performance requirements.

§ BS 6899 for GP rubber types, EPR types, and cross-linked polyethylene.

Table 31.5 Electrical properties of polymeric materials

Material	*Type*	*Volume resistivity (min.) at* 20°C (Ω-m)	*Permittivity at* 50 Hz	*tan δς at* 50 Hz
*Thermoplastic**				
Polyvinyl chloride	TI 1	2×10^{11}	6–7	0.1
Polyvinyl chloride	TI 2	1×10^{12}	4–6	0.08–0.1
Polyvinyl chloride	TI 2	2×10^{11}	6–7	0.09–0.1
Polyvinyl chloride	TI 4	1×10^{9}	5–6	0.07–0.13
Polyvinyl chloride	TI 5	5×10^{11}	6	0.9
Polyethylene LD	Pe 03	1×10^{16}	2.35	0.0003
Polyethylene LD	PE 2	1×10^{16}	2.35	0.0003
Polyethylene HD		1×10^{16}	2.35	0.0006
Polypropylene		1×10^{16}	2.25	0.0005
LSF‡⇐	BS 6724	1×10^{12}	—	—
Elastomeric†				
General-purpose GP rubber	EI 1	2×10^{12}	3–4.5	0.01–0.03
Heat resisting GP rubber	GP 1	7×10^{12}	3–4	0.01–0.02
Heat resisting GP rubber	GP 2	1×10^{13}	3–4	0.01–0.02
Heat resisting MEPR rubber	GP 4	7×10^{12}	3–4	0.01–0.02
Flame-retardant rubber	FR 1	5×10^{12}	4.5–5	0.02–0.04
Flame-retardant rubber	FR 2	1×10^{13}	4–5	0.015–0.035
OFR rubber	OR 1	1×10^{10}	8–11	0.05–0.10
Silicone rubber	EI 2	2×10^{12}	2.9–3.5	0.002–0.02
Ethylene vinyl acetate	EI 3	2×10^{12}	2.5–3.5	0.002–0.02
Hard ethylene propylene rubber		2×10^{13}	3.2	0.01
Cross-linked polyethylene		1×10^{14}	2.3–5.2	0.0004–0.005
Fluorocarbons				
Polytetrafluoroethylene		1×10^{16}	2	0.0003

* BS 6746 for PVC types and BS 6234 for polyethylene types.
† BS 6899 for GP rubber types, EPR types and cross-linked polyethylene.
‡‘LSF’ (low smoke and fume) is a generic term and the characteristics given are typical of the material which is still under active development and improvement to meet stringent performance requirements.

Although EPR can be produced in a hard grade (known as HEPR) with properties similar to those of XLPE, it is more expensive than XLPE and, as is common with rubber compounds, it contains a large number of ingredients. XLPE comprises merely polyethylene, an antioxidant and a cross-linking agent. The cross-linking can be accomplished by a variety of methods. Until recently the most common has been to mix an organic peroxide (dicumyl peroxide) with the PE and to extrude the insulated conductor into a large catenary tube containing steam under high pressure. For high-voltage cables, where minimum moisture content of the insulation may be important, radiant electrical heating may be substituted for steam and the CCV tube filled with nitrogen.

In the CCV extrusion process there is a wastage when starting and stopping; increasing use is being made of a process by which the extrusion can be in a conventional line, as for thermoplastic materials, and the cross-linking can be accomplished by a different chemical process. A silane and an accelerator are blended with the polyethylene and cross-linking is achieved by immersion of the insulated conductor in hot water. One method requires a separate process for preparing the graft polymer (which has a limited storage life), but in the Monosil process all the ingredients are blended in the hopper of the extrusion machine.

31.2.2.3 Impregnated paper

Layers of paper tapes are lapped around the conductor and the cable is dried and impregnated before application of the metal sheath which is required to keep the insulation dry and undamaged in service. The paper consists of a felted matt of long cellulose fibres derived from wood pulp. Washing of the fibres, both at the pulp stage and after formation of the sheet, is an important factor in the control of the properties of paper for cables. Large quantities of water are used, and for paper intended for the highest-voltage cables this water has to be deionised to ensure minimum power factor.

Impregnants have traditionally been based on mineral oils thickened with gum rosin to limit drainage from the insulation at service operating temperatures and to provide resistance to oxidation. It is now more usual to substitute materials such as microcrystalline waxes to obtain improved non-draining performance. For high-voltage internal pressure cables of the oil filled type it is necessary to use an impregnant with very low viscosity; hence, the highly refined mineral oils formerly used have been replaced by synthetic alkylates of dodecylbenzene type.

The electrical properties of the dielectric are not critical for voltages up to about 10 kV, but, at this level, ionisation in any air spaces becomes important and the impregnation process must ensure that butt gap spaces between paper tapes are well filled with impregnant. Impregnated paper itself, in sheet form, has a high electric strength, around 10 MV/m in short-time a.c. tests.

31.2.3 Armour

31.2.3.1 General wiring cables

When cables are not installed in conduit or trunking, they may require armour, most commonly provided by galvanised steel wire (GSW) helically applied in a single layer and known as 'SWA' (single-wire armour).

Pliable wire armour finds application for portable cables in quarries and mines. It consists of stranded seven-wire bunches of GSW applied helically in a similar manner to SWA, but with a shorter lay length, thus providing good mechanical protection with improved flexibility, enabling the cable to be moved without affecting performance.

Braided GSW armour is mainly used in cables for ships and off-shore applications. It has the advantage of easier installation in complex cable runs. For single-core cables in a.c. circuits, where magnetic effects could cause high losses with GSW, tinned phosphor-bronze wires are normally used.

31.2.3.2 Supply distribution cables

Most types of power cable require mechanical protection and/or an earth conductor to carry fault currents. For most distribution cables this is provided by SWA. Cables with aluminium sheaths seldom require armour. Cables with lead sheaths may be armoured with steel tape, which is cheaper, but SWA is preferred in the UK for the heavier, higher voltage cables for 10 kV and upwards because it increases corrosion resistance and the longitudinal strength of the cable for installation purposes. Steel tape is normally protected against corrosion by bitumen, but if better corrosion resistance is required, the tape may be galvanised.

When additional mechanical or tensile strength is needed, as for river crossings, coal mines or long vertical runs, a double layer of steel wires may be employed. Single-core a.c. cables are rarely armoured, but if armour is necessary, it can be provided by non-magnetic tape or wire, normally of aluminium.

31.2.4 Oversheaths and protective finishes

31.2.4.1 General wiring cables

The choice of sheathing material depends on its environmental performance. Matters for consideration are: ambient temperature; flexibility; resistance to abrasion, water, oil and other chemicals; performance under fire conditions; and compatibility with other materials with which a cable is in contact during its operational life. The sheathing material must also be chemically compatible with the other materials used in the cable both during and after processing. While insulants are chosen primarily for their electrical characteristics, sheaths are selected on their physical properties. Thus, not all insulants are suitable sheaths. However, in general, insulation and sheath materials are similar: e.g. a thermoplastic sheath protecting thermoplastic insulation.

Elastomeric sheathing materials include natural rubber, used for ordinary-duty domestic flexible cords; and synthetic rubbers such as NBR/PVC, PCP(OFR), CSP(HOFR). PCP is classed as an OFR material (oil resistant and flame retardant), and CSP as an HOFR material (heat resistant, oil resistant and flame retardant). These materials can be specially formulated to meet special requirements, e.g. improved water resistance, extra flame retardance, improved mechanical properties. Most elastomeric materials have a wider operational temperature range than thermoplastics and their superior performance under adverse environments (as found in mines and quarries) makes them first choice for such applications. Not only are they abrasion resistant but also they are flexible over a wide range of temperatures.

In thermoplastic insulated cables, e.g. general wiring cable and radiofrequency cables, the predominant sheathing material is PVC in various formulations. Polyethylene is chosen where cables are in water or operate at subzero temperatures. For more specialised applications nylon and polyurethane are also used. Nylon has application where the cable is likely to be attacked by hydraulic fluids and is also claimed to be a termite barrier. Polyurethane is being introduced into designs which call for a cable having good flexibility, abrasion and impact properties under arduous low-temperature conditions.

Cables with high-performance insulation materials can be sheathed with conventional thermoplastic or elastomeric materials, but more commonly they have sheaths similar to the insulation composition. PFTE, FEP and a combination of PTFE and polyimide/FEP tapes are often used. These are light in weight, and are resistant to abrasion and cut-through, even when applied with small radial thicknesses.

31.2.4.2 Distribution cables and transmission cables

Even when armoured cables are installed indoors above ground, it is unusual for the armour to be left bare, because (a) there are few environments where corrosion will not occur, (b) without an outer covering the armour layer may become disturbed during installation, and (c) few cables are above ground for their whole length. Nowadays, to make cables easy to handle and provide a clean finish, a polymeric sheath is usually applied overall.

Where cables are buried, the soils can be aggressive and cable life may depend on the degree of protection provided. For many years, lead sheathed cables depended on bituminised textiles as a bedding *under* the armour, and a serving of two layers of hessian or a layer of helically applied jute strings *over* the armour. Bitumen is provided over each layer, as well as to flood the armour, and use of the optimum grade of bitumen at each stage is important. Today, extruded sheathing—usually either PVC or sometimes (for toughness) polyethylene—has largely replaced bitumen finishes, even on lead sheathed cables, and all new designs of cable introduced during the last three decades have had PVC or polyethylene oversheaths. The successful introduction of aluminium for sheathing depends entirely on adequate protection with extruded plastic oversheathings.

For exposure to sunlight, plastic oversheaths should be black, but in other situations colours are sometimes used as a means of identification. To this end the sheaths are also embossed with the words 'Electric Cable', the voltage and sometimes further details of the cable construction, the manufacturer and the year of manufacture. Even though very tough, plastic oversheaths can be damaged during installation and, if they protect an aluminium sheath, it is important to carry out an inspection before backfilling. In the case of expensive transmission cables it is usual to test for damage by applying a graphite coating over the plastic and then to carry out a 10 kV d.c. test between this electrode and the metal sheath.

One of the disadvantages of PVC concerns cables in buildings or in tunnels. Although PVC is basically flame retardant, if a serious fire develops, it can transmit flame and will decompose, with evolution of noxious acidic fumes and dense smoke. Alternative synthetic compositions—sometimes known as 'low smoke and fume' (LSF)—are now becoming widely available to overcome this hazard.

31.3 General wiring cables and flexible cords

31.3.1 Wiring system cables

A wiring system cable is usually regarded as the final link in the transmission network which begins in the power station and ends at the socket outlet in the home, office or factory work-bench. Although rubber was for many years the insulant for wiring cables, it has been superseded almost entirely by PVC, or for some applications by mineral insulated cables. Although several alternatives have been tried, plain annealed copper remains the sole conductor material for wiring cables up to 16 mm^2.

PVC was originally developed in Germany in the 1930s. It took many years for PVC to be universally accepted for use in wiring cables. As PVC compounds improved, the thickness of insulation was progressively reduced. It now approaches half of that originally used and is acceptable for twice the operating voltage. The insulation is designed to have a higher tensile strength, resistance to deformation and a higher insulation resistance than the sheath. Sheathing compounds are usually formulated to provide good abrasion resistance and yet have easy-tear properties to facilitate stripping at terminations.

When vulcanised rubber insulation was used with copper conductors, it was necessary to tin the copper to prevent chemical reaction with the rubber. With PVC insulation, plain conductors became universally accepted. The introduction of cables to metric standards in 1969 achieved a greater degree of international alignment.

PVC wiring cables in common use in the UK are of two basic designs. One is the single-core unsheathed cable, used in conduit or trunking, the numbers of cores varying from 2 up to as many as 38. The other is the insulated and sheathed cable, available in single-core, two-core and three-core versions, the two-core and three-core cables (made in a flat formation) having the option of a bare earth continuity conductor.

In the late 1980s the need for improved fire performance wiring cables led to the introduction of an alternative insulating and sheathing material; special characteristics of these cables are the low emission of smoke and corrosive gases when affected by fire and a rapidly increasing use of these materials is foreseen.

Whilst EPR is used as the insulation for flat-twin festoon lighting cables and insulated, textile braided and compounded single-core conduit wire, another elastomeric insulation having low smoke and fume characteristics is now becoming widely available for wiring (conduit) cables (BS 7211). These cables have an upper temperature limit of 90°C, and although voltage drop may limit their advantages in some situations, it is anticipated that they will progressively replace PVC insulated cables in the next decade.

31.3.2 Flexible cords

Most flexible cords are designed for and used in domestic premises as the supply lead from the socket outlet to portable, and some fixed, appliances. The range of domestic appliances covers such items as can openers, food freezers, hairdryers, vacuum cleaners, microwave ovens, towel rails and washing machines. There is a flexible cord to suit each one.

The major cable insulants used in flexible cords are natural or synthetic rubber and PVC. Where flexibility is a prime requirement (for instance, on an electric iron), a rubber insulated type is most suitable. However, PVC is less expensive, has an attractive surface finish and is available in a larger range of colours. It is therefore the first choice for the supply lead to most domestic appliances.

Nowadays, many appliances carry the approval mark of the British Electrotechnical Approvals Board (BEAB) for House-hold Equipment, which means that in most cases the supply lead should meet the requirements of BS 6500: Insulated flexible cords. Independent auditing of flexible cords to BS 6500 by the British Approvals Service for Cables (BASEC) provides an assurance of the integrity and reliability of the cablemakers' products. This standard specifies cords having elastomeric, thermoplastic and glass-fibre insulants. These types are designed for a wide range of applications with upper temperature limits varying from 60 to 185°C. The natural rubber insulated types are rated at 60°C and cover the twin-twisted and textile braided cord in use for many years for pendant flexibles. The recommendations for this application are now 85 and 150°C rubber or glass-fibre insulated types. The UDF cord is used on electric irons and consists of rubber insulation and a thin rubber sheath, over which is applied a semi-embedded textile braid.

Currently under technical evaluation are cords employing cross-linked PVC for applications such as electric irons, where the ability to withstand contact with a hot surface is desirable combined with flexibility and abrasion resistance.

The PVC insulated and sheathed cords in BS 6500 are rated at 70°C and are available in light- and ordinary-duty versions for such applications as table lamps, television sets, washing machines and refrigerators. For areas where temperatures are likely to exceed those satisfied by standard PVC, but not in excess of 85°C conductor temperature, BS 6141 defines a range of heat resisting PVC cords. Examples of use are connections to immersion heaters and night storage heaters. For situations involving contact with oil and grease, there is an elastomeric insulated and sheathed flexible cord, rated at 85°C and having an oil resisting sheath. The glass-fibre insulated types in BS 6500 are primarily intended for use with light fittings (luminaires) and other situations where the cord is not subject to mechanical damage or continuous flexing. They can be used at temperatures up to 185°C.

Some flexibles are available in cut and trimmed form as an aid to the handyman, i.e. the cores are cut to length and the sheath and insulation are stripped. A few are also available in coiled extensible form, the most common type being for electric shavers.

There are also flexible cables designed for more harsh industrial environments. Three common types have a copper wire or galvanised steel wire braid, or a spiral steel strip, over the inner sheath, and an outer oversheath. The copper wire braided type is mainly used for portable hand lamps and where a flexible cable is required in certain flameproof installations. The steel wire braided and steel strip types are utilised where both flexibility and mechanical protection are required.

For situations where a limited degree of flexibility at low temperatures is required, special PVC formulations are available. Areas such as temporary supplies to traffic lights and portable tools on building sites are examples.

Some, but not all, designs of flexible cord used in the UK have been harmonised in accordance with the CENELEC procedure and for these the same design is standardised throughout the Common Market countries.

31.3.3 Control and instrumentation cables

Whereas power cables are the 'arteries' of industry, control and instrumentation cables are its 'nerves' and are used for the control of equipment and data collection. They range from single-core cables used in the wiring of control panels and switchgear, to the complex control and instrumentation cables used in power stations and petrochemical sites.

At one end of the scale are the single-core cables used within machine tools and switchgear. Where normal ambient temperatures are involved, PVC insulation is employed. At the other end of the scale is, for example, a North Sea oil terminal utilising 500 km of cable and connecting as many as 2000 instruments measuring flow rates or liquid levels in storage tanks. Control cables have copper conductors and are laid up in multicore or multipair formation, each core being separately identified.

Thermocouple cables (*Figure 31.1*) are used for connecting the thermocouple to its measuring instrument. The term 'thermocouple cable' is often used to describe both extension and compensating cables. Extension cables utilise conductors of the same alloys or metals as the thermocouple itself, while compensating cables utilise conductors of cheaper material although having similar thermoelectric characteristics. The normal conductor materials or alloys used are constantan, copper, iron, copper–nickel, nickel–chromium and nickel–aluminium.

Several national standards exist for the colour identification of insulation and sheath. Unfortunately, there is not yet a recognised international standard.

To prevent electrical interference in both control and thermocouple circuits within and between the cables, metallic screens (usually in the form of tapes) are applied over the individual pairs and/or the laid-up cores.

The finish of any type of cable to be buried in a petrochemical environment has to be given special consideration because of the presence of hydrocarbons in the soil. The steel wire armour which is normally applied as a mechanical protection on underground cables is not a barrier to the ingress of hydrocarbons into the heart of the cable: this is best achieved by applying a lead sheath. Thus, such control and instrumentation cables contain a PVC bedding, lead sheath, another PVC bedding, single wire armour and PVC oversheath. BS 5308 gives details of control cables with this type of protection. North American practice is to use an aluminium sheath.

Safety in hazardous areas has to be carefully considered. Intrinsic safety is a protective technique which ensures that any electrical sparking which may occur is incapable of causing an ignition of gas or vapour. Although a cable itself will rarely cause an explosion, it is possible for gas or vapour to percolate along the interstices of cable from a hazardous zone to a non-hazardous one. This problem can be cured by use of a stopper box or sealing gland. Reference should be made to BS 5345 for specific details. The British Approvals Service for Electrical Equipment in Flammable Atmospheres (BASEEFA) test and certify intrinsically safe and flameproof equipment.

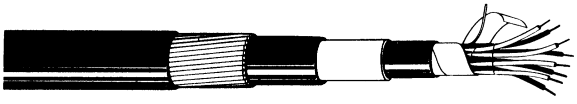


Figure 31.1 Multi-pair thermocouple cable with PVC insulation, pair screening, lead sheath and wire armour

31.3.4 Cables for electronic applications

Cables for electronic applications are single-core and multicore equipment wires, and radiofrequency cables. Equipment wires are usually regarded as insulated single conductors with or without a screen and sheath. They can also be supplied in flat formation as multicores or multipairs, often with a transparent backing. The space between conductors is precisely controlled to ensure consistent electrical characteristics and to assist in termination. Multicore equipment wires generally have PVC or PTFE insulation and sheath in a range of conductor sizes and number of cores, unscreened and screened.

Radiofrequency cables transmit high-frequency signals at minimum loss. Calculation of the optimum cable design involves operating frequency, capacitance, velocity ratio, impedance and attenuation. With the exception of a few twin and special designs, radiofrequency cables are normally *coaxial*; they have an inner conductor, insulation, outer concentric conductor forming the screen, and sheath. Also used for radiofrequency applications are PE and PTFE insulated multipair cables for use in computer interfacing.

For conductors, plain annealed copper is most common but even this is available in several grades, each conferring special properties. It is possible to draw copper to extremely fine sizes because of its good ductility. By leaving copper in the hard-drawn condition, additional tensile strength may be obtained at the expense of elongation. Where d.c. conductivity is of secondary importance, composite conductors incorporating a high-tensile-steel core are employed in miniature cables.

Insulants used in electronic cables vary between PVC, PE, PPE, PVF_2, PVF, ETFE, FEP and PTFE (see *Table 31.4*). The choice of insulant is an optimisation of performance and economics. In radiofrequency cables, where electrical performance is paramount, the insulation is usually either PE or PTFE. However, chemical, mechanical and thermal performance must also be considered.

Screens are applied to prevent electrical interference between circuits or to control the amount of pickup by, or leakage from, a cable. Braided and lapped wires, tapes, tubes, foils and films are among the screening materials used, depending on the application.

Sheaths are applied to act as protection. A thermoplastic or fluorocarbon material is most common. In more demanding environments it may be necessary to have additional protection over the sheath in the form of steel wire braids or armour.

Although there are many specifications covering electronic cables, manufacturers have their own standards.

In the optical fibre cable, the signals are transmitted optically rather than electrically. The conductor is made of a high-quality glass-fibre which transmits light. The main advantages of optical fibre cables compared with conventional metallic conductor cables include low weight, small volume, increased system capacity, freedom from electromagnetic interference and improved security.

31.3.5 Arc welding cables

BS 638: Part 4 Specification for welding cables, includes single-layer EPR, CSP or CPE (HOFR) 85°C rubber to

BS 6899 and a two-layer covering, EPR and CSP or CPE to BS 6899.

The conductor is made up of a large number of small copper or aluminium wires, usually with a separator between conductor and insulation to make the cable supple. Single-conductor cables meet the majority of requirements for connection to electrode holders, arc-welding guns or leads for both manual and automatically controlled metal arc-welding equipment, or to form extension or return leads. Multicore cables are sometimes required for connections to the distribution boxes of multi-operator equipment.

Because of the variable periods of operation, current ratings have to be derived specifically for arc-welding cables and are contained in BS 638: Part 4. The period during which current flows varies from periodic to continuous, according to the application. The longer the period of use, the greater the conductor heating effect, so that current ratings are reduced as the operating cycle lengthens. The operating (or duty) cycle is defined as the time a cable operates in each 5-min period expressed as a percentage: e.g. up to 1.5 min operation in the 5-min period is 30%. Duty cycles are classified as shown in the table below.

Duty cycle	*%*
Automatic	≤100
Semi-automatic	30–85
Manual	30–60
Intermittent or occasional	≤30

Excessive voltage drop can occur when long cable lengths are required between the set and the electrode; conductors of larger current rating must then be selected. For flexibility, the final length of cable to the electrode can revert to the area appropriate to the current rating.

31.3.6 Offshore and ship cables

A wide variety of installation conditions and extremes of temperature are experienced in tankers, refrigerated vessels, ferries, trawlers, tenders, passage vessels, dry cargo ships, etc. In the UK the cables used are largely standardised as ethylene propylene rubber (EPR) insulated and chlorosulphonated polyethylene (CSP) sheathed and produced to BS 6883 (*Figure 31.2*). The operating temperature of 85°C provided by this combination has been proved satisfactory over a number of years and has been used extensively for applications on North Sea oil platforms, although in this case braid armoured cable with an outer CSP sheath has been the main standard.

EPR has excellent electrical properties, good corona resistance and good low-temperature flexibility. The CSP sheath provides a tough outer surface which is resistant to weather, oil resistant and flame retardant.

The increasing use of higher power generation on board ship has meant that systems with voltages of 3.3 and 6.6 kv are now in operation. Again the use of suitable designs of EPR insulated CSP sheathed types have proved fully adequate for this service. For oil rig platforms 13.8 kV cables are also in regular operation.

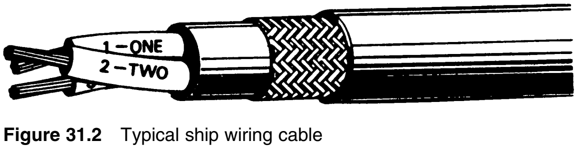


Figure 31.2 Typical ship wiring cable

The conductors of ship wiring cables are of a more flexible construction than comparable land-based cables because of the problem of installing them through complex structures characteristic of shipboard installations. The reactance of a cable operating in an a.c. system depends on many factors, including, in particular, the axial spacing between conductors, and the proximity and magnetic properties of adjacent steel-work. This latter point is of crucial importance in ships and oil rigs. It is desirable to minimise the effect of magnetic induction by means of adequate spacing between cables and steel-work, minimum spacing between conductors, and the avoidance of magnetic materials between single cores in the same circuit.

There are many classification authorities that will approve cables to international and to some national standards. When a ship installation is contemplated, the local surveyor of the chosen classification authority should be consulted before a decision is taken on the particular design to be used.

31.3.7 Aircraft cables

Cables used for wiring aircraft are continuously under development and, as technology improves, new materials are used in this most exacting of applications. The ambient temperature variation within an aircraft is wide and provision has to be made for cables to operate down to −75°C and up to 260°C. In addition, various special fire resisting cables are needed for use in aircraft-engine fire zones.

For the lower temperature a combination of special PVC, glass braid and nylon has been used where a conductor temperature of 105°C is deemed appropriate. Since the mid-1960s, when a miniature range of cables was introduced both for multicore and screened versions, much development work has been undertaken and greater attention has been paid to the effect of cable weight on the performance of the aircraft; as a result, cables with polyimide insulation are being widely used in the aircraft industry for general airframe wiring. This material has been chosen because of its excellent mechanical and electrical properties and its resistance to the various chemical contaminants present in aircraft. Depending upon the type of conductor and coating used for colouring purposes, polyimide insulated cables are approved to operate at conductor temperatures of 150 and 210°C.

The higher temperature ranges of cable consist largely of combinations of PTFE, glass and polyimide, and when using a nickel plated conductor, are approved for operation at conductor temperatures up to 260°C.

Special fire resisting cables for engine-bay wiring are available and comprise combinations of silicone rubber, glass-fibre, quartz fibre, polyimide and PTFE to ensure that circuit integrity can be maintained for a short period during a fire.

31.3.8 Cables for railways

The main application for general wiring cables for railway use is for track signalling. Multicore signalling cables are laid along the trackside. The insulation is a combination of natural and synthetic rubber. A natural rubber layer is applied next to the conductor to provide electrical integrity, and the outer layer is of polychloroprene (PCP) to give each insulated conductor some oil resistant properties. The cores are collected together and covered overall with a thick

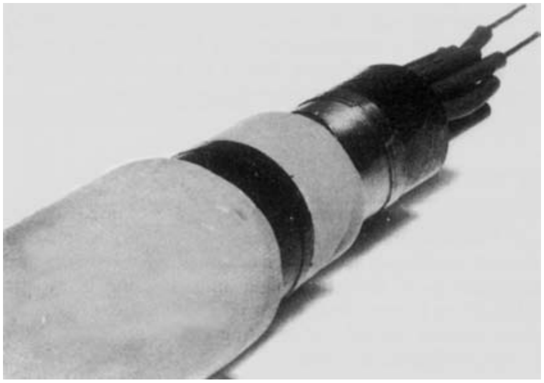


Figure 31.3 Multi-pair signalling cable with materials and construction for optimum flame retardance and freedom from smoke and fumes in a fire

sheath of heavy-duty PCP compound specially chosen for toughness, and weather and abrasion resistance.

Cables for track power feeds on electric systems at medium and low voltage are insulated with EPR and sheathed with chlorosulphonated polyethylene (CSP). Cables for traction and rolling stock are conventionally of EPR/CSP composite insulation, but recent developments have resulted in new materials having low smoke properties and greater resistance to the oils and fluids used in traction and rolling stock.

In the underground system operated by London Transport Executive particular emphasis has been placed on minimising hazards resulting from fire. Designs are now approved which, in a fire, give off fewer toxic products and far less smoke than did previous designs (*Figure 31.3*).

Two-way communication between a control centre and moving trains is now becoming common, even when the trains are in tunnels. It is accomplished by installing suitably designed electric cables near to the track, either at ground level or overhead, to act as elongated aerials. The most common cable for this purpose is the so-called ‘radiating coaxial’ or ‘leaky feeder’.

31.3.9 Cables for mines and quarries

Cables for metalliferous mines and quarries have been standardised for many years. They are essentially flexible and tough, as they need to withstand all the rigours of service in a rugged and rough environment. The range of cables, given in BS 6116, incorporates ethylene propylene rubber (EPR) as insulation, a rubber undersheath, a layer of stranded galvanised steel wires applied as an armour, and a tough weather resistant outersheath of polychloroprene (PCP). Cables are available for 600/1000 V, 3.3 kV and 6.6 kV systems. Some higher voltage installations have been made using cables with individually screened cores and a thick overall PCP sheath, but their use is subject to permission from the relevant authorities.

Cables for use by equipment connection in underground coal mines are manufactured to specifications issued by British Coal and use an EPR compound insulation specially formulated to give good impact strength and crush resistance. Individual core screening is the norm as part of the safety measures necessary for operation at the coal face. The sheath is of PCP, which has excellent mechanical properties and is flame retardant. Low-voltage pliable armoured cables similar to those in quarries are also used for portable supply cables to conveyor loaders, etc., and similar cables suitable for 3.3 kV and 6.6 kV systems are used for connection to transformers.

Thermoplastic insulated cables are also widely used for power, lighting and signalling purposes, and British Coal has now standardised power cables insulated with XLPE and EPR.

An increasing demand for improved communication services in mines and quarries is reflected in the expanding use of two-way mobile radio, radio paging and radio control systems. However, mines and quarries often present situations where free-space propagation is not possible: for instance, propagation in tunnels can be restricted to only a few hundred metres. One solution is the radiating cable or leaky feeder in which signals radiate from the cable rather than from a conventional aerial. Radiating cables have special screens which provide and control the electromagnetic field around the cable so that signals can be picked up by nearby mobile receivers, with communication in the reverse direction also possible. These cables use low-permittivity dielectrics, which, with the special screens, ensure good radiation and transmission characteristics unaffected by the arduous external environmental conditions encountered in mines.

31.3.10 Mineral insulated metal sheathed cables

Mineral insulated metal sheathed cables generally consist of copper conductors insulated with compressed mineral powder, typically magnesium oxide (MgO), and enclosed in a copper sheath.

The traditional method of manufacture is to position the required conductors within a large-diameter tube and fill them with powder using a ramming process, the filled cable then being drawn to the required final diameter by cold drawing through numerous dies with inter-stage and final furnace annealing. More recently, continuous manufacturing methods have been developed in which the sheath is produced from rolled and welded strip into which powder and conductors are introduced, the assembly being reduced to its final diameter by rolling or a combination of rolling and die drawing with inter-stage and final annealing by induction. The annealing of the copper is carried out to restore ductility and conductivity, lost due to work hardening. Moisture ingress into the dielectric is inhibited by control of the chemical activity of the powder, or by additives.

Normally, mineral insulated cables need no further protection over the copper sheath, will withstand high service temperatures, and are impervious to oil and water. Being composed of inert inorganic materials they are incombustible and non-ageing. However, for aesthetic appeal, identification or corrosion protection of cables buried underground or inaggressive industrial environments a thermoplastic outer covering may be applied, which may be typically PVC, but more recently low smoke and fume sheathing materials have been introduced.

Mineral insulated cables are made with 1, 2, 3, 4, 7, 12 or 19 conductors in light-duty (500 V) or heavy-duty (750 V) grades, with conductor areas ranging from 1 to 400 mm^2, and current ratings of 11 A to over 1000 A depending on the cable size and installed conditions.

The current rating of mineral insulated cables, unlike other types of cable, is not determined by the temperature withstand of the insulation, but by the temperature attained by the copper sheath. For normal applications and where the cable has an outer covering, the limiting temperature is

70°C. As for other types of cable installed in buildings, the published ratings are based on an ambient temperature of 30°C with adjustments being needed for other ambient temperatures and grouping as detailed in Section 31.6.

However, bare mineral insulated cable with a copper sheath can operate continuously at sheath temperatures up to 250°C, depending on the type of end seal applied, and when so rated the current rating capacity is greatly increased. At temperatures above 250°C, progressive oxidation of the copper sheath occurs, although the cable can function for limited periods with sheath temperatures in the region of 1000°C.

To make connections the copper sheath is removed at each end and seals provided by the cable's manufacturer are used to seal the face of the insulant and to insulate the exposed conductor tails. The copper sheath provides an excellent low resistance circuit protective conductor, connection to which is achieved by use of compression glands or by use of special end seals with integral earth tails.

For specialised applications such as thermoelectric cables, heating cables, down-well and other transducer cables or for continuous operation at temperatures higher than 250°C, stainless steel or alloy sheaths are also used with nickel, steel, alloy and many other special conductors.

31.3.11 Cables in fire hazard

An area of electric cable technology where much research and development work has been concentrated in recent years is that of the behaviour of cables in fires. Although they may overheat when subject to current overloads or mechanical damage, electric cables in themselves do not present a primary fire hazard. However, cables are frequently involved in outbreaks of fire from other causes which can eventually ignite the cables. The result can be the propagation of flames and production of noxious fumes and smoke. This result, added to the fact that cables can be carrying power control circuits which it is essential to protect during a fire to ensure an orderly shutdown of plant and equipment, has led to a large amount of development work by cablemakers. This work has included investigations on a wide range of materials and cable designs, together with the establishment of new test and assessment techniques.

Although PVC is essentially flame retardant, it has been found that, where groups of cables occupy long vertical shafts and there is a substantial airflow, fire can be propagated along the cables. Besides delaying the spread of fire by sealing ducts at spaced intervals, an additional safeguard is the use of cables with reduced flame propagating properties. Attention has also been focused on potential hazards in underground railways, where smoke and toxic fumes could distress passengers and hinder their rescue. Initially, compounds with reduced acidic products of combustion were incorporated in cables which have barrier layers to significantly reduce the smoke generated. In the meantime, other cablemaking materials have been developed which contain no halogens and which also produce low levels of smoke and toxic fumes as well as having reduced flame propagating properties. These are now incorporated in British Standards such as BS 6724 and BS 7211.

A different requirement in many installations, such as in ships, aircraft, nuclear plant and the petrochemical industry (both on and off-shore), is that critical circuits should continue to function during and after a fire. Amongst the cables with excellent fire withstand performance, mineral insulated metal sheathed cables are particularly suited for use in emergency lighting systems and industrial installations where 'fire survival' is required. As fire survival requirements on oil rigs and petrochemical plants become more severe, new control cable designs have been developed to meet fire tests at 1000°C for 3 h with impact and water spray also applied, and also to have low smoke and low toxic properties.

Figure 31.4 Heat sensor cable

Another novel approach to fire protection in power stations and warehouses is the use of fire detector cables (*Figure 31.4*). These are used in a system which both detects and initiates the extinction of a fire in the relatively early stages of its growth. These cables have also been installed in shops, offices and public buildings, where the cables can be used to operate warning lights or alarms.

The present position in relation to materials is that problems due to smoke and objectionable fumes are dealt with in the case of insulation by heavy additions of aluminium trihydrate in EPR and EVA. For beddings and oversheaths similar addition may be made to ethylene acrylic elastomer, but such compounds do not have the toughness and oil resistance of CSP and PCP compounds.

For aircraft engine components, where cable weight is important, the cable construction is based on silicone rubber plus quartz with PTFE coverings.

For ships' cables, silicone rubber is also used, and where a glass braid is also included, the silica ash enables the IEC 331 test at 750°C to be met. By use of special EPR compounds, the withstand temperature may be increased to 1000°C. The use of mica/glass tapes on conductors provides good high-temperature insulation which is cost effective in comparison with silicone/glass and mineral insulated designs.

31.4 Supply distribution cables

For underground public supply systems and mains distribution in factories, paper insulation has given way to synthetic insulation, except for certain sectors of the public supply.

Since the early 1960s, PVC has been the major insulant for industrial cables up to 3 kV, but this is now changing and XLPE is increasingly finding favour because of its potentially higher operating temperature (90°C). Similarly, for higher voltage industrial applications, XLPE is now the preferred dielectric.

In Europe and elsewhere, consumers are supplied at around 240 V single-phase and 240/415 V three-phase, as required. From the outset, the system for urban areas has been underground with direct burial of multicore cables, and three-phase transformers feeding large groups of consumers through cables along the whole length of every road. The step-down from the transmission grid has moved towards voltages of 19/33 kV and 6.35/11 kV, but international standardisation for cable specifications caters for a full range of 0.6/1, 1.8/3, 3.6/6, 6/10, 8.7/15, 12/20 and 18/30 kV r.m.s. The rounding off of voltages to whole numbers allows for the fact that the designs cater for 20% variation of voltage.

In the USA, and other countries following American practice, cable designs and voltage standards are the same, but the types favoured and the practical utilisation tend to be very different. The supply to the consumer caters for both 110 and 220 V or thereabouts, and except for the innermost areas of cities the distribution is largely by overhead lines. Instead of three-phase transformers, the local supply is from single-phase units at 10–15 kV or higher, using appropriate transformers to obtain the dual consumer voltages. Conventionally, such transformers are pole mounted, and small, as they feed only a few consumers.

Undergrounding on the American system tends to be a replica of the overhead practice by continuation of the use of similar small transformers and merely adding insulation to the overhead line conductors. Extruded polyethylene or XLPE is convenient for this purpose for both low-voltage and high-voltage requirements, and this is why the interest first developed on the American continent. Very simple single-core cable constructions meet the requirement and much high-voltage cable has been installed in which the neutral conductor comprises copper wires applied over the insulation with no outer sheath. In recent years a concerted effort has been made to get away from the unsightly poles and overhead distribution lines with emphasis on 'underground residential distribution' (URD). This embraces the concept of small single-phase transformers outlined above but more emphasis is now being placed on direct burial of cables instead of installation in ducts. With the dual voltage requirement for consumers, the URD concept, together with the use of single-core cables, provides a way of undergrounding overhead networks at minimum cost. It seems unlikely that it will be adopted in countries where systems have long been geared to other practices.

31.4.1 Paper insulated cables

From its introduction at the end of the last century, impregnated paper has given excellent service to the cable industry. Under normal conditions, users have been able to install the cables and then forget about them. Ultimate lives of 50–60 years are common, and the majority of cables have been replaced only because they became too small for the load. The UK supply industry depreciates paper cables over a 40-year life—surely towards the maximum for any industrial plant.

While the basic dielectrics have changed little throughout this century, there have been considerable improvements in quality of materials and manufacturing techniques, and these have led to successive reductions in thickness over the years.

31.4.1.1 Belted and screened constructions

In multicore cables a greater insulation thickness is required between conductors than from conductor to metal sheath. The most economic construction, therefore, is to apply part over the individual conductors and then a small thickness as a 'belt' over the laid-up cores (*Figure 31.5*). The spaces between the cable cores under the belt are filled with jute or paper, but whereas the main insulation consists of paper tapes applied in a controlled manner, the filler insulation has to be softer and less dense to be compressed into the space available. It is therefore weaker electrically, and it will be seen from the pattern of flux distribution in *Figure 31.6* that significant stresses arise in the filler spaces. An even more important effect, to be seen in *Figure 31.6* is that, in addition to the radial stresses through the layers of paper, there is also a tangential stress component along the paper surface. In the tangential direction the electric strength of impregnated paper is only one-tenth of that radially.

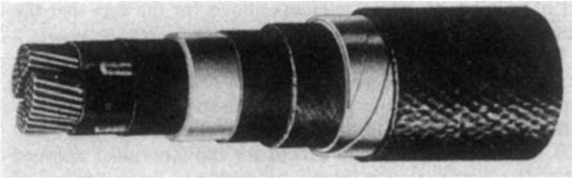


Figure 31.5 Four-core, 1 kv, paper insulated, lead sheathed cable

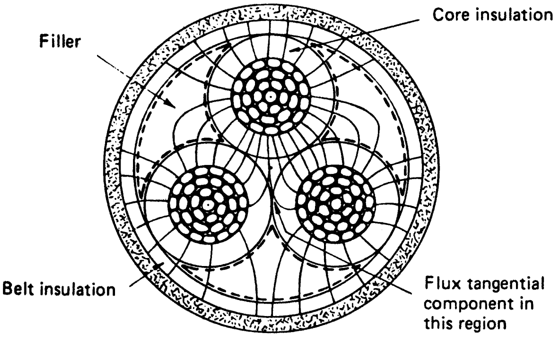


Figure 31.6 Flux distribution in paper insulated belted cable with top conductor at peak potential

When supply voltages were increased to 22 kV and 33 kV in the 1920s, many cable failures occurred due to lack of appreciation of this fact. Hochstädter identified the need for an earthed metallic layer over the insulation to create a purely radial field, a construction subsequently known as 'H'-type or screened. Very little metal is required for the purpose, and while thin copper tapes have been used, the most common form nowadays for multicore cables is a layer of thin aluminium tape or of metallised paper, consisting of aluminium foil on a paper backing. The latter is usually pinpricked to facilitate passage of oil during impregnation. The cores and fillers are held together by a binder of 'copper woven fabric tape' (CWF) containing a few thin wires woven into the weft. This gives protection against scuffing and provides electrical contact between the screen and the metallic sheath. Another construction, used mainly in continental Europe, is the 'HSL' or 'HSA' type, which denotes three lead or aluminium sheathed single cores laid up together and then armoured overall.

The screened construction is optional at 11 kV, but mandatory for higher voltages. Because the dielectric has much better electric field distribution the operating temperature can be increased and higher current ratings obtained. Some 11 kV users find that these factors justify the somewhat greater expense and the extra skill required in jointing.

31.4.1.2 Insulation

The insulation comprises layers of paper tapes, of thickness in the range 0.7–1.9 mm, carefully applied to maintain controlled butt gap spacings and optimum registration between layers. The stress is highest at the conductor surface and may be increased locally, owing to the conductor profile or lack of smoothness. To improve this situation at voltages of 6.35/11 kV and above, a layer of semiconducting carbon paper is applied over the conductor to exclude from the field the small spaces between the wires of the outer layer which otherwise could be sites for discharge.

The thickness of insulation has to be determined by both mechanical and electrical requirements, the former being dominant at the lower voltages, e.g. to withstand bending and to resist damage due to impact. Similarly, at 11 kV, while impregnated paper itself has an a.c. breakdown strength of the order of 10 MV/m, the actual cable design stress is only 2 MV/m, the effects in butt gap spaces being one of the most important factors.

The impregnation of the paper is carried out before application of the metallic sheath by the 'mass-impregnation' process. The cores, on drums or rewound into trays, are inserted into large tanks. These are first evacuated to remove all the moisture in the paper. Hot impregnating compound is then admitted, and the tank is maintained under pressure for a period which depends on voltage, and then cooled slowly to ensure that contraction voids are not present within the insulation.

To obtain good impregnation, the compound viscosity at 120°C should be low, but in the operating temperature range of the cable it needs to be as high as practicable, so that no drainage occurs into the inevitable space under the metallic sheath and into joints. Traditionally, the compound consisted of mineral oil thickened with gum rosin. A problem with such compounds was that the viscosity at maximum operating temperature was not high enough to prevent drainage when cables were installed vertically or on hilly routes, thus leaving the already relatively weak butt gap spaces devoid of impregnant. In the 1950s, BICC developed the 'mass-impregnated non-draining compounds' (MIND) which subsequently became standardised in the UK. In these compounds the viscosity control is obtained by the addition of such materials as microcrystalline waxes and polyethylene to mineral oil.

When single- or three-core cables are operated at 11 kV or higher, some discharge may occur in the space between the insulation and the inside of the metallic sheath. Although this is not unduly detrimental, it is eliminated by the inclusion of a carbon paper over the insulation.

31.4.1.3 Lead sheath

Unalloyed lead is suitable for the majority of armoured cables but is prone to fatigue cracking if subjected to vibration or to high expansion and contraction, as when cables are suspended on hangers or are in manholes. In the UK, when moderate improvement of fatigue strength is required, it is usual to adopt alloy 'E' to BS 801 (0.4% tin, 0.2% antimony). Alloy 'B' (0.85% antimony) has higher fatigue strength and is desirable for conditions involving severe flexing, such as aerial cable installations and cables on bridges. Other alloys are available and are preferred in some countries.

The use of very high-purity lead is detrimental because it can give rise to large grain size and low fatigue strength. Hence, it is always preferable to use lead with impurities up to the limit of 0.1% as permitted by BS 801. Tin and antimony are frequently added for this reason.

31.4.1.4 Armour

The use of armour fulfils a variety of functions, primarily to supply mechanical protection during cable handling and installation, and subsequently in service. Steel taping is the most common, but in the UK a layer of galvanised steel wire is often applied for 11 kV and higher to increase the longitudinal strength of the cable. Galvanised steel tape is popular in tropical countries to provide greater resistance to corrosion; narrow steel strips are often preferred in continental Europe. In general, the resistance to damage is proportional to the armour thickness, and steel strips or tapes are less effective than steel wire.

31.4.1.5 11 kV aluminium sheathed cables

The replacement of a lead sheath by an aluminium sheath with good corrosion protection, such as an extruded plastics oversheath, provides a very economic cable construction eliminating armour. For cables of a type offering other advantages, e.g. to provide a concentric conductor as in the Consac CNE type cable described later, and h.v. cables operating under internal pressure, aluminium sheaths have been widely used since the mid-1960s.

For other types of paper insulated cables, however, aluminium sheaths have not found favour, one factor being the somewhat greater skill required for sheath plumbing when joints other than the cast resin type are used. An exception is that in the UK, the public supply authorities did standardise almost universally in the mid-1970s on 11 kV aluminium sheathed cables and, until recently, XLPE insulated cables were unable to compete on price. There are now clear signs that XLPE will be increasingly used on electricity company networks at 11 kV. However, whilst they were using aluminium sheathed cables, there was a variation of design, some preferring a smooth sheath, whilst others favoured the corrugated form. With a weight reduction of 50% compared to lead sheathed and armoured cables, both types were easy to handle, although there was a problem with pulling 240 and 300 mm^2 cables with smooth sheaths into ducts. Subsequently the preference was for corrugated sheaths.

A significant factor is the effect of thermomechanical forces at straight joints. With the flexible corrugated sheath the position does not greatly differ from that with lead sheathed cables. However, if full load is to be carried regularly, it is desirable with smooth aluminium sheathed cable to employ joints filled with cast resin to overcome possible problems at the plumbs and buckling of cores within the joint sleeve. Most of the 11 kV cable installed by the public supply authority in the UK operates at less than maximum rating because of factors of which the most pertinent is that the cable network is in open rings and is only required to carry full load when the ring is closed because of a fault near to the transformer.

Typical 11 kV aluminium sheathed constructions as used by some authorities are shown in *Figure 31.7*; the design

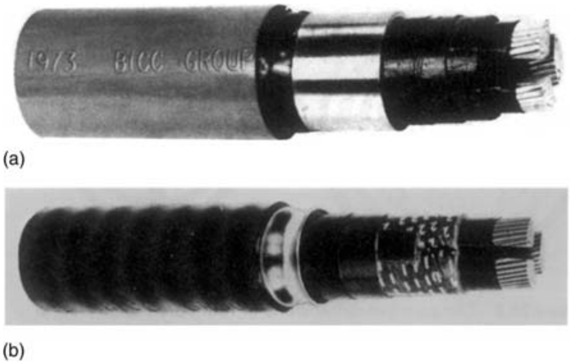


Figure 31.7 Smooth (a) and corrugated (b) aluminium sheathed, 11 kV, paper insulated cables

which was finally standardised was a belted version with corrugated aluminium sheath. In this construction there is a large space between the outside of the insulation and the inside of the sheath. It is important that this should be partially (but not completely) filled with impregnating compound.

31.4.2 CNE cables for PME systems

In this heading CNE denotes a 'combined neutral and earth' conductor in cable construction and PME signifies 'protective multiple earth' applied to a network.

31.4.2.1 PME systems

Traditional practice in UK buried systems involved earthing of the neutral conductor at one point only, at the substation. This meant that the supply cables along the streets required five conductors, three phases, one neutral and one earth (the lead sheath). Consumers normally obtained a satisfactory earth by connection to buried lead water pipes, but when conditions were difficult, as with overhead line sections in rural areas, a practice of multiple earthing of the neutral (MEN) was introduced in the 1940s by burying suitable metal adjacent to poles. This gradually extended to the PME concept, which basically implies that consumers are provided with an earth terminal connected to the supply neutral conductor.

In the 1950s further problems arose when lead water pipes began to be replaced by plastic pipes. One solution was to earth consumers' plant to the lead sheath of the supply cable, but this was only satisfactory if across all straight and branch joints the lead sheaths were plumbed to the jointing sleeves. In many cases, however, the joints were of the mechanical grip type in cast-iron boxes, and these had such a high resistance that the earth path to the substation was inadequate. By additional earthing of the neutral conductor, nowadays usually only at the remote end of the run, and by using this conductor also as the protective earthing conductor, the consumer earthing problems were overcome. Moreover, the supply cable required one fewer conductor and by developing a new range of cables very considerable savings were obtained. *Figure 31.8* shows the much lower material utilisation achieved with one form of CNE cable.

Initially, PME systems with CNE cables were kept separate from the existing networks, but by appropriate and simple bonding between neutral conductors and lead sheaths, all existing systems can be converted to PME. The UK network has largely been modified in this way. CNE cables can, therefore, be installed indiscriminately for replacements and additions. A point requiring attention is that, if consumers are given a PME earthing facility, all exposed earthed metal within the installation which may be touched must be suitably bonded to provide an equipotential background.

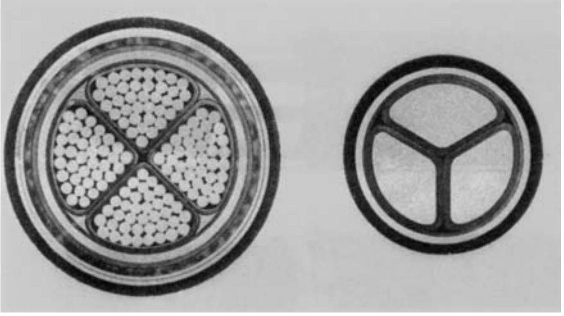


Figure 31.8 Comparative dimensions of four-core PILS/STA and Consac CNE cables of equal rating

31.4.2.2 CNE cable types

The Consac cable (BS 5593), first introduced in the mid-1960s and shown in *Figure 31.9*, maintains the use of paper insulation. Aluminium sheathing provides the neutral. The use of solid aluminium conductors with 1 kV paper insulation had already become established with four-core PILS cables before the development of Consac. Because of the total use of aluminium as conductor metal and the small amount of other material, Consac has a very economic construction, but in the early days some of the cost advantage was lost by the extra difficulty of plumbing the aluminium sheath. This was later overcome by the simple techniques involving the use of mechanical fittings and cast resin filling.

One of the important features in the design of any CNE cable is that, in the event of cable failure, there should be no loss of the important protective neutral conductor. It has also to be recognised that, with the growing use of mechanical excavating equipment, the main source of cable failures is now third-party damage. If the PVC oversheath on Consac is damaged, local corrosion of the aluminium sheath will follow and water entering the insulation will produce detectable cable failure before there is any severe reduction in conductance of the neutral.

The Waveform cable type (*Figure 31.10*) is also known as Waveconal and Alpex. Although introduced some years after Consac, and a little more expensive, it has become more extensively used than Consac, because of its simple jointing techniques. The XLPE insulation represented the first departure by the UK regional electricity companies from paper insulation for mains cables. In Waveform cable the neutral conductor comprises aluminium wires applied with a sinusoidal lay, and, as with all modern cable designs, there is an outer PVC oversheath. For making service joints, no cutting of the neutral is involved and the wires can readily be formed into two bunches for mechanical jointing. A key point in the design is that the wires are spaced and encapsulated between two layers of unvulcanised rubber so that each wire is separately embedded in the rubber. In the event of local damage to the oversheath, the entry of groundwater is thus limited; again an important factor in preventing loss of the neutral conductor. In Germany this form of sinusoidal lay neutral construction is

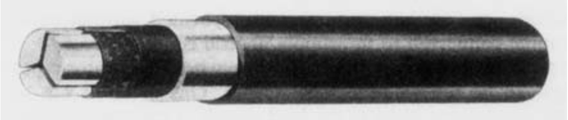


Figure 31.9 Consac cable

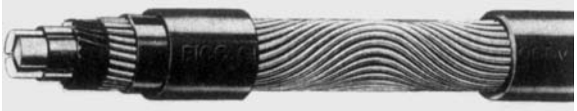


Figure 31.10 Waveconal cable

known as 'Ceander', but has only been employed with copper wires and without the rubber bedding. The use of copper wires as the neutral together with a single layer of unvulcanised rubber has recently found favour with the distribution companies in the UK, and many of those currently using Waveconal and Consac have now indicated their intention to adopt the cable having a copper wire neutral.

In Scotland some use has been made of 'Districable', a type which is also found in France. This is a four-core construction, again with XLPE insulation on the phase cores, but the neutral/earth circular or shaped conductor has a lead sheath to protect it from corrosion. Two thin galvanised steel tapes are applied as a binder and to provide a metal screen in contact with the neutral/earth conductor. A PVC sheath is applied overall.

Ultimate simplicity in CNE cable design and ease of jointing can be achieved by the use of four shaped aluminium conductors, insulated with XLPE and then provided with a PVC or polyethylene oversheath with no outer metallic protection. This type has already replaced all other public supply mains cable types in Germany. In the UK it has been rejected, however, even though a metallic envelope is not mandatory below 650 V and spiking tests have shown that danger from flash or shock is little greater than with the Waveconal, Consac, Districable or lead sheathed paper cables. The reason is that, in the event of mechanical damage, there could be exposure of the aluminium in the neutral to groundwater, with the possibility of undetected loss of this conductor. Damage to phase conductor insulation could also give rise to currents in the ground, flowing through other buried metalwork.

31.4.3 Service cable

Prior to the introduction of CNE mains cables, the service cables were of the three-conductor split-concentric type. This design is still used for consumers where it is not practicable to provide a PME earth terminal connected to the supply neutral/earth. The single-core (or multicore for three-phase supply) phase conductors are insulated with PVC or XLPE. In a helically applied concentric layer around the phase core or cores, some of the copper wires are bare to form the earth conductor and some have a thin layer of PVC coating to comprise the neutral conductor, the two portions being separated by PVC strings.

For a PME system, the construction is further simplified to a two-conductor design, the concentric layer consisting of bare copper wires.

As an alternative design of service cable, some users of Waveform mains cable prefer to adopt the same construction, i.e. sinusoidal lay aluminium wires embedded in rubber, as the neutral/earth conductor.

In addition to use for house service, all these cables find applications for such requirements as street lighting, traffic signs and complete individual routes for motorway lighting.

31.4.4 PVC insulated power cables

Although used for public supply cables in some overseas countries, PVC insulation has never been adopted in the UK for this purpose, other than for the service cables described above. The reason is associated with its thermoplastic nature and resultant softening at elevated temperatures. Thus, at 1 kV, ratings are restricted by a maximum temperature of 70°C, whereas paper can be operated to 80°C and XLPE to 90°C. More important, however, is that, in the event of a short overload, severe thinning may occur due to deformation by conductor thrust at bends, whereas paper or XLPE insulation would be relatively unaffected.

Close fusing to give cable protection is usually impracticable in public supply systems, but presents no great problem in industrial applications. From the late 1950s, therefore, PVC insulated cables were almost universally applied in this sector for voltages up to 3.3 kV. More recently, XLPE, with its superior overload characteristics has become increasingly more popular. Close fusing was defined by the 14th edition of IEE Wiring Regulations as an excess-current operating device which operates within 4 h at 30% excess rated value for cables direct in the ground, or 50% excess for cables in ducts or in air. In the 15th and recent 16th editions the whole concept has been changed (see Section 31.6.7).

Figure 31.11 shows typical 1 kV cable and 6 kV cable, the latter as used in coal mines.

As PVC insulated cables are little affected by moisture, no metal sheath is required, and this contributes greatly to ease of handling as well as simplifying jointing and terminating procedures. No precautions have to be taken to prevent entry of moisture.

BS 6346 caters for conductors of stranded copper or solid aluminium, but not stranded aluminium. The solid form was chosen because it provides the most economic cable construction and is particularly suitable for manufacture with PVC insulation. Solid conductors are also very much better for either soldering or mechanical jointing techniques. Stranded aluminium conductors are often preferred by overseas users and are also used for power supply cables in coal mines, as they facilitate coiling for taking the cables down the mine shafts.

Except for the smallest sizes, the conductors are shaped, and uniform thickness of insulation is obtained by extruding the PVC as a slightly oversize tube which is made a snug fit on the conductor by a combination of conductor feed speed control and internal vacuum. For multicore cables, the cores fit tightly together, leaving few gaps, but when these are of larger size, non-hygroscopic fillers are included so that the laid-up cores are reasonably circular.

For the armour bedding, there is a choice between PVC tapes and a layer of extruded PVC. The latter is more expensive but provides a robust cable which is preferred for cables with circular conductors, for cables laid

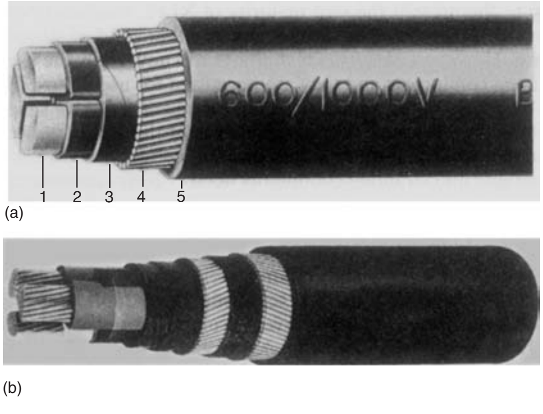


Figure 31.11 Pvc insulated cables: (a) three-core, 1 kV, SWA for industrial use; (b) British Coal three-core, 6.6 kV DWA

underground, and when it is desirable for terminating glands to provide a seal on to the bedding.

While any form of armour can be supplied (e.g. steel tape or strip, aluminium strip or galvanised wire (GSW)), BS 6346 covers only GSW or aluminium strip. GSW is normally preferred, as it gives optimum mechanical protection and adequate earth conductivity. Aluminium strip armour is now usual only when extra earth conductance is required and it is then important that suitably designed aluminium terminating glands be used. Aluminium armour is also necessary for single-core cables, because steel armour, being magnetic, increases losses, with an adverse effect on ratings.

A PVC sheath is usually applied overall and no bitumen is normally included over the armour. When this was done in the early years, following conventional practice with textile servings, it was found that the bitumen extracted plasticisers from the PVC, creating a mobile black liquid which would bleed from the cables at terminations below vertical runs.

The early choice of PVC by British Coal for mining cables operating at 3.3, 6.6 and, to some extent, at 11 kV was associated with the fact that the resilience of the insulation was found to provide better resistance than paper insulation to damage by rock falls. Although PVC is also satisfactory for higher voltages and has been used extensively in Germany at 20 kV, the electrical losses tend to be high. Better materials such as XLPE are now available when polymeric insulation is preferred.

Although the relative hardness of PVC at ambient temperature can be modified considerably by the choice and proportion of the plasticisers used, these cannot exert a significant effect on deformation at maximum operating temperatures. Heat resisting grades are defined in BS 6746, and such grades can even be formulated to allow PVC to operate for limited periods up to around 100°C without serious degradation due to chemical factors. However, they do little to improve deformation resistance and not much use has been made of them for power cables.

31.4.5 XLPE insulated cables up to 3.3 kV

Polyethylene has never found much application outside the USA for power cables, largely because PVC became established and polyethylene suffered from the same disadvantage of thermal deformation. XLPE completely overcomes this problem, and in the voltage range up to 3.3 kV it provides an advantageous alternative with cable constructions which are essentially identical. The main difference is that, as it is a much tougher material, the insulation thickness can be reduced, in the case of 1 kV cables to the minimum which can be extruded satisfactorily.

XLPE has now firmly established itself as an attractive alternative (both technically and economically) to PVC for industrial cables in the UK. XLPE has positive advantages because it is a better insulating material with much lower dielectric loss factor; more particularly, it can be operated satisfactorily to 90°C, with corresponding improvement in cable ratings. These factors have clearly provided an incentive for XLPE to be considered as a competitive material throughout the whole range of power cables up to the highest voltages and it is now being used in all spheres of application. There are competitors, such as ethylene propylene rubber, which may have advantages for specific cable types, but are unlikely to be economic over the whole range.

Up to 3.3 kV, therefore, XLPE is now superseding both paper and PVC insulation. In comparison with PVC, the continuous current rating advantage is usually more apparent than real, because cable size is dictated by voltage drop rather than current rating. The short-circuit rating based on 250°C instead of 160°C is likewise a bonus only infrequently required. Where XLPE does gain is in that, when ambient temperature is high, such as in tropical countries, the benefit from a smaller derating factor can be substantial. XLPE is not flame retardant, as is PVC, but as flame retardancy is governed more by the oversheath than the insulation, this is not normally significant.

In comparison with paper insulation, XLPE also has a small continuous rating benefit, but the main advantage is the absence of a metallic sheath and the availability of cable which is much cleaner and easier to handle in laying and jointing, together with lower permissible bending radii. The simpler jointing techniques, without any need for plumbing, provide strong attraction for developing countries where such skills are not readily available, and, to date, this is probably the area where XLPE has made the greatest impact.

Another field for XLPE is for self-supporting 240/415 V cables for overhead distribution, as a replacement for bare conductors. Four insulated circular stranded aluminium conductors are twisted together with a long lay and used with special fittings. Following widespread use in Europe, this application has now proved to be an economic alternative in the UK.

Some European countries have extended this application for higher voltages (up to 15 kV), but so far the only use in the UK has been in pilot schemes at 11 kV.

31.4.6 PE and XLPE cables for 11 kV to 45 kV

The excellent dielectric properties of PE and XLPE brought these materials into prominence in the early 1960s for higher voltage applications and an increasing scale of effort has been devoted to them ever since. In some countries, particularly the USA, they came into regular use at 10–20 kV at an early stage, and, in spite of a very poor initial service performance in comparison with paper insulation, they have since virtually replaced it for many years at voltages up to 45 kV.

The most important single factor which has caused problems is that, as with paper insulation, internal partial discharges occur at voltages of 5 kV upwards at any irregularities within or at the surface of the insulation. However, whereas paper insulation has fairly good resistance to such discharges and the effects in butt gap spaces can be minimised by oil or gas pressure, polyolefines such as PE and XLPE are particularly weak. Both PVC and EPR are rather better, but have other limitations.

While it was recognised that the insulation must be extremely clean and free from voids, and that screening at both surfaces of the insulation was necessary, many cables were put into service without adequate testing to ensure freedom from discharge. It was also not until the mid-1970s that ideal forms of screening were developed which could be readily removed for jointing and adequately deal with thermal expansion and contraction. Then, in the succeeding years, the final problem was to identify and find solutions to problems caused by effects of water in contact with the insulation. Water has minute solubility in PE and XLPE, but ‘tree-like’ structures were found in the insulation and it was eventually established that these could lead to electrical breakdown.

Although UK manufacturers supplied some of the first cables used in the early 1960s, the acceptance of XLPE at home has been slow to come. Three main reasons account for this. First, for the types of cable used, there was, until recently, no clear economic incentive in comparison with paper insulation. Second, in the countries where polymeric materials were quickly adopted, one of the prime reasons was to enable jointing to be carried out with less-skilled resources. In the UK this was not necessary. Third, the higher operating temperature of cross-linked insulation has particular benefit in reducing the derating penalty in countries having high ambient temperatures. Nevertheless, cable manufacturers export a large proportion of their production and so have been obliged to produce competitively.

In the problems that have arisen, polyethylene has no advantages over XLPE and, because it is a thermoplastic material, has great disadvantages in current ratings. Even in the USA, where the use of PE was substantial, it has now given way to XLPE. The remainder of this section refers to XLPE only.

For many years, IEC Specification 502 formed the basis for cable construction for UK manufacturers. However, the issue of BS 6622, which generally follows IEC 502 but is somewhat more demanding, covers the voltage range from 6.6 kV up to and including 33 kV. Above this voltage there is no international specification, but for cables above 30 kV and up to 150 kV, IEC Specification 840 gives detailed requirements for their test performance.

31.4.6.1 Conductors

Up to the present the vast majority of XLPE insulated cables have employed circular conductors in either solid or stranded form. However, extrusion techniques will now allow the use of shaped conductors and they are permitted for use up to 11 kV in BS 6622. At the moment there is limited demand for shaped conductors. Because of the importance of the screen between conductor and insulation, a smooth conductor surface is desirable and stranded conductors need to be well compacted.

31.4.6.2 Conductor screens

Many of the early cable failures were due to imperfections resulting from the use of semiconducting fabric tapes as conductor screens. A thin layer of extruded semiconducting polymeric material is now mandatory, and to ensure a clean interface it is normally extruded in tandem with the main insulation and cured with it. In the case of stranded conductors, a semiconducting tape may be applied between the conductor and extruded screen to prevent penetration between the wires and facilitate removal for jointing.

31.4.6.3 Insulation

Extrusion and curing can be carried out by a variety of processes, but a cardinal feature of all of them is that good material handling to avoid dirt and contamination is vital. The most common method of extrusion for cables up to 20 kV is the Monosil (or similar) process, whilst for 30 kv and above it is the continuous catenary vulcanising (CCV) method in which the curing is carried out by radiant heating and nitrogen under pressure (although it has limitations, curing is sometimes carried out by the application of steam pressure and cooling in water).

As explained earlier, the Monosil type process involves curing the cores in hot water; despite this the insulation performance and subsequent dielectric moisture content is equivalent to that of a ‘dry-cured cable’.

31.4.6.4 Insulation screen

One of the main factors concerning the dielectric screen is that it should be easily removed for jointing. A layer of semiconducting polymeric material, compatible with the insulation, can readily be extruded and cured in the same operation and techniques and materials are now readily available to enable the manufacturer to produce either a firmly bonded screen or a ‘strippable’ screen, which—although in intimate contact with the insulation—can readily be removed without recourse to special tools. Up to 33 kV the majority of cables use strippable screens, whilst bonded screens are occasionally used for 33 kV and almost always for higher voltages.

The older, taped, form of insulation comprising a layer of semiconducting varnish followed by an easily removed semiconducting tape is now used less often.

Both forms of semiconducting screen are usually followed by either a copper tape, applied helically, or a concentric layer of copper wires. The amount of metal in these screens must be related to what is required for earth fault current-carrying capacity; if tapes are used for three-core cables the metal tapes can be supplemented by copper wires in the filler spaces.

31.4.6.5 Finish

A typical three-core cable is shown in *Figure 31.12*, in which the copper taped cores are laid-up, then provided with a PVC extruded bedding, galvanised steel wire armour and PVC or PE oversheath. A single-core cable would have a

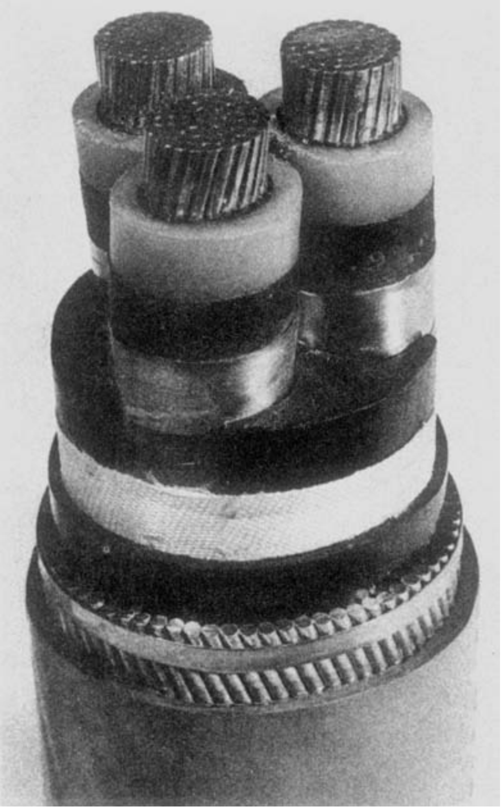


Figure 31.12 Three-core 8.7/15 kV XLPE insulated steel wire armoured cable

PVC or PE bedding over the concentric copper earth wires, then aluminium wire armour and PVC or PE oversheath.

Bearing in mind the faults that can be experienced due to contact between groundwater and insulation, and acknowledging the fact that plastic oversheaths may be damaged during installation or subsequently, new designs with components or special layers to restrict movement of water within cables are being produced. These will utilise some form of conductor blocking, usually in conjunction with blocking (swelling) tapes applied radially under the outer layers. Other designs use a powder which swells on contact with moisture.

31.4.6.6 Dielectric deterioration by treeing phenomena

No discussion on polymeric insulation at high voltage would be complete without some reference to deterioration caused by treeing mechanisms. These are related to a prebreakdown characteristic which gradually spreads through the dielectric under electrical stress through paths which, when visible or made visible, resemble the branch structure of trees. Trees are of two basic types:

(1) *Electrical trees* These are trees in a dielectric consisting of permanent channels having dendritic or branching patterns due to partial discharges during application of a.c., d.c. or impulse electrical stresses. The channels originate at sites of high stress due to non-uniform electrical fields from imperfections such as protrusions at an insulation interface, a void or a contaminant.

(2) *Electrochemical trees* This is a class of tree generated in a dielectric during application of electrical stress in the presence of liquid water or water vapour—hence, often known as ‘water trees’. They consist of fine water channels which can be seen under a microscope after staining. They disappear if the sample is dried, but reappear after boiling in water. Electrochemical trees are formed at stresses which are much lower than those required to produce electrical trees, and the rate of growth may be very slow. The tree patterns appear generally at opaque areas in the translucent polyethylene. If the dielectric or screen is in contact with soil water containing such minerals as sulphides, the water may have a characteristically coloured stain. The initiation of electrochemical trees is at the same types of site as indicated above for electrical trees. Characteristic names are often given to them according to origin, e.g. ‘bow-tie’ trees from contaminants (*Figure 31.13*(*a*)) and ‘bush’ or ‘broccoli’ from surface imperfections. *Figure 31.13*(*b*) shows an electrical tree which is developing in an area where electrochemical treeing has become extensive.

It is this treeing phenomenon which is the important reason for the insulation to be free from all irregularities and for the surfaces to be smooth and in good contact with the screens. Cables may operate for many years before a tree size is generated which will contribute to ultimate breakdown. The presence of water is a requisite for treeing to be initiated, but a very small amount suffices, and for the highest voltages it is desirable to ensure that all moisture is excluded, e.g. by provision of a metallic sheath.

31.4.7 Cable tests

Full details of the tests and procedures required are given in the cable standards listed at the beginning of the section. IEC 55, IEC 502 and IEC 840 are the most relevant documents and values quoted below are taken from these standards. A complete summary would be lengthy and it is

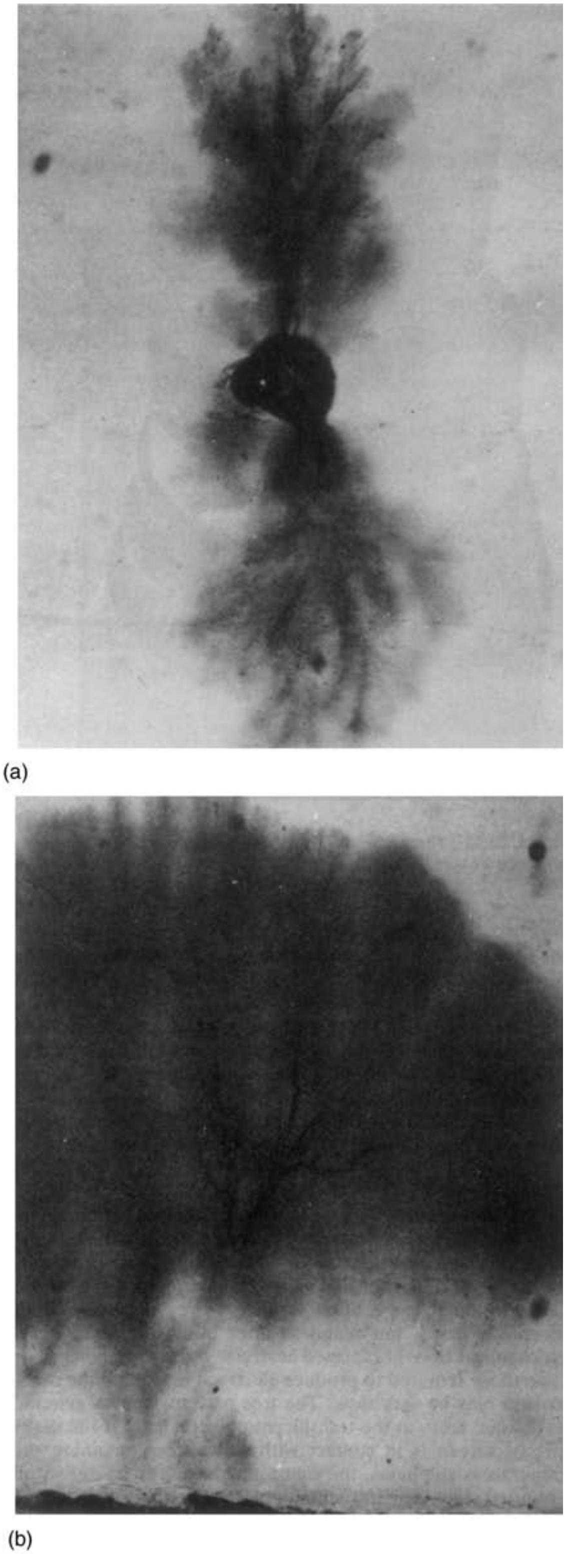


Figure 31.13 (a) Bow-tie tree at an inclusion; (b) Electrical tree in an area of extensive water tree development

only possible to give a brief outline to illustrate the general basis for the more important tests.

31.4.7.1 Manufacturing tests

Tests during manufacture are restricted to those which are not possible on finished cables and comprise a.c. spark tests on polymeric insulation and sheaths.

31.4.7.2 Tests of completed cables at works

Tests of completed cables at works comprise the following:

(1) Measurement of the thickness of insulation and other prescribed components.
(2) Conductor resistance test.
(3) An a.c. test for 5 min (30 min for cables above 30 kV) at a voltage which is usually 2.5 $U_0 + 2$ kV for cables rated up to 3.6/6 kV, and $2.5U_0$ for cables of 6/10 kV and above. For multicore non-screened cables the test is required between conductors and also between any conductor and sheath. For cables with individually screened cores it is from conductor to sheath only.
(4) For paper insulated cables with rated voltage U_0 of 8.7 kV and above, a dielectric power factor/voltage test is required to determine compliance with prescribed limits for maximum power factor at 0.5 times U_0 and maximum difference in power factor from 0.5 to 1.25 times U_0 and from 1.5 to 2.0 times U_0.
(5) A partial discharge test is required for cables insulated with PE and XLPE of rated voltages above 1.8/3 kV and on cables insulated with EPR and PVC of rated voltages above 3.6/6 kV. The magnitude of discharge at $1.5U_0$ must not exceed (a) 20 picocoulombs (pC) for EPR, PE and XLPE for cable up to 30 kV, (b) 10 pC for PE, EPR and XLPE above 30 kV and up to 150 kV, and (c) 40 pC for PVC.

31.4.7.3 Tests after installation

(a) Paper cables: a 15 min d.c. test at a voltage of 70% of the values given in (3) above.
(b) Polymeric cables: a 15 min d.c. test at a voltage of approximately $4U_0$ for cables up to 30 kV and $3U_0$ for cables above 30 kV up to 150 kV.

31.4.7.4 Special and type tests

(1) A bending test at a radius much more severe than stipulated for installation, followed by a voltage test. For paper cables the diameter of the test cylinder varies, according to the cable rated voltage and type, from 12 to 25 times the diameter of the cable plus the diameter of the conductor $(D + d)$. Three cycles of bending are required and maximum limits are stipulated for tearing of individual paper tapes. For polymeric cables up to 30 kV two cycles of bending are required over a test cylinder of $20(D + d)$ for single-core cables and $15(D + d)$ for multicore cables. For polymeric cables above 30 kV and up to 150 kV three cycles are required over a test cylinder of $25(d + D) + 5\%$ for cables with metal sheaths and $20(d + D) + 5\%$ for others.
(2) A drainage test for non-draining paper cables at the maximum continuous operating temperature for the cable. The maximum permissible drainage is 2.5–3% of the internal volume of the metal sheath.
(3) A dielectric security test for paper cables comprising sequential bending, impulse and a.c. tests. The impulse withstand requirement is 95 kV for $U_0 = 8.7$ kV, 125 kV for $U_0 = 12$ kV and 170 kV for $U_0 = 18$ kV. The a.c. application is $4U_0$ for oil/resin impregnation and $3U_0$ for non-draining impregnants.
(4) A power factor/temperature test for paper cables of $U_0 = 8.7$ kV and above to a temperature 10°C above rated temperature. Limits are 20–60°C, 0.0060; 70°C, 0.0130; 75°C, 0.160; 80°C, 0.0190; 85°C, 0.0230.
(5) An electrical test for PE and XLPE cables above 1.8/3 kV and PVC or EPR cables above 3.6/6 kV. This requires sequential application and/or measurement of partial discharge, bending, power factor/voltage, power factor/temperature, load cycles, partial discharge, impulse withstand and a.c. high voltage. For synthetic insulated cables above 30 kV and up to 150 kV the sequential application and/or measurement is bending, partial discharge, power factor/temperature, load cycles, partial discharge, impulse withstand and a.c. high voltage. The impulse requirement is $U_0 = 3.6$ kV, 60 kV; $U_0 = 6$ kV, 75 kV; $U_0 = 8.7$ kV, 95 kV; $U_0 = 12$ kV, 125 kV; $U_0 = 18$ kV, 170 kV; $U_0 = 26$ kV, 250 kV; $U_0 = 36$ kV, 325 kV; $U_0 = 64$ kV, 550 kV; $U_0 = 76$ kV, 650 kV; and $U_0 = 87$ kV, 750 kV. The a.c. test comprises 4 h at $3U_0$ for cables up to 30 kV and 15 min at $2.5U_0$ for cables above 30 kV up to and including 150 kV.
(6) Tests on the component materials before and after ageing and, in the case of polymeric cables, on the complete cables after ageing.

31.5 Transmission cables

31.5.1 Historical development sequences for a.c. transmission

31.5.1.1 Problems due to partial discharges within paper insulation

Reference has already been made to the work by Hochstädter which led to the 'H' type or 'screened' radial field design. Such constructions were quite satisfactory at 33 kV and to a limited extent at 66 kV. Failures at 66 kV and higher voltage were found to be due to discharges in minute vacuous voids formed by expansion of the impregnating compound with insufficient subsequent contraction to fill all the space available. Emanueli in the late 1920s pioneered the first solution, which was the oil filled cable. The basic requirement was either to eliminate completely the possibility of voids being created, as in the pressurised oil filled cable, or to ensure that they were always under a high gas pressure. Gas may be admitted directly into the insulation or exerted externally on a flexible sheath over the insulation, in which case the void suppression principle is more akin to that of the OF cable.

As the void formation mechanism was also clearly related to the temperature excursions of the insulation, the operating temperature limit of the solid (i.e. non-pressure) insulation could be raised from 65 to 85°C for gas filled cables and 90°C for oil filled cables, with consequently much improved cable ratings. Even more important was the fact that the a.c. operating electrical stress of impregnated paper insulation could be increased from 4 MV/m to about 16 MV/m and reductions in dielectric power factor were achieved.

31.5.1.2 Types of paper-insulated pressurised cables

Many different types of pressuring are possible and may be classified into the two basic constructions indicated in *Table 31.6* which lists those currently in service.

Table 31.6 Pressure cables and voltage in commercial service

Design	*Voltage range* (kV)
Fully oil impregnated	
Lead or aluminium sheath	
Low-pressure OF	30–525
Steel pipe	
High-pressure OF	30–500
External gas pressure with diaphragm sheath	30–275
Internal gas pressure	
Lead or aluminium sheath	30–275
Steel pipe	30–132

OF, oil filled.

A fact which emerges from *Table 31.6* is that oil pressure can be used up to the highest voltages at present required (525 kV). Gas pressure has a limitation primarily related to a lower electrical breakdown strength. Although gas pressure cables have some advantages in terms of the associated accessories and equipment, the oil pressure cables are usually more economic.

31.5.1.3 Transmission system requirements

When Emanueli first developed the pressure cable technique, he was fulfilling a need for the requirement of the early 1930s to transmit in the range of 30–132 kV with conductor sizes of around 200 mm^2. With the growth in the usage of electricity during the next 30 years, cable voltages and ratings had to keep pace with and match the overhead-line circuits. Developments quickly proceeded to find solutions for the higher voltages with conductors up to 2500 mm^2.

Bulk transmission in the UK began in the 1930s with a circuit requirement of 110 MVA at 132 kV; 275 kV followed in the late 1950s with a winter circuit rating of 760 MVA; and by the late 1960s the circuit demand had increased to 2600 MVA at 400 kV. To obtain this from a single cable circuit meant that design had to be pushed towards the limit for paper insulation in relation to electrical features, diameter and coiling on drums. To match the increases in overhead-line ratings, it has become necessary for the heat generated in cables to be removed by more sophisticated engineering means involving cooling pipes.

31.5.1.4 Alternatives to impregnated paper insulation

The oil filled cable has been most successful in meeting all requirements up to 525 kV. At some future date (possibly not until the next century in the UK, but earlier in some other countries) there will be a need for undergrounding parts of transmission lines operating at 800–1000 kV. To produce cables within a diameter (say 160 mm) which is practicable for manufacture and handling, and to keep losses within an economic limit, it seems essential to use higher stresses (possibly 25 MV/m) and dielectrics with lower power factor and permittivity than can be achieved at present with impregnated paper. Possibilities are discussed later.

There has been much recent interest in materials such as polyethylene and XLPE, which have good potentialities for very high voltages and other advantages for the lower voltage range. To date, the enthusiasm has stemmed not primarily from low losses, but from the possibility of a much simpler cable construction, the promise of less complicated jointing requirements and, above all, the elimination of pipework and pressurising equipment. In some overseas countries, where installation and maintenance skills are not readily available, this is an important factor which could well justify a higher intrinsic cable cost. Location and repair of oil leaks can be troublesome. However, some of the problems with polymeric insulation for distribution cables have been discussed, and for voltages greater than 132 kV the usage so far of PE and XLPE cables has been relatively small.

31.5.2 Types of cable

31.5.2.1 Basic requirements

Apart from absolute consistency and freedom from defects, the essential requirements of high-voltage dielectrics are:

(1) High impulse strength, because this is the ultimate design stress requirement and determines dimensions.
(2) Low dielectric power factor in order to keep the heat generation to a minimum. When conductor size is at the maximum possible, much expense may have to be devoted to means of cooling the cable to obtain an economic circuit.
(3) Low permittivity to reduce both the electrical losses and charging current.
(4) Ease of bending during installation without sustaining damage which could affect service life.

Impregnated paper under oil pressure is the only dielectric which so far has met all these requirements up to about 525 kV. In relation to electrical losses, however, it is reaching its limit at this voltage without forced cooling. Impregnated paper with gas pressure is technically satisfactory up to 275 kV but is not generally economically competitive with oil filled cables.

Low-pressure oil filled cable (*Figure 31.14*) is used almost universally in the UK throughout the voltage range and it is predominant throughout the world. High-pressure OF cable is favoured in the USA.

The influence of the impulse strength requirement on design stress can be seen from *Table 31.7*.

Conventional oil filled cable has a safe impulse stress of around 100 MV/m and a.c. stress of 30 MV/m, a ratio of about 3/1. This has to be compared with the service performance requirement of between 10.2/1 and 6.2/1 according to voltage, i.e. cables must be designed on an impulse breakdown stress basis and then they will have a large safety margin for a.c. performance. The reverse would soon lead to breakdown. *Table 31.7* also illustrates that, because the impulse/a.c. ratio reduces with increasing voltage, higher design stresses can be adopted as voltage increases: typical

Table 31.7 Working and impulse voltages

System voltage (kV)	*Working voltage* (kV)	*Impulse-test voltage* (kV)	*Impulse/working ratio*
33	19	194	10.2
66	38	342	9.1
132	76	640	8.4
275	160	1050	6.6
400	230	1425	6.2

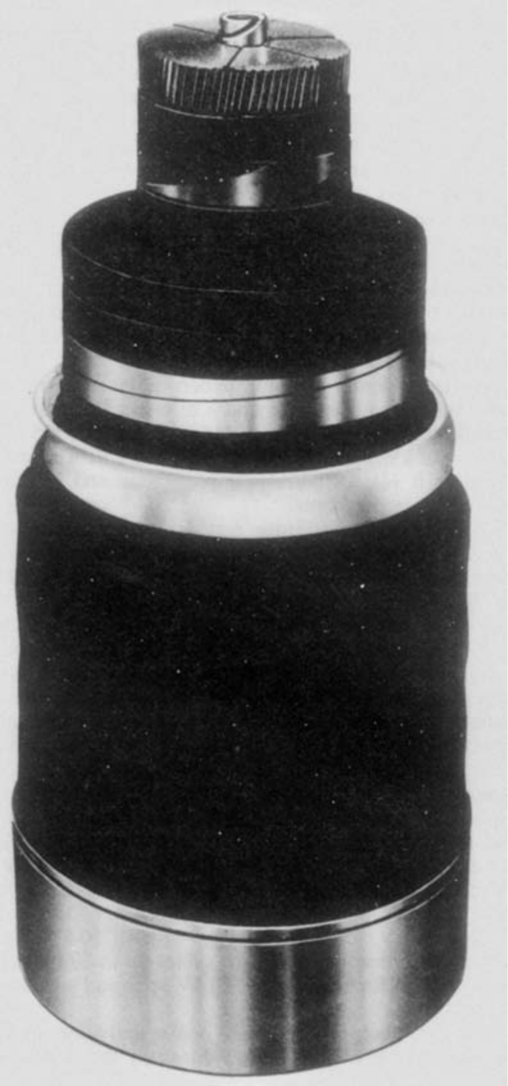


Figure 31.14 A 400 kV single-core oil filled cable

values are 7.5 MV/m at 33 kV; 12 MV/m at 132 kV; and 15 MV/m at 275–400 kV.

31.5.2.2 Low-pressure oil filled

Right from the beginning the low-pressure oil filled cable has been well to the fore and has been the only type of cable widely used in the UK at 275 and 400 kV. Single- and three-core designs are available from 33 to 132 kV but, because of diameter limitations, only single-core cables can be produced for the higher voltages. *Figure 31.15* illustrates how oil channels are provided within the cable. In single-core cables the oil flow is normally through a duct in the centre of the conductor, but for short lengths used to terminate three-core cables the design may incorporate an annulus formed by the provision of longitudinal ribs on the inside of the lead sheath. In triple-core lead sheathed cables and aluminium sheathed cables with circular conductors, a duct is placed in the fillers between the cores. Alternatively, with a corrugated aluminium sheath (CSA) it is possible to omit the ducts and fillers. At 33 kV the conductors may be of oval shape and the construction is known as ductless shaped oil filled, whereas for higher voltages with circular conductors it becomes ductless circular oil.

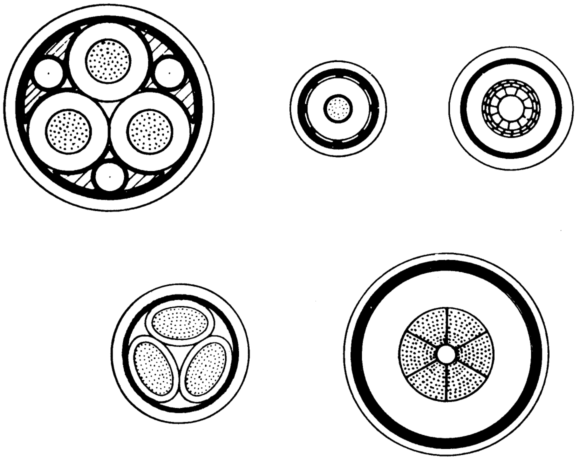


Figure 31.15 Cross-sections of typical oil filled cables

As the cable heats, the oil expands and is forced out of the cable through pipes at joints or terminations into a tank reservoir having internal pressurised capsules so designed that, on cooling, there is a feedback of oil into the cable. *Figure 31.16* illustrates the system. Tanks are of sizes to suit the route length and volume of oil in the cable. They are pressurised to take into account variations in height along the cable route. By the inclusion of stop joints between lengths of cable the circuit may be split into several oil sections. The designed static pressure within the cable is 5.25 bar, but transient pressures up to 8 bar can occur during periods of rapid heating due to increasing load. Optimum planning of the oil feed and sectionalising arrangements is a very important part of the economic design of a cable system.

From the time the cable is filled with oil during manufacture, the oil pressure must be continually maintained. A small tank is fitted on the cable drum; it remains connected during cable laying and even during jointing a flow of oil is maintained. When cables are installed in vertical shafts, e.g. for pumped storage stations, special arrangements are necessary. The Cruachan 275 kV pumped storage scheme in Scotland has a vertical head of 325 m with a consequent hydrostatic pressure of approximately 30 bar. The cable had to be partially drained under vacuum to limit the oil flow while making the lower stop joint, and reimpregnated before making the upper sealing end.

Lead sheaths will not withstand significant internal pressures, and are reinforced to withstand a continuous pressure of 5.25 bar for normal installations by the helical application of bronze tapes. In spite of the reinforcement, the lead sheath is subject to some expansion under creep stress. British practice favours the use of $\frac{1}{2}$C alloy (0.2% tin, 0.075% cadmium). Aluminium sheaths have technical and economic advantages and nowadays they are of corrugated design. This enables the thickness to be reduced and also provides greater flexibility for handling.

To maximise the efficiency of oil flow and the length of individual oil sections, the viscosity of the oil needs to be as

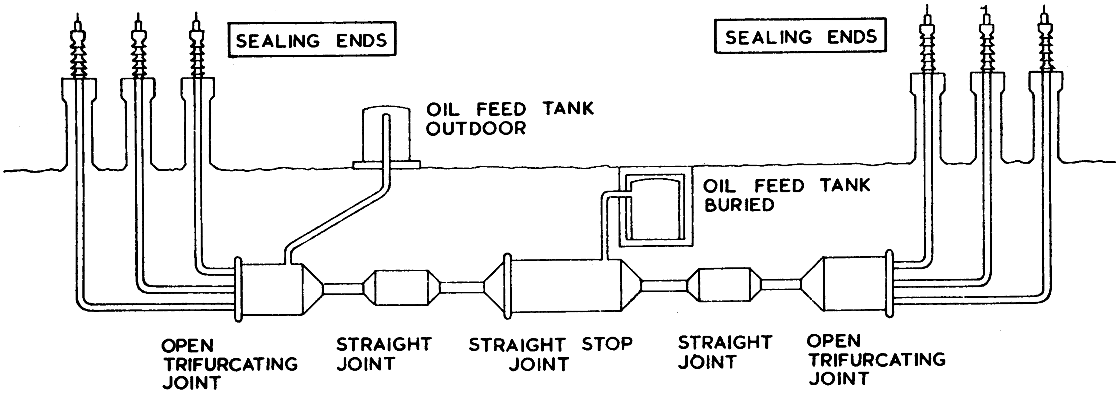


Figure 31.16 Diagrammatic layout of a typical three-core OF cable system

low as possible, consistent with low power factor and electrical strength. Until recently, mineral oils were used with a viscosity of about 12 centistokes maximum at 20°C. However, current practice is to use synthetic alkylates of dodecylbenzene type which have better gas absorbing properties under electric stress. Such impregnants are intermediates in detergent manufacture.

Apart from the impregnant, the insulation for oil filled cables also differs significantly from that for lower-voltage paper cables. To keep the power factor as low as possible, the paper needs to be more thoroughly treated to remove impurities. For example, the water used in papermaking and washing for very-high-voltage cables may be deionised. For electrical stress reasons, and to obtain good bending performance without disturbance of the dielectric by wrinkling, etc., the papers tapes are graded from thin adjacent to the conductor (where the electrical stress is highest) towards much thicker and wider on the outside (to withstand the higher mechanical stress). With the large insulation thickness (up to 30 mm) required for the highest-voltage cables it is necessary to control the design and paper lapping parameters to allow individual paper layers to slide over one another on bending. This is done by shrinking the papers by predrying before lapping, and carrying out the lapping in a low-humidity atmosphere with careful control of tension.

31.5.2.3 High-pressure oil filled cable

The high-pressure oil filled cable is a type of cable (*Figure 31.17*) developed in the USA and used extensively only in a few countries. It evolved from the predominant American practice of installing cables in buried ducts, the steel pipe being essentially a duct which can be installed a short length at a time without need for long trenches to be kept open. Either the cable cores have a temporary lead sheath which is stripped off as the cable is pulled into the pipe, or the unsheathed cable is delivered to site on a specially sealed and protected drum. So that the cores are not damaged during the pulling operation, D-shaped skid wires are applied helically over the insulation. After jointing, the pipe is evacuated and filled with oil to a pressure of 14 bar and the pressure is maintained by automatic pumping stations. The relatively large volume of oil and the high pressure enable a viscous impregnant to be used.

Pipe type high-pressure oil filled cables tend to be more expensive than self-contained oil filled cables laid directly in the ground, but in built-up inner city areas, or where robustness is desirable, they can be advantageous. Except for terminations, they always consist of three single-core cables pulled into a single pipe. The proximity of the cores and the high electrical losses in the pipe impose lower ratings than for self-contained cables.

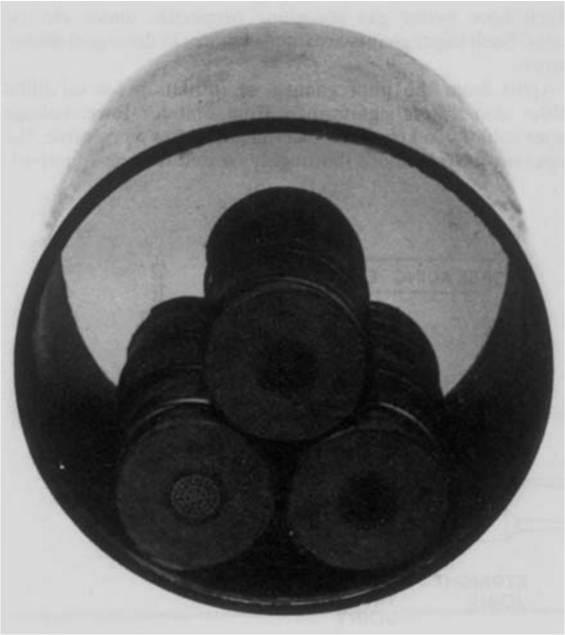


Figure 31.17 Pipe type 230 kV high-pressure OF cable

31.5.2.4 Gas pressure cables

During the 1930s–1940s, many designs became established to utilise the principle of gas pressure to suppress partial discharge in voids.

In the ‘internal gas pressure’ cable and in one form of pipe type cable the gas was admitted directly into the cable insulation and held by the metal sheath or steel pipe. The ‘gas compression’ cable worked on a different principle. Insulated oval conductors were sheathed with either lead or polyethylene and the individual cores or the three laid-up cores were then covered with a pressure retaining metallic sheath or pulled into a steel pipe. The space between the inner sheath and the outer pressure retaining member was filled with high-pressure gas, usually nitrogen. Expansion and contraction of the relatively viscous impregnating compound was compensated for by the inner sheath acting as a diaphragm.

With one exception, gas pressure designs are obsolete mainly because they cannot match the technical performance of oil filled cables through the voltage range. At 275 kV and above, low power factor and high breakdown strength (a.c. and impulse) become increasingly important, and oil filled cables can be operated to higher design stresses.

The exception is the 'pre-impregnated gas filled cable', useful for applications where problems exist in creating practicable oil sections, e.g. on hilly and undersea routes. Other advantages are: (a) there is no need for specialised oil equipment and (b) long continuous lengths suitable for submarine use can be manufactured. The total length for installation is determined by what can be coiled down in a ship, as joints between lengths can be made either in the factory or on the ship. In this gas filled cable system the paper is impregnated with a special greasy compound before being lapped on to the conductor. With modern designs the impregnated cores are covered by a smooth aluminium sheath and gas is admitted directly into the insulation after installation. There is a minimum of impregnating compound, and although voids do exist from the outset, the high nitrogen pressure provides good electrical strength. As with other forms of gas cable, however, operating voltages are usually limited to 132 kV.

31.5.2.5 Cables with polymeric insulation

Mention has already been made of the increasing use of polymeric insulation, largely XLPE, for distribution cables up to 33 kV. The low power factor of around 0.0005 which is attainable with such cables is also clearly attractive in comparison with the minimum of about 0.002 which is possible with the best oil filled cables. It seems likely that a new phase of cable transmission is emerging. Incentives are simpler jointing techniques, and reduced maintenance as pressurising equipment and oil leaks are avoided. An economic consideration relates to the voltage limit above which it is desirable to have a metal sheath over the insulation to prevent contact with water.

It was reported in 1988 that 230 km of lead sheathed 225 kV cable with thermoplastic polyethylene insulation was in satisfactory service in France, the first lengths having been installed in 1969. Electricité de France also claimed that the overall economics were favourable. This is despite the use of a lead sheath and limitation with straight polyethylene of the operating temperature to 70°C compared with 90°C with paper insulation.

Following consolidation of satisfactory experience at 33 kV, it is at around 132 kV that large-scale experience will first be obtained and it is now generally accepted that for voltages of 66 kV and above, a metallic sheath is required. XLPE is now well established as a dielectric at 132 kV and long length installations at 275 kV are just beginning. Many small installations at 132 kV are in service with a design stress of 7–9 MN/m. The problems have been enumerated in Section 31.4.6 and are primarily concerned with the production of clean insulation and screen interfaces plus possible incidence of water treeing in service. The French experience with straight polyethylene indicates that the former can be overcome, and only time will prove whether a metal sheath is necessary to prevent degradation by water tree mechanisms.

Much development work is proceeding on the optimum method of curing XLPE. Although dry-cured material has a lower content of microvoids than steam-cured material, there is little difference between the two levels in water tree growth if water is in contact with the insulation. If kept dry, treeing ceases to be a problem and the improved short-term breakdown strength of the dry-cured material enables it to be operated at a higher stress.

31.5.3 Submarine power cables

The engineering of submarine cable links is complex. Apart from the choice between a.c. and d.c. transmission, cable design has to take account of the maximum depth on the projected route, the potential hazards caused by shipping, corrosion and possibly marine borers. There are many submarine power cable installations giving satisfactory service at voltages up to ±400 kV d.c. and 420 kV a.c.

The choice of cable for a submarine crossing is influenced by the system voltage, the maximum depth and the length of the crossing. Solid type paper insulated cables are suitable for voltages up to 400 kV d.c. and 33 kV a.c. For deep-water installations special design features are necessary to enable this type of cable to resist the external water pressure. Solid type cable has been used successfully at 550 m depth on a d.c. link between Norway and Denmark. For higher voltages self-contained pressure assisted cables (oil filled and gas filled) are used, the internal pressure being maintained above the external water pressure at the deepest part of the route. Although not so far used for major submarine transmission schemes, polymeric insulated cables may be attractive for future a.c. links.

It is preferable that the cable be manufactured in continuous lengths without joints. As this is not always possible, proven techniques have been developed for the construction of flexible joints in all types of power cable, to facilitate laying in a continuous operation. Manufacture has to be arranged so that the cable can be coiled down on land and then reloaded directly into the cable laying vessel.

Experience on the ±100 kV d.c. cable circuit between England and France (1961) and on the Sweden–Denmark ('Konti-Skan') ±250 kV d.c. link (1965), indicates that cables laid directly on the sea-bed across busy shipping lanes or fishing zones are liable to suffer frequent impact damage caused by dragging anchors and trawls. A significant increase in circuit security can then be obtained by embedding the cables. Techniques have been developed for cutting trenches in the sea-bed and for accurate positioning of the cables within the trenches. On many cable routes adequate security can be obtained by burying the cable at the shore approaches only.

31.5.4 D.c. transmission

Table 31.8 indicates the main d.c. schemes in operation, all of them being submarine. There are advantages in cable cost, but expensive terminal conversion stations may make such schemes uneconomic unless there are other overriding considerations. These arise when large national systems need to be interconnected and occasionally when large blocks of power have to be transmitted within a network. Most of the existing schemes are for submarine links where the charging currents for a.c. cable systems would be excessive.

D.c. cables can be operated at much higher design stresses than a.c. cables. For example, a typical 250 kV oil filled cable could have a maximum design stress of 33 MV/m, whereas a comparable 275 kV a.c. cable would have

Table 31.8 D.c. cable schemes

Link	*Voltage* (kV)	*Approximate installation date*	*Type of cable*	*Approximate route length* (km)
Gotland (Sweden)	100	1954	Solid	100
UK–France (Cross Channel Link)	±100	1965	Solid	52
Sardinia–Corsica (Italy)	±200	1965	Solid	104
Cook Strait (New Zealand)	±250	1965	Gas filled	40
Mainland to Vancouver Island	±300	1969	Solid	26
Mainland to Vancouver Island	±300	1976	Oil filled	36
Skagerrak (Norway–Denmark)	±250	1977	Solid	127
Tsugaru Strait (Japan)	±250	1978	Oil filled	45
Sweden to Gotland	150	1983	Solid	90
UK–France (Cross Channel Link)	±270	1985	Solid	45
Sweden to Finland	±400	In progress (1991)	Solid	200
Hawaii to Maui Alenuihaha Channel	±300	Planned (1991)	Oil filled	61

a design stress of 15 MV/m. Although the partial discharge in voids mechanism does not apply in d.c. operation, there are other factors, such as stress distribution and transient voltages arising from rectifier malfunction, which have to be taken into account.

In a.c. cables the stress distribution in the insulation is determined by the geometry and the permittivity of the dielectric. It is usual to assume a uniform permittivity, as this property is affected only to a minor degree by changes in cable temperature and voltage. In d.c. cables, however, the steady state stress distribution is dependent on the geometry and resistivity of the dielectric. If the latter remains uniform, the stress distribution is the same as that for a.c. However, the resistivity of the dielectric is highly dependent on the dielectric temperature and to a lesser degree on the applied stress. When the cable is carrying load, there is a temperature gradient across the dielectric, the effect of which is to reduce the stress adjacent to the conductor and to increase it at the outside. It is possible to arrive at the conditions where the stress at the outside exceeds that at the conductor and the insulation must be designed to cater for these changing stress conditions. *Figure 31.18* illustrates the principles involved.

Nevertheless, as pressurising of the insulation is not so necessary and does not give much advantage in direct voltage and impulse breakdown strength, it is possible to use mass impregnated solid type cables for much higher equivalent direct voltages. This assumes that the dielectric is not weakened by migration of impregnating compound, i.e. the insulation is 'non-draining'. For the highest voltage levels, oil filled cables are used and, as with a.c. but for a different reason, clean paper obtained by the use of deionised water is desirable. Excessive conductivity due to ionic impurities can lead to thermal instability and breakdown.

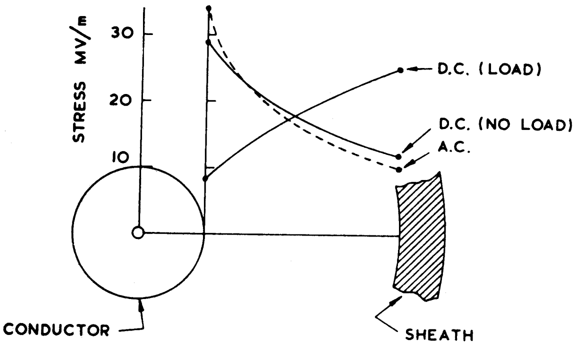


Figure 31.18 Stress distribution in d.c. and a.c. cables

A power of 500 MW can be transmitted by *three* single-core a.c. cables with 1000 mm^2 conductors at 275 kV. The same power can be conveyed by *two* single-core 800 mm^2 d.c. cables at ±250 kV. To transmit 1500 MW would require a double circuit comprising *six* naturally cooled 2000 mm^2 a.c. cables, but still only *two* d.c. cables, reducing by two-thirds the number of substantially identical cables required.

31.5.5 Cable ratings and forced cooling

The considerations in the general section on current carrying capacity are applicable also to transmission cables, but because of the much greater power carried, the effects of heat dissipation in the ground are of particular importance. First, it is necessary to inspect the soil to determine its thermal resistivity: 1.2°C m/W is taken as a representative value, but it may be much higher in sand, shingle or made-up ground, or if the soil is likely to be permanently dry. The moisture content is a significant factor in ground thermal resistivity; this became apparent when cables were loaded continuously so that moisture could not seep back during reduced load periods. If the ground surrounding the cable reaches a temperature of around 50°C, there is a considerable danger, with certain types of soil, of reaching a 'runaway' condition: complete drying out, high thermal resistivity, excessive temperature rise in the cable and breakdown.

When there is doubt about the thermal properties of the backfill, it is safer to surround the cable with imported material having known thermal resistivity in the dry condition. This means creating a dense mass with little air space by a controlled mixture of sand and gravel, particle sizes being blended to obtain good packing. Laboratory control of composition and compaction is important. An alternative is to use a mixture of selected sand and cement in the proportion of 14:1.

During the early 1960s, following the first failure due to ground drying out, several installations were completed in the UK with cooling pipes laid adjacent to each cable, water being circulated through a closed heat exchanger. The latter was air cooled or in some locations water cooled by supplies from bore-holes. Initially, aluminium was used for the pipe but this was later changed to high-density polyethylene.

In these later installations four pipes were used with an internal bore of approximately 66 mm. Specially selected sand was used around the cables and pipes. Some 160 circuit km of 275 kV with a winter rating of 760 MVA was installed in this way. It was later found that this rating, together with 1100 MVA at 400 kV, could be achieved by naturally cooled cables with a stabilised backfill and more realistic assumption of ambient ground parameters.

Separate pipe cooling came back into prominence in 1977 when overhead line ratings were further increased by raising the operating temperature to 65°C. This required a winter rating of 2038 A, equivalent to 970 MVA at 275 kV and 1410 MVA at 400 kV, which could not be achieved with the 2500 mm^2 maximum conductor size and stabilised backfill. A similar need for additional cooling also arose with the 400 kV cables in the Dinorwic pumped storage scheme in North Wales. Improved water pipe systems were adopted, the emphasis being on the use of larger pipe diameters and special arrangements for water cooling of the joints (*Figure 31.19*).

A rating factor which is particularly important with high-voltage single-core cables is to prevent the very high losses in metallic sheaths and reinforcement if these were solidly bonded at both ends of a feeder. The losses accrue from currents induced in the low-impedance sheath circuit and are related to the conductor current and separation between phases. Without elimination of such losses the use of aluminium sheaths would not generally be economic. In some cases it may be possible to bond earth at one end only, but modern practice is to employ a transposition method (*Figure 31.20*) in which the metallic sheaths are interrupted every few hundred metres with cross-connection at jointing positions. Voltages are then balanced at every third joint and usually kept below 65 V under full-load conditions. High transient voltages can occur; and to check that the sheath insulation is satisfactory, a 10 kV d.c. test is carried out after laying.

When two circuits are laid on a common route, the current rating will be reduced by mutual heating unless thermal independence can be obtained by a separation of about 2 m at 132 kV, with progressive increases for higher voltages. To obtain the most economic solution, it is necessary to examine the cost of larger conductor cable, extra trenching and external cooling. In the case of a single-core circuit the mutual heating effects between the phase cables must also be taken into account. With single-point bonding and cross-bonding to eliminate sheath circulating currents, a flat formation with spacing between cables of 150–300 mm is usually beneficial to avoid unduly high sheath voltages. If single-core cables are bonded and earthed at both ends, it is necessary to install the cables in trefoil formation, because with wide spacing the increase in sheath losses would more than offset the reduction in mutual heating.

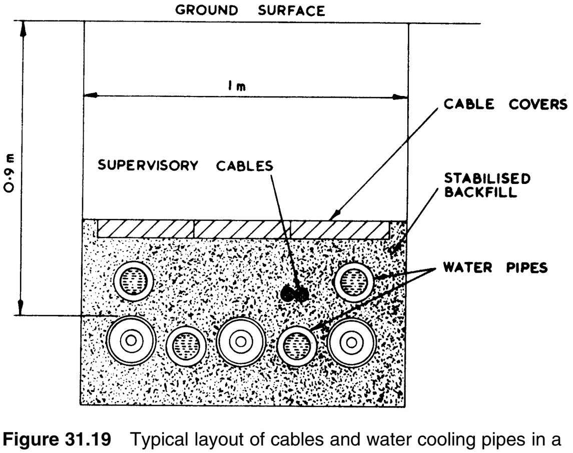


Figure 31.19 Typical layout of cables and water cooling pipes in a trench

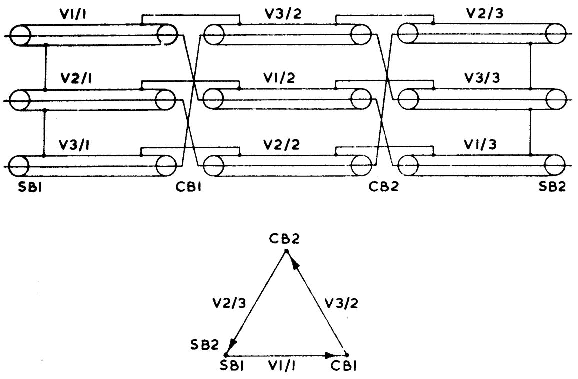


Figure 31.20 Cross-bonding of cable sheaths to provide transposition for reducing sheath losses

31.5.6 Future development

It was mentioned earlier that the low-pressure OF cable was nearing its limit of performance at the present maximum service voltage of 525 kV. However, overhead lines will soon be in operation at around 1000 kV, and at some future date there will be a need for cables to match them.

A major factor in the development of such cables is the need to keep the diameter of the cables down to a size that will enable drums of completed cable to be transported on existing road systems. As 525 kV cables are already approaching this limit, only a small increase in insulation thickness can be permitted. This will result in the insulation of the higher voltage cables operating at a much higher electrical stress. Experimental work has shown that this should be feasible provided the minimum oil pressure is substantially increased.

While higher design stresses are possible with impregnated paper insulation, they create problems insofar as dielectric losses are concerned. The dielectric loss of a cable which occurs whenever the cable is energised can be expressed as follows:

$$\text{Dielectric loss} = 27.7\,Vfds\varepsilon\delta\varsigma \times 10^{-4}\,(\text{W/m})\Leftarrow$$

where V is the phase voltage (kV), f is the frequency (Hz), d is the conductor screen diameter (mm), s is the maximum design stress (kV/mm), $\varepsilon\varsigma$ is the relative permittivity of the insulation, and $\delta\varsigma$ is the dielectric loss angle.

For a given conductor size, frequency of supply and insulation characteristics, the dielectric losses are proportional to the product of the operating voltage and design stress. The increase in the dielectric loss means that the current dependent losses, i.e. the current rating, must be reduced to prevent the maximum design temperature being exceeded. At a voltage of approximately 850 kV, the cable reaches its maximum design temperature purely by voltage energisation and, therefore, has no current rating.

To overcome these problems, designers have been looking for alternative materials which have lower dielectric losses than paper but which have the required mechanical and physical characteristics. Although much research work has been devoted to insulation consisting of all-plastic films, the material now being introduced into commercial service is a laminate consisting of a film of polypropylene between two layers of paper (PPL). The proportion of paper to polypropylene is approximately 50:50. It has been found that this type of laminate overcomes many of the problems associated with all-plastic films.

The use of PPL reduces the dielectric loss to about 25% that of conventional paper insulation and extends the useful operating voltage range above 1000 kV. It is also beginning to find use in low- and high-pressure oil filled cables with voltage ratings down to 345 kV.

Polymeric insulation such as XLPE also has very low dielectric losses and, therefore, is attractive for the higher operating voltages. Until recently these insulations operated at much lower electrical stresses than paper insulation and therefore were not suitable for use in cables at voltages in excess of 275 kV. However, much development work has been undertaken with the object of increasing the operating stress. This has been mainly directed at reducing the contaminant level in the insulation and providing extremely smooth conductor and dielectric screens. Short lengths of 500 kV cable are now in service to obtain operating experience.

A completely different approach is to make use of superconduction: very large currents can be carried by small conductors without the generation of much heat. Until recently, it was necessary to use liquid helium to cool the conductor to the required temperature.

The engineering problems are not inconsiderable, but experimental cables have been made and trials undertaken to demonstrate the practicability of such schemes. It has, nevertheless, been established that the economics are such that this form of transmission can only be justified for ratings of the order of 5–10 GVA. In the UK such requirements are several times what can immediately be envisaged.

The subject has attracted renewed interest with the recent discovery of high temperature superconduction. At the time of writing discoveries are still taking place and, therefore, it is not possible to assess its full impact on underground transmission. It would appear that with these new superconductors, liquid nitrogen is a possible coolant which is considerably cheaper than liquid helium. However, the materials so far discovered are of a brittle nature and will require special designs of conductors.

31.6 Current-carrying capacity

The continuous current rating of a cable is dependent on the way heat generated in the conductor, insulation and metallic components is transmitted through the cable and then dissipated from its external surface. For convenience the conductor temperature is taken as the reference datum for the cable. A notional maximum cable rating can then be calculated from the permissible temperature rise from a standard base ambient or ground temperature to the maximum temperature that the particular type of insulation will withstand with a reasonable margin of safety. Adjustments to this notional rating have to be applied to cover many factors, which include a different base temperature and variations in heat dissipation from the cable surface: e.g. dissipation from a cable in a duct is lower than from a cable in free air.

The difference between conductor temperature and external or ambient temperature is directly related to the total heat losses and the law of heat flow, using a conduction current analogue. This analogy may be extended into the type of circuit diagram in *Figure 31.21*, which shows how the heat input at several positions has to flow through

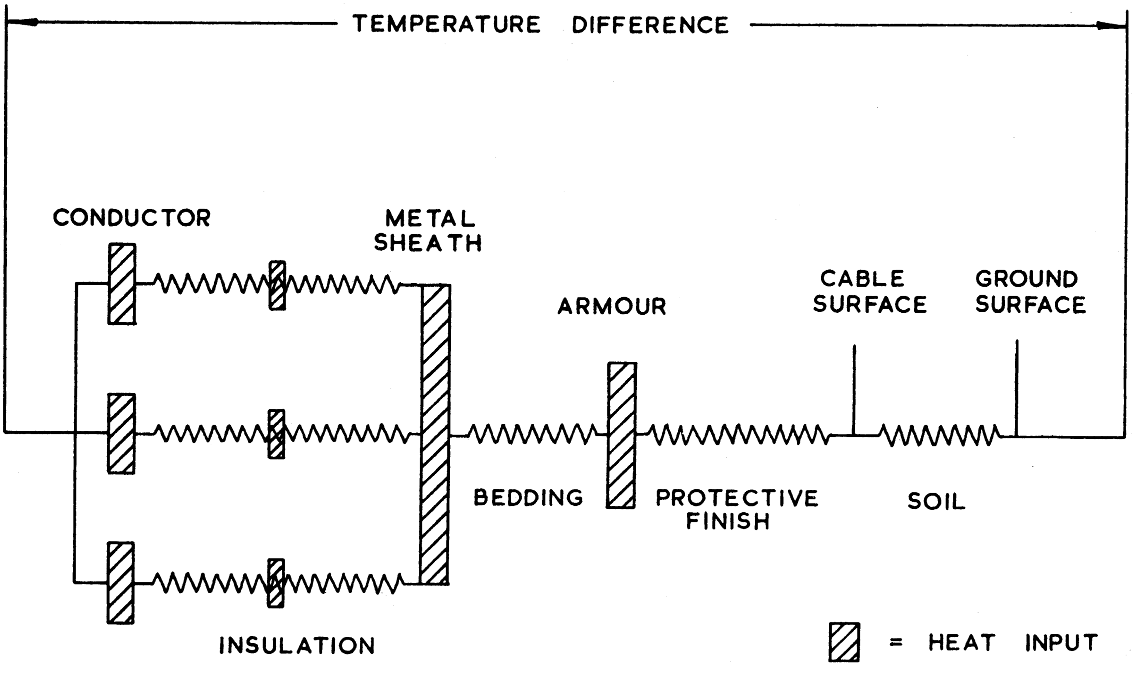


Figure 31.21 Equivalent circuit diagram for the heat flow in a three-phase belted cable

various layers of different thermal resistance. To make calculations, values of thermal resistivity have to be measured for all the materials involved. Thermal resistivity is defined as the difference in degrees Kelvin between opposite faces of a metre cube caused by the transference of 1 J/s of heat; the SI unit is kelvin-metres per watt (K-m/W).

31.6.1 Availability of continuous ratings

Cable users frequently need to ascertain the rating for a particular type of cable at a given voltage and with a range of copper or aluminium conductor sizes. The most common sources of reference are:

1. The IEE Regulations for the Electrical Equipment of Buildings. This source covers cables of all standard types (up to 1 kV). A difference from the others listed below is that the ratings quoted are lower, as they are calculated from a base ambient temperature of 30°C (compared with 25°C) and, hence, a lower permissible temperature rise. Only 'in air' ratings are included.
2. Report ERA 69-30 (8 parts), published by ERA Technology Ltd, provides ratings for paper cables up to 33 kV, PVC cables up to 3.3 kV and thermosetting insulated cables up to 3.3 kV. Guidance is also provided for cyclic and emergency ratings, together with further information relating to cables in typical specific installation conditions.
3. Manufacturers' catalogues.
4. Lower ratings may be selected for specialised installations because of particular environmental conditions. For example, the IEE Regulations for the Electrical and Electronic Equipment of Ships stipulate an ambient temperature of 45°C and somewhat lower maximum temperatures for continuous operation (80°C for thermosetting insulation).
5. The above cover installations based on British cable practice. When USA types of cable and system are applicable, reference may be made to ICEA publications.
6. IEC Publications 364-5-523 and 287 (*Table 31.1*): 364-5-523 provides ratings for unarmoured cables and 287 gives the basic methods for calculating ratings using the standard data included. Values prepared by all other bodies are almost always derived in accordance with this specification.

In general, these documents provide tabulated figures for copper and aluminium conductor cables installed in air, in ducts and buried directly in the ground. The data quoted are for standard conditions, and multiplying factors are given for variations in the conditions.

A feature of USA practice is that data provision is made for limited periods of operation with emergency overload for a specified number of hours per year to a higher temperature. While it is recognised that such operation could affect the life of the cable, the conditions are chosen to ensure that only limited ageing is likely to occur. British practice has not yet included this feature in published recommendations.

Another important aspect relates to the fact that the published ratings are quoted for 'continuous' or 'sustained' operation. Few cables are loaded for the whole of their life to full rating, and allowance is made for this. Nevertheless, the derivation of ratings is a most complex subject and many large users, such as the UK distribution companies, have developed ratings which allow for their own circumstances, such as cyclic operation and the emergency conditions which can arise with 11 and 33 kV cables normally installed as open rings (Electricity Council Engineering Recommendation P 17—Current Rating Guide for Distribution Cables).

A standard rating for the particular cable and specified installation conditions having been determined, factors have then to be applied to obtain the actual rating for the individual conditions. The references quoted provide these factors for variations such as ambient or ground temperature, depth of laying, thermal resistivity of soil and mutual heating due to cables being installed close together.

31.6.2 Factors in cable ratings

31.6.2.1 Temperature

As previously mentioned, ratings are governed primarily by the permissible temperature rise from a declared base temperature to a maximum for the particular cable design. The base temperature is normally 15°C for buried cables and either 25 or 30°C for cables in air. At the maximum continuous temperature the heat generated in the cable equates with the heat dissipation from it, which is dependent on the thermal resistance of the cable components and the surroundings.

The internationally recognised limits for conductor temperatures with the common types of insulation and cable design are shown in *Tables 31.9* and *31.10*.

In the case of the insulation materials, it is not usually chemical degradation which is the main aspect. With paper insulation the permissible temperature is reduced with increasing voltage and this ensures that there is not undue expulsion of impregnating compound for the duty required.

Table 31.9 Paper cables: conductor temperature limits

Rated voltage, U_0/U (kV)	*Design*	*Temperature* (°C)*
0.6/1, 1.8/3 and 3.6/6	Belted	80
6/10	Belted	65
6/10	Screened	70
8.7/15	Screened	70
12/20 and 18/30 MIND	Screened	65

* For continuous operation. Temperature for short-circuit conditions is 160°C, except for 0.6/1 kV cable, for which the limit is 250°C, subject to the accessories being suitable.

Table 31.10 Polymeric cables: conductor temperature limits

Insulating compound	*Temperature* (°C)*	
	Continuous	*Short-circuit*
Polyvinyl chloride	70	160†
Polyethylene	70	130
Butyl rubber	85	220
Ethylene propylene rubber	90	250
Cross-linked polyethylene	90	250

* Temperature limits are based on intrinsic properties and do not take account of variations in cable and accessory design. Short-circuit ratings are affected by (a) reduction of thickness of PVC and PE by thermomechanical forces; (b) conductor and core screens; (c) design of accessories (e.g. soldered conductor joints are unsuitable).

† 140°C for conductors above 300 mm^2.

Similarly, screened paper cables are more independent of compound effects in the filler spaces and can be operated to a higher temperature. Thermoplastic insulation softens significantly with increasing temperature and the limit is governed by deformation. Thermosetting materials can withstand much higher temperatures without undue deformation. Limitation by ageing effects is a factor with natural rubber compounds.

31.6.2.2 Conductor losses

With the exception of some higher voltage transmission cables, the I^2R conductor losses represent the major source of heat produced. These also have to be dissipated through the longest radial path in the cable. When the conductors are large, the effective resistance may also be increased because of skin effect. The increase is negligible for sizes up to about $185\,\mathrm{mm}^2$ and can be reduced by the use of the Milliken construction described earlier.

Proximity effects may be caused by the interaction of magnetic fields associated with adjacent current-carrying conductors and these, too, can cause further redistribution, as with skin effect. The proximity effect occurs with small spacing and so is most significant for low-voltage cables of large conductor size.

31.6.2.3 Dielectric losses

Dielectric losses are reasonably negligible for paper and XLPE cables up to about 60 kV and for PVC cables up to 6 kV. They are mainly of importance for high-voltage transmission cables. One reason that PVC has not found much application in the 10–20 kV field is that the losses are high in comparison with paper and XLPE. Even so, they represent only around 6–8% of the conductor losses in 11 kV cables. With XLPE the figure is around 0.1%.

31.6.2.4 Sheath and armour losses

Losses in metallic sheaths are of great importance for large conductor single-core cables bonded and earthed at both ends. As explained in Section 31.5.5, they can be avoided by cross-bonding. Although they make some contribution to total losses, the effect is not very significant for multicore cables. Similar remarks apply to armour, but losses due to magnetic effects are dominant for single-core cables and it is usually necessary to use non-magnetic armour material.

31.6.2.5 Internal thermal resistance

Thermal resistance within the cable is related to: (a) cable design and construction, e.g. the number of separate layers and the volume; and (b) the thermal resistivities and thicknesses of the individual materials. Values (in Km/W) included in IEC 287 are:

Impregnated paper, solid cables	6.0
Oil filled cables	5.0
Polyethylene, XLPE	3.5
Polyvinyl chloride	5.0/6.0
Ethylene propylene rubber	3.5/5.0
Bituminous textiles	6.0

31.6.2.6 External thermal resistance

For cables in free air the heat dissipation is related to the degree of exposure and to the surface emissivity, which depends on surface condition. Published ratings assume shading from the sun, and if this is not provided, derating may be necessary.

31.6.3 Sustained ratings

Table 31.11 indicates a typical example of published ratings for one type of cable installed in air, in ducts and buried directly in the ground. When cables are installed in ducts two other thermal resistances are introduced—namely, an air space between the cable and the duct, and the duct itself. As can be seen from *Table 31.11*, these cause a heavy rating penalty. *Table 31.11* also illustrates that while ratings for cables in air and buried direct in the ground are broadly similar, in air they are lower for small conductor sizes and higher for larger sizes. These differences are related to heat dissipation as a function of surface area.

Cables installed under water have the lowest external thermal resistance and highest ratings. However, there is always a danger that in time a layer of silt may build up, and investigations in canals have shown that such layers can have high thermal resistivity.

31.6.4 Short-time and cyclic ratings

A cable on load will show an exponential temperature rise/time relationship and, if starting from a low temperature, may take many hours to reach stable condition at maximum temperature. It can, therefore, carry more than maximum continuous rating for a limited time, the factor for overload depending on the extent of initial loading.

For cyclic loadings some increase of rating, compared with continuous, may be applied to an extent which will vary with the shape of the load curves. Calculations to take advantage of this possibility are rather tedious, but guidance may be obtained from ERA 69–30: Part IV (published by ERA Technology Ltd).

31.6.5 Short-circuit ratings

Often the conductor size necessary is related to short-circuit current rather than continuous current requirements. The

Table 31.11 Sustained current ratings and volt drop for triple-core copper, 0.6/1 kV, XLPE insulated, armoured cables

Conductor area (mm^2)	*Rating, dg* (A)	*Rating, sd* (A)	*Rating, air* (A)	*Volt drop* (mV/A/m)
16	119	96	107	2.5
25	152	124	134	1.7
35	182	149	165	1.2
50	217	177	201	0.87
70	266	218	256	0.60
95	319	263	316	0.45
120	363	300	369	0.37
150	406	338	423	0.30
180	458	382	489	0.26
240	529	442	582	0.21
300	592	496	672	0.19

Depth of laying, 0.5 m.
Soil thermal resistivity, 1.2 Km/W.
Ground temperature, 15°C.
Ambient air temperature, 25°C.
Maximum conductor temperature, 90°C.
Rating: dg, direct in ground; sd, in single-way ducts; air, in free air.

short-circuit current, which may be 20 or more times normal, produces thermal and electromagnetic effects proportional to the square of the current. So far as the cable insulation itself is concerned, much higher conductor temperatures can be allowed because the heating and cooling are very rapid and the full temperature will not be sustained for significant time by the insulation. Figures are included in *Tables 31.9* and *31.10*.

Short-circuit ratings are not published for individual cables in any official documents, because of the large number of the cable types and sizes involved and the fact that they have to be related to the duration of short circuit which applies to the particular circuit. *Figure 31.22* illustrates a typical example of the graphs available from manufacturers to provide information on a basis of a maximum conductor temperature with a range of durations.

Other factors may dictate a lower rating for a particular design of cable or installation condition. A short circuit in a cable produces electromagnetic forces which could burst the cable if the cores are not adequately bound together (single-core cables are a special case of this). The accessories must also be designed to withstand both electromagnetic and thermomechanical forces; and accessories must be compatible with the cable in this respect. Solderered joints impose limitation of the short-circuit temperatures to 160°C.

The method of installation may also limit permissible short-circuit current. Local pressure due to clamping may lead to high forces, with deformation of cable components. Longitudinal expansion can also be considerable and has to be absorbed uniformly. When cables are buried, the cable is restrained and these forces must be accommodated by joints and terminations.

31.6.6 Voltage drop

Voltage drop may be of great significance for 0.6/1 kV cables but is not usually important at higher voltages. A typical requirement in the IEE Regulations is that the voltage drop in a cable run should be such that the total drop in the circuit, of which the cable forms a part, does not exceed 4% of the nominal voltage.

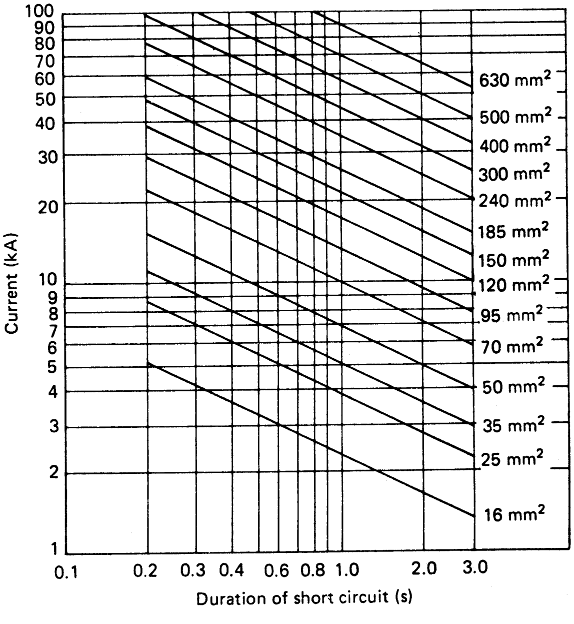


Figure 31.22 Short-circuit ratings for copper conductor XLPE insulated cables (temperature rise 90 to 250°C)

As the actual power factor of the load is seldom known, a practical approach is to assume the worst condition, i.e. where the phase angle of the load is equal to that of the cable. Cable manufacturers issue tabulated figures for volt drop, as in *Table 31.11*, based on this assumption. If the actual current differs greatly from the tabulated current the figure may be approximate only. From a table such as this, a suitable cable size may be selected but it must also be able to carry the current.

31.6.7 Protection against overload current

For some types of cable, particularly wiring cables, the required current rating of the cable must be determined by the overload protective device rather than the circuit current. The rating of the device must not be less than the circuit current and, of course, such ratings are in discrete steps. In the 16th edition of the IEE Wiring Regulations the protective device must satisfy the requirements of:

$I_B \leq I_n \leq I_z$

$I_2 \leq 1.45 I_z$

where (a) the nominal or current setting I_n shall not be less than the design current I_B of the circuit; (b) I_n does not exceed the lowest of the current-carrying capacities I_z of any of the conductors in the circuit; (c) the current causing effective operation of the protective device I_2 does not exceed 1.45 times the lowest of the current-carrying capacities I_z of any of the conductors of the circuit.

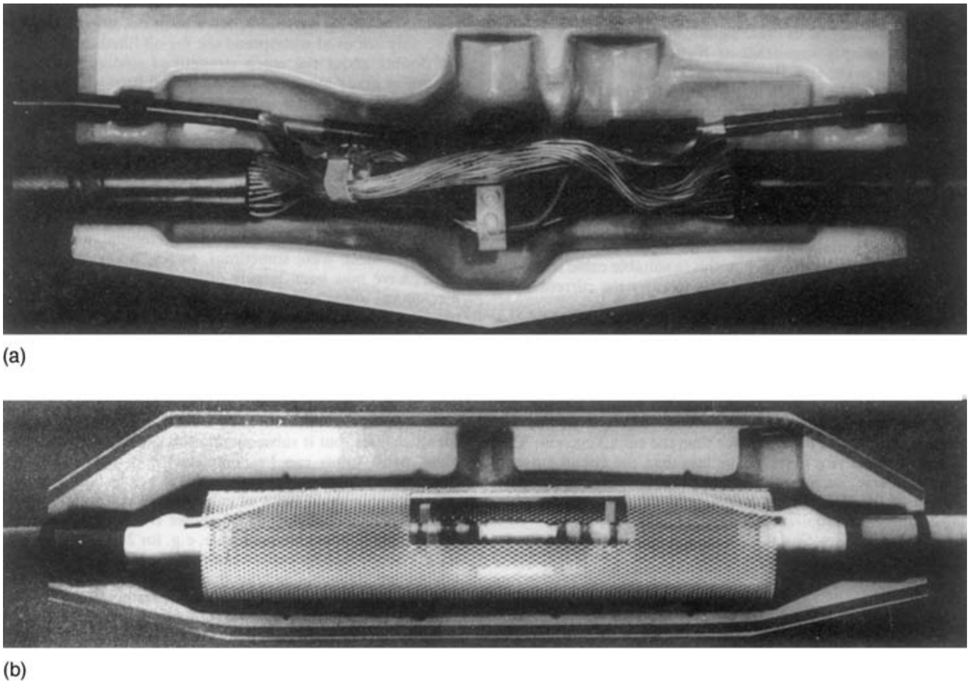
31.7 Jointing and accessories

31.7.1 Aluminium conductor jointing

The technique involving a flux for soldering with aluminium is essentially the same as for copper, but more care is required and strict observance of temperature limits is important. A special flux is necessary to remove the oxide skin and its composition should be without dermatitic risk. Solid conductors have an advantage in that a tinned layer can more readily be produced by the abrasion tinning procedure.

Problems with aluminium conductors, however, have largely been overcome by development of improved and simple compression methods suitable for both solid and stranded conductors. Probably the only remaining difficulty is that, when making straight joints on multicore tables, the larger separation necessary between cores for inserting the tool head makes the joints more bulky. Test requirements are given in BS 4579: Part 3.

Reference has been made to the problems with aluminium conductors in the tunnel type terminations of house-wiring fittings and these emphasise the care necessary in designing any mechanical fittings for aluminium. Suitable approved fittings are available for specific requirements such as the concentric neutral conductor in Waveform distribution cable, and for aluminium conductors in house-service cut-outs. Alternatively, for terminations, fittings are available for connecting a short piece of copper to an aluminium conductor.


(a)

(b)

Figure 31.23 Cast resin type joints before resin filling: (a) service joint on Waveconal cable; (b) straight joint on paper insulated cable

Where maximum strength and minimum volume of joint are essential, welding techniques have been developed, but they have only achieved widespread use for oil filled cables. There are doubts about the creep strength of soldered joints and jointing is made more difficult by the oil flow which must be maintained. The metal inert gas (MIG) welding procedure, which is essentially a casting process, proved to be practical and reliable.

31.7.2 Joints for distribution cables

The traditional practices of the past, which required highly skilled jointers for soldering and plumbing and involved materials which were sometimes, heavy, bulky and cumbersome, have now been largely dispensed with. By the use of mechanical joints and cast resin filling into simple shell type plastics moulds, all the components can be packed as a convenient complete kit and only simple tools are needed on site. The resin is poured while cold, and when mechanical fittings are used, no heating of any kind is required. At the time of mixing on site, the resin is fluid and penetrates well to fill all cavities, but it subsequently sets quickly to a hard solid mass. Most joints can be completed in 1–$1\frac{1}{2}$ h and can be energised immediately. The basic principles of cast resin jointing apply to all types of cable for all voltages up to and including 10 kV. *Figure 31.23* shows two typical examples. Where 'live' jointing is practicable, e.g. for 240/415 V services, it can equally well be carried out with these designs. The main advantages, however, accrue from the lower skill demanded and the short time required for jointer training.

Figure 31.24 shows a cross-section of a service joint on CNE cable taken through one of the phase connectors. With this type of connector it is not necessary to remove

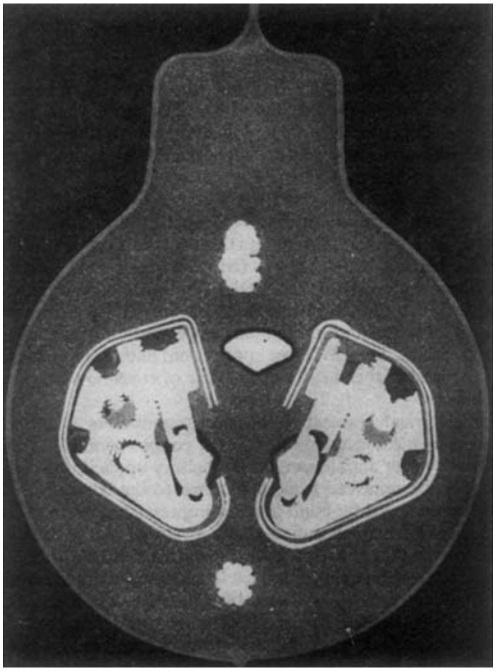


Figure 31.24 Section of cast resin service joint on Waveconal cable, showing the insulation piercing phase conductor connector

the insulation from the phase conductor, even if it consists of XLPE, because knife edges on the connector pierce the insulation and establish firm contact with the conductor.

31.7.3 Joints for transmission cables

Special fittings are required to joint and terminate transmission cables, the detailed design of which depends on the type of cable system employed. The following description of those used for low-pressure oil filled cables gives an indication of the important categories.

Straight joints The main requirements for joints connecting adjacent lengths of cables are: (a) to provide electrical continuity for the cable conductors; (b) to maintain the electrical insulation; (c) to provide an oil-tight connection between the sheaths; (d) in the case of cross-bonded single-core installations, to provide an insulating barrier between adjacent cable sheaths with facilities for cross connection; and (e) to maintain the insulation of the sheaths in the case of cross-bonded cable installations.

The electrical connection between conductors is usually provided by a compression ferrule for copper conductors and MIG welding for aluminium. As the factory-applied insulation has had to be removed to permit access to the conductors, it has to be replaced by hand-applied impregnated paper tapes and rolls. These are not electrically as strong as the factory-applied insulation and must be built up to a greater diameter. The different diameters for the cable and hand-applied insulation introduce longitudinal electric stresses in the joint. The stresses (and, hence, the shape of the boundaries) must be carefully controlled, as the electric strength of paper insulation along the laminations is only about 1/15 of that normal to the surface of the paper. *Figure 31.25* illustrates a typical straight joint for 33 kV triple-core cable.

Trifurcating joint This type of joint resembles a straight joint but is used to connect a three-core cable to three single-core cables, usually for terminating purposes.

Stop joint A stop joint provides all the functions of a straight joint but, in addition, separates the two adjacent cables hydraulically. It is used to limit the hydraulic pressure in a cable system installed when there are significant changes in elevation. This requirement has a major effect on the design, as it is necessary to introduce a different material in the electrical insulation to be capable of withstanding the hydraulic pressure difference. In modern designs of stop joint, this function is carried out by a moulded barrier of epoxy resin with a mineral filler. A stop joint for 132 kV single-core cable is shown in *Figure 31.26*.

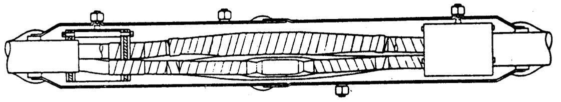


Figure 31.25 Straight joint for a 33 kV, three-core, OF cable

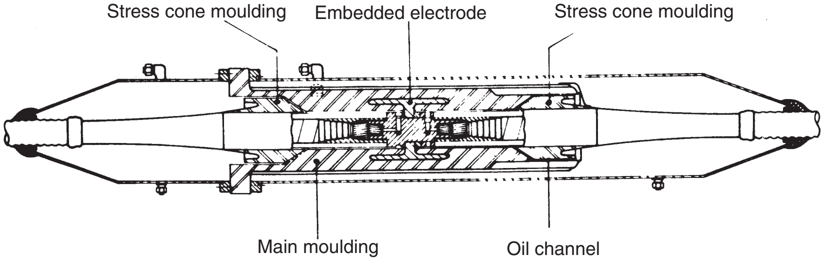


Figure 31.26 Single-core, 132 kV, OF cable stop joint

Outdoor termination The termination is enclosed in a porcelain housing, which retains the oil pressure and provides protection against climatic conditions. The porcelain has a long creepage path to allow for contamination by dirt, rain, fog and snow. As for straight joints, care must be taken to control the longitudinal stresses in the paper insulation. In very-high-voltage cable systems this is usually provided by means of a 'capacitor cone' similar to that used in high-voltage bushing.

Oil immersed termination This is similar to an outdoor termination but is used to make connection to oil immersed equipment. As there is no surface contamination from the atmosphere the length of the termination is appreciably shorter than the equivalent outdoor form.

SF_6 immersed termination Increasing use is now being made of SF_6 gas instead of oil for insulation. *Figure 31.27* shows a cable termination into SF_6 insulated metal-clad equipment. The cable is of the single-core 132 kV oil filled type with a copper conductor. The insulator is of the plug-in connector type, which employs a solid cast-in electrode, unpierced by the need for seals. The electrical stress control is of the capacitor cone type and is composed of cylinders of aluminium foil embedded in oil impregnated paper rolls.

31.8 Cable fault location

Irrespective of the type of cable, fault location demands a systematic approach if time and cost are to be minimised. Current practice is to adopt the four-step approach of diagnosis, preconditioning, prelocation and pinpointing.

31.8.1 Diagnosis

Diagnosis is used to confirm the existence of a fault and to determine its character. The dividing line between low- and high-resistance faults is based on the assumption that prelocation will be by a modern technique such as the

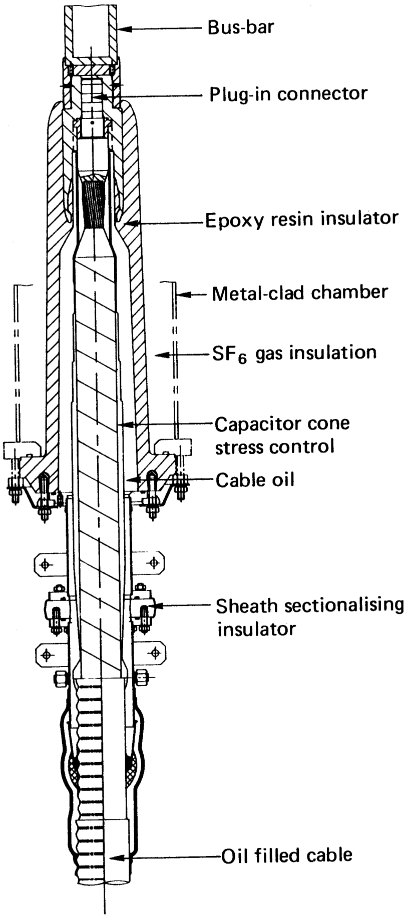


Figure 31.27 SF_6 plug-in sealing end for 132 kV OF cable

pulse-echo method, rather than by any of the classical bridge methods. It is essential that the fault resistance be measured using an ohmmeter and not an insulation tester such as a 'megger'. Cable continuity may be checked using a pulse-echo set, rather than an ohmmeter, with the advantage that breaks in the cable sheath as well as in the cable cores will be detected.

31.8.2 Preconditioning

Preconditioning (often referred to as 'fault burning') is used to change high-resistance faults to low-resistance ones which can then be prelocated using the pulse-echo method. Fault burners are usually designed to give various combinations of voltage and current output at ratings of up to 5 kVA. While reasonably successful on paper insulated cables (except on flashing and intermittent faults), fault burning has not proved to be effective on cables employing polymeric insulation, particularly XLPE. Prelocation techniques, such as the impulse current method, described below, have eliminated the need to precondition faults.

31.8.3 Prelocation methods

Prelocation is the application of a test at the terminals of a cable to give an indication of the distance of the fault from the test point. While the measurement should be as accurate as conditions allow, the primary purpose of the terminal tests is to give an indication, as quickly as possible, of the vicinity in which to commence the final pinpointing tests.

For many years cable fault prelocation methods were based on d.c. or low-frequency measurements of conductor resistance or capacitance. Most of these tests were performed using some form of bridge circuit which, when balanced, would indicate the ratio of the resistance, or capacitance, of the faulty conductor to that of a healthy one in the same cable. Versions of these 'classical' methods are still in use today but they have been largely replaced by 'modern' methods which are based on travelling-wave phenomena—the first to appear being the pulse-echo or radar method.

31.8.3.1 Pulse-echo method

Pulse-echo fault locators generate a short-duration pulse which is injected into the cable and which travels to the fault point at D (metres) and back in a time t (microseconds):

$$t = \frac{2D}{v}\,(\mu\mathrm{s}) \Leftarrow$$

where v (m/μs) is the speed at which the pulse travels along the cable and is determined almost totally by the permittivity of the cable dielectric ε, according to the formula:

$$v = \frac{300}{\sqrt{\varepsilon}}(\mathrm{m}/\mu\mathrm{s}) \Leftarrow$$

where 300 m/μs is the velocity of light in vacuo.

The advantages of the pulse-echo method, compared with bridge methods, are that it requires access to only one end of the cable and it is not necessary to perform any 'equivalent length conversions', provided the cable insulation is the same throughout the whole cable route. It is thus far less dependent on accurate cable records than are the bridge methods.

In the earliest instruments, the injected pulse and its reflection(s) were displayed on a cathode-ray tube but modern versions now use dot-matrix liquid crystal display (LCD) panels. Simplified instruments, for cable-length checking as opposed to cable fault location, may not provide a graphical display of the 'echogram', but only give an (automatic) readout of the time/distance to the reflection from the end of the cable.

Pulse-echo fault locators do, however, have one serious limitation in that the amplitude of the reflection produced by a fault depends on the ratio of the fault resistance R_f to the surge impedance of the cable Z_0, as shown in *Figure 31.28*. Series faults, i.e. faults affecting the continuity of one or more conductors of a cable, generally present no difficulty as they are generally complete open circuits producing 100% reflections (of the same polarity as the injected pulse). Shunt faults, i.e. faults affecting the cable insulation, produce (reversed polarity) reflections with amplitudes of less than 5% if the fault resistance is greater than 10 times the surge impedance of the cable. A reflection amplitude of 5% is about the smallest which can be easily identified from amongst the other 'background' reflections which arise on a jointed cable route. Comparison of the trace obtained from the faulty phase with that from a healthy phase in the same cable can simplify the interpretation and

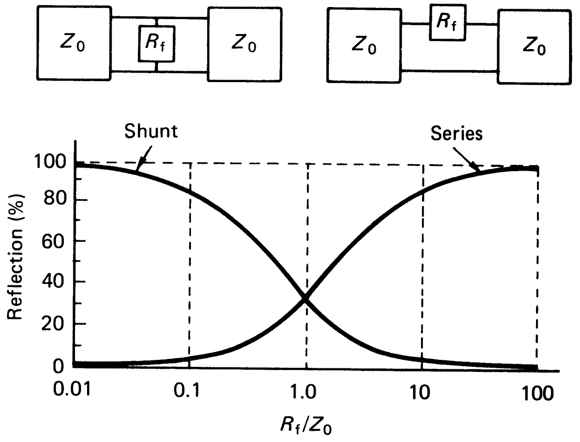


Figure 31.28 Relative amplitude of reflective pulse as a function of the ratio of fault resistance (R_f) to surge impedance (Z_0) for both series and shunt faults

extend the sensitivity to possibly 1%. With typical power cable surge impedances lying in the range 15–50 Ω, it is frequently impossible to locate shunt faults without the use of 'fault burning' or 'fault re-energisation'.

31.8.4 Cable fault characteristics

Representing cable faults simply as resistors is misleading, and a more realistic equivalent circuit, where the fault resistance is paralleled by both a spark gap and a capacitor, is shown in *Figure 31.29*. The values of all the elements of the equivalent circuit can vary widely and are completely independent of each other. The breakdown voltage V_b of the spark gap is determined by the distance between the two metallic boundaries of the fault which may be bridged by carbonised insulation in the case of a shunt fault or air spaced for an open circuit series fault. The value of the resistance is directly related to the degree of carbonisation of the insulation, whilst the value of the capacitance varies with the amount of moisture present. Based on the equivalent circuit shown in *Figure 31.29*, and the '5% reflection' limit for the pulse-echo method, cable faults can be conveniently classified as shown in *Table 31.12*.

Table 31.12 Classification of cable fault types

Fault type	R_f	*Spark gap*
Series	$\rightarrow\infty\Leftarrow$	Breakdown under impulse or d.c.
Low resistance	$<10 \times Z_0$	Breakdown under impulse, provided R_f is not too low
High resistance	$>10 \times Z_0$	Breakdown under impulse
Flashing	$\infty\Leftarrow$	Breakdown under impulse or d.c.
Intermittent	$\infty\Leftarrow$	Breakdown under prolonged d.c.

The existence of the fault 'spark gap' has been well known for many years, and it has been extensively exploited in the acoustic method of pinpointing—based on applying a high-voltage impulse to the cable to create an audible flashover at the fault point.

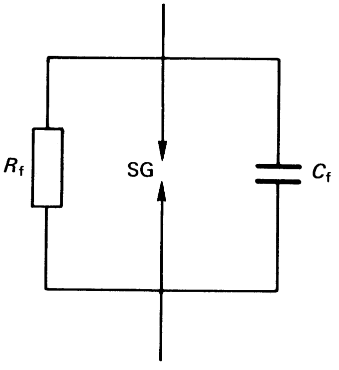


Figure 31.29 Equivalent circuit of a cable fault. SG, spark gap

31.8.4.1 Impulse-current method

The impulse-current method of fault location also exploits the existence of the spark gap, but in this case it is the electrical transient created by the breakdown, rather than the acoustic transient, which is of interest. This method overcomes the limitation of the pulse-echo method since it is applicable to every type of fault. Faults are located by detecting and recording the current signals flowing in the impulse generator circuit using a high speed digital transient recorder.

Low resistance shunt faults and open circuit series faults are located by identifying the direct reflections of the applied impulse in much the same way as with a standard pulse-echo instrument—albeit with reversed reflection polarities. Most faults, however, are located from the transients created by the flashover of the spark gap at the fault point which always occurs some time after the arrival of the voltage impulse due to a phenomenon known as 'ionisation delay'.

Figure 31.30(a) shows how a typical current transient is created under the application of an impulse voltage from a capacitor discharge generator where the impulse reflects from the open circuit end producing 'voltage doubling' which (eventually) results in breakdown of the fault spark gap. An actual recorded transient, produced under the conditions illustrated in *Figure 31.30(a)*,is shown in *Figure 31.31*. The transient can sometimes be simplified by increasing the applied voltage sufficiently to reduce the ionisation delay so that the fault breaks down before the reflection from the open-circuit end arrives back at the fault point—this is shown in *Figure 31.30(b)*. On flashing or intermittent faults, an even simpler transient can be produced by closing the impulse generator contactor and raising the voltage on the cable and the generator capacitor at the same time until flashover occurs, thereby producing the transient shown in *Figure 31.30(c)*.

The impulse-current method has now been adopted by every cable fault location equipment manufacturer in Europe, and is in widespread use throughout the world. As a method, it can accommodate a far wider range of fault and cable types than the pulse-echo method and requires only a simple and inexpensive 'linear coupler' to connect the recording instrument to the impulse generator. In one respect, however, it is inferior to the pulse-echo method, since the transient phenomena are far more complex and

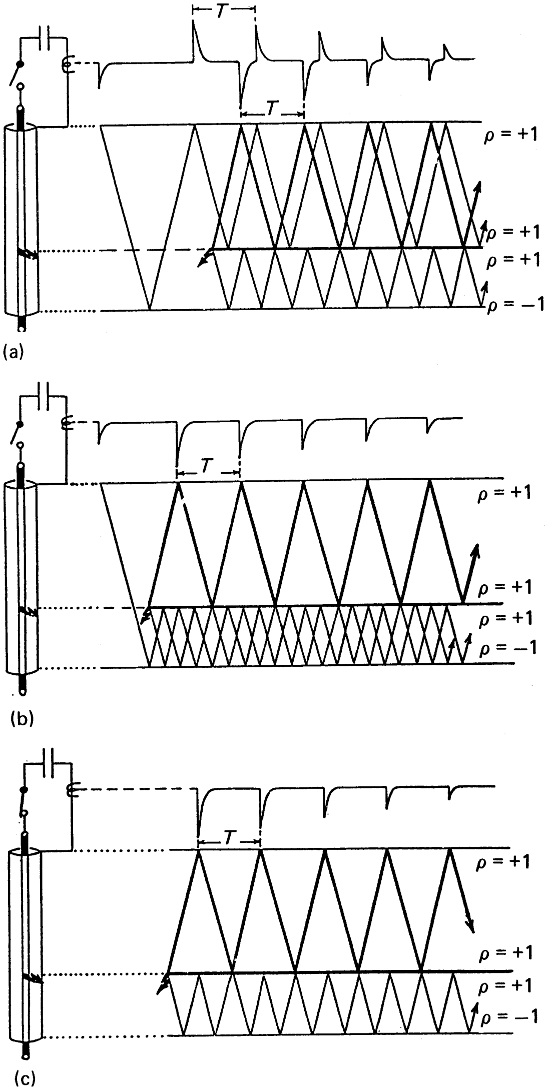


Figure 31.30 Current transients created by the application of an impulse voltage from a charged capacitor: (a) voltage doubling induced breakdown; (b) fault breakdown before reflected pulse; (c) simpler transient produced by closing the generator contactor

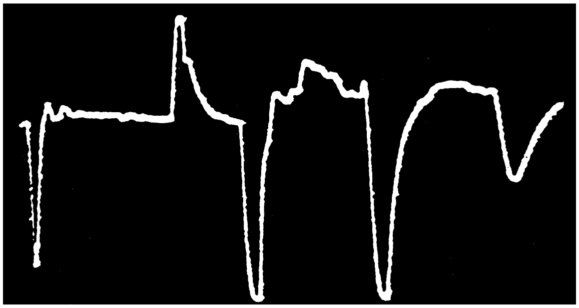


Figure 31.31 Recorded transient corresponding to the condition shown in *Figure 31.30*(*a*)

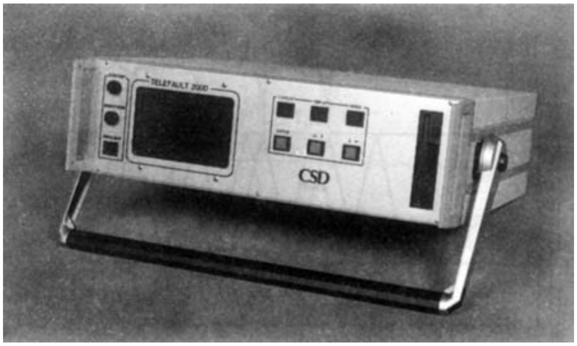


Figure 31.32 Cable fault analyser

require more experience to interpret. One solution to this problem has been the introduction of computer-aided equipment such as the cable fault analyser shown in *Figure 31.32.*

31.8.4.2 Secondary-impulse method

An alternative solution to the problem of the complexity of impulse current waveforms has been the development of the secondary-impulse method. Here the high voltage impulse generator is used to create a flashover of the fault spark gap and then, whilst current is still flowing through the ionised fault, a (second) low-voltage signal from a digital pulse echo instrument (*Figure 31.33*) is applied to the cable via a special high-voltage isolation and filter unit.

The secondary-impulse method cannot cope with as wide a range of cable and fault types as the impulse-current method, but interpretation of the waveforms is much simpler—a very important aspect when locating faults on the multi-branched networks frequently used in low-voltage distribution systems. A further complication of fault location on low voltage cables is that consumers' loads must often be assumed to be still connected to the cable, thereby precluding the application of any abnormal voltage

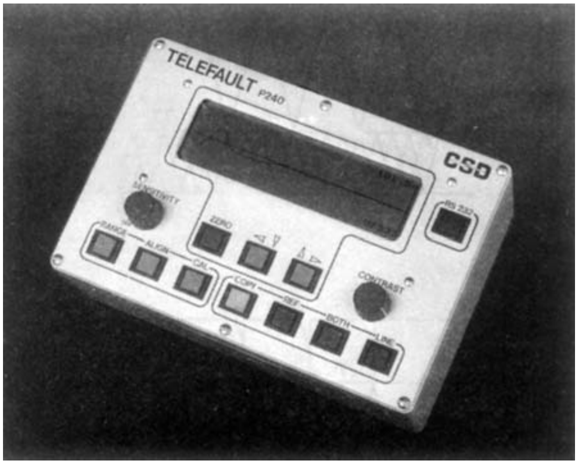


Figure 31.33 Digital pulse echo instrument

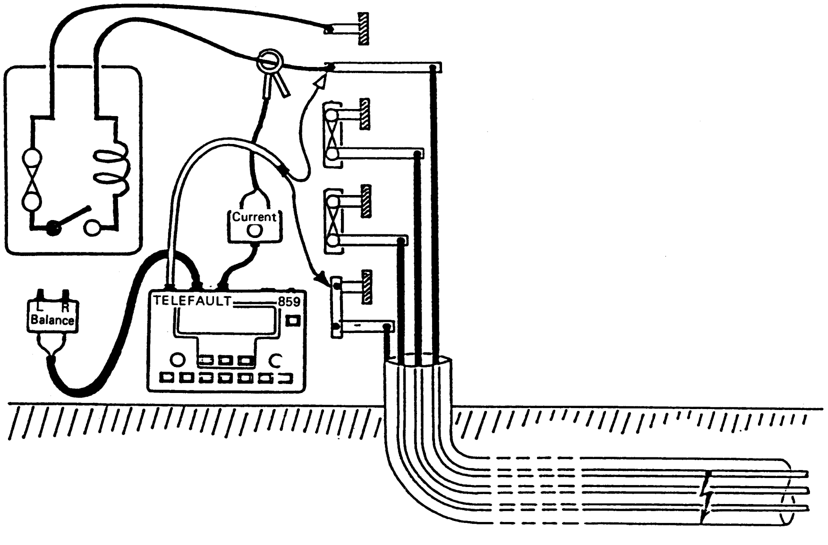


Figure 31.34 Test lead connected to the cable termination

to 'condition' faults—many of which may only be 'active,' and therefore detectable, when the cable is energised.

Connecting a pulse-echo set to an energised low-voltage power cable requires careful attention to safety, e.g. the provision of suitably fused test leads, and to methods of reducing the complexity of the waveforms—particularly the unavoidable mismatch which exists where the test lead is connected to the cable termination (*Figure 31.34*). A second test lead is connected to a 'balancing network' which is adjusted to present a complex impedance as close as possible to that created by the stray inductance of the main test lead connections onto the cable tails in combination with the surge impedance of the cable under test. The signals produced on the test lead connected to the cable are subtracted from those produced on the test lead connected to the balancing network by a hybrid transformer. With an effective method of removing the outgoing pulse, and its associated 'ringing' within the test lead connections, it is possible to increase the width of transmitted pulse without any loss of short-range discrimination. Wide pulses propagate further than narrow pulses as they not only contain more energy but also the spectral distribution of the energy is shifted towards lower frequencies where the cable attenuation is also less.

When using the secondary-impulse method to locate unstable faults on low-voltage power cables, it is necessary that the pulse echo set injects its signal into the faulty cable after the fault arc has been established. During the arc the effective resistance of the fault will be almost zero, producing a reflection amplitude approaching 100%. Before the arc is struck, or after it has extinguished, there is no reflection from the fault, but the reflections from all the other points of mismatch along the cable remain the same. One method of detecting when the fault arc exists is to use a clip-on current transformer, as shown in *Figure 31.34*. For convenience, and safety, the cable should be re-energised via a switching device which must include a series 'wavetrap' to prevent the injected pulse from reaching the substation busbar, and thereby all the other cables connected to it. When the re-energising device closes, there is usually a delay before the fault responds and the cable termination triggers the pulse-echo set. As can be seen from *Figure 31.35*, the operator's task is simply to compare the waveform recorded during the time when fault current is flowing with a waveform obtained when fault current is not flowing.

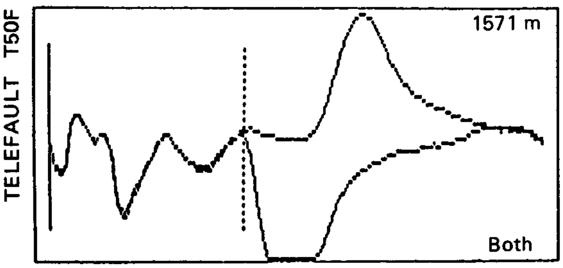


Figure 31.35 Typical secondary impulse method result

31.8.5 Pinpointing

Pinpointing is essential on direct buried cables if the location and repair of a fault is to be accomplished with a

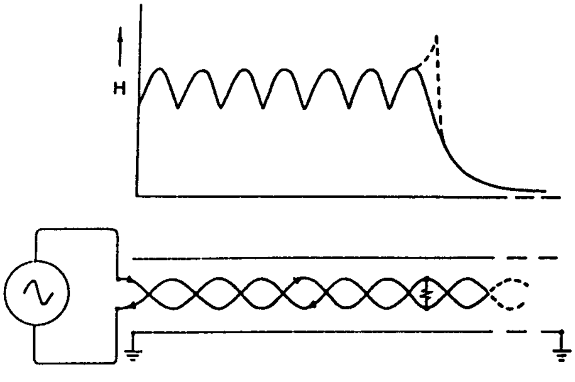


Figure 31.36 The AF induction method

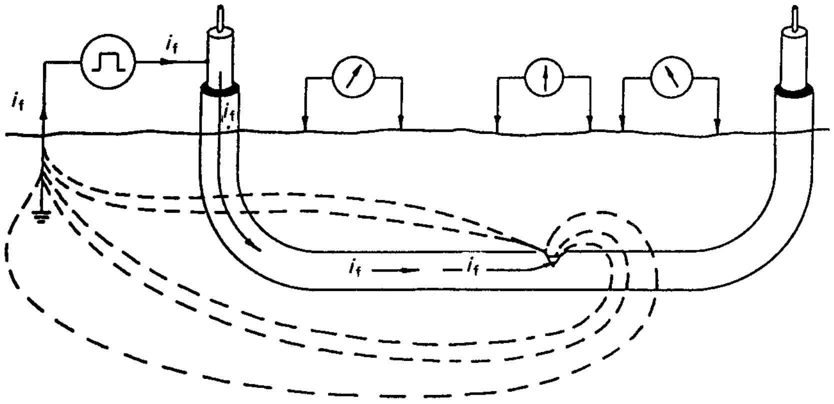


Figure 31.37 Pool of potential method for serving fault location

single excavation. The majority of faults on high-voltage power cables are pinpointed by detecting the acoustic signal generated when the fault spark gap breaks down either from the application of a voltage impulse from a surge generator or a high direct voltage from a pressure test set. In some cases the acoustic signal can be detected without any special equipment but, in general, it is an advantage to use a 'ground microphone' and amplifier to pick up the mechanical shock wave. The importance of the pinpointing stage cannot be overemphasised, and anything which might jeopardise the generation of the acoustic signal, such as prolonged preconditioning, should be avoided.

Once a fault develops into a very low-resistance 'welded' condition, it will 'short out' the spark-gap, making it impossible to generate an acoustic signal. Multiple excavations may then be unavoidable unless the fault is between two conductors or both conductors are welded to the cable sheath. If a low-resistance path exists between one conductor and another, it is possible to pinpoint the fault using the 'AF induction' or 'Bimec' method (*Figure 31.36*). Between the signal generator and the fault the signal induced in the search coil exhibits a characteristic rise and fall owing to the lay of the cable cores, while beyond the fault the signal either disappears completely or, more likely, becomes constant. The lay effect is the only positive means of identifying that the received signal is emanating from the faulty cable and is not caused by re-radiation from other adjacent buried metallic services. The induction method of pinpointing can be applied, subject to the necessary fault conditions, without any prior knowledge of the cable route and without prelocating the fault. As the appropriate fault conditions for the 'induction' method occur relatively infrequently, the main use of the technique is in cable route tracing when an artificial core-to-core fault is applied at the far end of the cable.

Both the acoustic and the induction pinpointing techniques require a complete metallic path in which the signal currents can flow. An alternative approach, used in pinpointing serving faults on insulated sheath transmission cables, is to apply a signal between the metallic sheath and the general mass of ground so that a voltage gradient is established in the earth in the vicinity of the fault. The voltage gradient or pool of potential is detected using a sensitive voltmeter connected to a pair of probes and the fault position is pinpointed accurately (*Figure 31.37*). Probing over the complete length of a long transmission cable is time consuming, and many serving faults are sectionalised, using a magnetometer to trace the current flowing in the sheath up to the fault point.

When cables are accessible, being installed above ground or still on the drum prior to installation, it is often possible to pinpoint faults by exploiting their 'microphonic' characteristics. A high-gain amplifier, a.c. coupled to a faulty cable, will pick up the small 'noise' voltages generated when the fault is subjected to mechanical vibration. Partially broken conductors can be detected and pinpointed by circulating constant d.c. through the cable; insulation faults can be detected using a high-voltage source to pass a small polarising current through the fault resistance.

Acknowledgements

The author is grateful to his past colleagues at BICC for their assistance in the preparation of this chapter, and in particular to Tony Arkell for his help in updating the section on transmission cables.

Further reading

General books

BUNGAY, E. W. G. and McALLISTER, D., *Electric Cables Handbook*, 2nd edn, BSP Ltd, London (1990)

GRANEAU, P., *The Science, Technology and Economics of High Voltage Cables*, Wiley, New York (1980)

HEINHOLD, L., *Power Cables and their Applications*, Siemens Aktiengesellschaft, Berlin (1990)

WEEDEY, B. M., *Underground Transmission of Electric Power*, Wiley, New York (1980)

Conductors

ANON., 'New wiring cable has copper-clad aluminium conductors', *Elec. Times*, 43–45 (2 July 1970)

BALL, E. H. and MASCHIO, G., 'The a.c. resistance of segmented conductors used in power cables', *IEEE Trans.*, **PAS-87** (1968)

CHATTERGEE, S., 'Failures of terminations on aluminium conductor cables in domestic wiring installations', *Paper No. 6*, Electrical Contractors Association of Eastern India (August 1977)

GRIESSER, E. E., ‘Sodium as an electrical conductor’, *Wire*, 2006–2014 (December 1966)

HUMPHREY, L. E. *et al.*, ‘Insulated sodium conductors’, *IEEE Trans.*, **PAS-86** (1967)

KALSI, S. S. and MINNICH, S. H., ‘Calculation of circulating current losses in cable conductors’, *IEEE Trans.*, **PAS-99** (1980)

McALLISTER, D., ‘Aluminium cables accepted by industry’, *Elec. Times* (10 September 1971)

McALLISTER, D., ‘Terminations for aluminium conductor power cable’, *BSI News* (August 1976)

Polymeric materials

BLOW, C. M., *Rubber Technology and Manufacture*, Newnes-Butterworth, London (1975)

BRYDSON, J. A., *Plastics Materials*, 4th edn, Butterworth Scientific, London (1982)

TOWN, W. L., ‘Using extrudable materials for cable insulation’, *Elec. Times* (23 February 1979)

Cables in fires

BENNETT, H. R., ‘Cables for limited fire performance’, *Elec. Times* (11 March 1977)

DAY, A. G., ‘Oxygen index tests: temperature effects and comparison with other flammability tests’, *Plast. Polym.*, **43** (106), 64 (1975)

LINDSTROM, R. S., *et al.*, ‘Effects of flame and smoke additives in polymer systems’, *JFF/Fire Retardant Chem.*, **1**, 152 (August 1954)

NATIONAL MATERIALS ADVISORY BOARD, *Flammability, Smoke, Toxicity and Corrosive Gases of Cable Materials* (*Publication NMAB-342*), National Academy of Sciences, Washington (1978)

NESS, D. E. M. and BLACK, R. M., ‘New materials for cable sheathing’, *IEE Electron. Power*, 698 (1982)

PHILBRICK, S. E., BUNGAY, E. W. G., BARBER, M. D. and WILLIAMSON, A. E., ‘Cables for new power stations’, *IEE Second Int. Conf. Power Cables and Accessories 10 kV to 180 kV* (*IEE Conference Publication No. 270*) (1986)

SULLIVAN, T. and WILLIS, A. J., ‘Reducing the hazards of cables in fires’, *Elec. Times* (17 Nov. 1978)

WHITE, T. M., ‘Cable fires in power stations’, *IEE Electron. Power* (February 1977)

General wiring type cables

ANON., ‘Wiring cables and flexible cords—new harmonised standards’, *BSI News* (February 1976)

ANON., ‘Selecting cables for the hostile conditions on construction sites’, *Elec. Times* (2 December 1977)

ANON., ‘Compatibility of PVC cables’, *Elec. Times* (4 August 1978)

BUNGAY, E. W. G. and HOLLINGSWORTH, P. M., ‘Progress with harmonisation of cable standards with CENELEC’, *Elec. Times* (2 April 1976)

HOLLINGSWORTH, P. M., ‘The effect of application on the choice of cable materials’, *Elec. Rev.* (24 April 1964)

HOLLINGSWORTH, P. M. and TOWN, W. L., ‘Trends in wiring cable design and installation’, *Elec. Rev.* (11 May 1973)

LATHAM, W. B., ‘Mineral insulated cables in hazardous areas’, *Elec. Times* (2 April 1976)

RODGERS, J. A., ‘Good wiring guide’, *Elec. Wholesaler* (October 1972)

SECCOMBE, G. H., ‘EEC and cable standards: where are we going?’, *Elec. Times* (2 January 1975)

TAYLOR, F. G., ‘Cables for electronics’, *Electrotechnology* (March 1973)

TODD, D., ‘Electric cable for signalling and track to train communications’, *I. Mech. E.R.E.J.* (September 1975)

TOWN, W. L., ‘Wiring cables’, *Elec. Times* (22 June, 6 July 1972)

TOWN, W. L., ‘A guide to the selection of electrical and electronics wires and cables’, *OEM Design* (April 1974)

WILSON, I. O., ‘Magnesium oxide as a high temperature cable insulant’, *Proc. IEE, Part A*, **128**(3), 159–164 (April 1981)

Power cables (general)

BALDOCK, A. T. and HAMBROOK, L. G., ‘Regulations relevant to the design and utilisation of distribution cables for the Electrical Supply industry and the consumer’, *IEE Conf. on Distribution* (May 1976)

BLECHSCHMIDT, H. H. and GOEDECKE, H. P., *Cables with Synthetic Insulation in the Federal Republic of Germany*, CIRED, Liege (1975)

BUNGAY, E. W. G., *et al.*, *The Development of 11 kV Cable Systems*, CIRED, Liege (1975)

BUNGAY, E. W. G. and PHILBRICK, S. E., ‘Paper-insulated 11 kV aluminium-sheathed cables’, *IEE Conf. on Distribution* (May 1976)

GOSDEN, J. H., *Reliability of Overhead Line and Cable Systems in Great Britain*, CIRED, Liege (1975)

GOSDEN, J. H. and WALKER, A. J., ‘The reliability of cable circuits for 11 kV and below’, *IEE Conf. on Distribution* (May 1976)

HOWARD, R. S., JENKINS, T. and BROOK, R. T. ‘Operating experience with 11 kV polymeric cable systems in one U.K. Area Board’, *IEE Second Int. Conf. on Power Cables and Accessories 10 kV to 180 kV* (*IEE Conference Publication No. 270*), 31–36 (1986)

HYDE, H. B., LE POIDEVIN, G. J. and PHILBRICK, S. E., ‘The development of a single core polymeric cable for 33 kV distribution systems’, *IEE Second Int. Conf. on Power Cables and Accessories 10 kV to 180 kV* (*IEE Conference Publication No. 270*), 46–50 (1986)

REYNOLDS, E. H. and ROGERS, E. C., ‘Discharge damage and failure in 11 kV belted cables’, *Trans. S. Afr. Inst. Electr. Eng.*, **52**(10), (October 1961)

ROSS, A., ‘Cable practice in (UK) Electricity Board distribution networks: 132 kV and below’, *Proc. IEE*, **121**(11R), 1307–1344 (November 1974)

SCHMELTZ, J., ‘EDF distribution cables’ *8th Cables Seminar, Hanover*, (*Electricity Council Translation OA 2116*) (28 September 1976)

SWARBRICK, P., ‘Developments in 11 kV underground cable systems. Paper-insulated aluminium sheathed cables and resin filled joints’, *Elec. Rev.* (14 Dec. 1973)

WHITE, T. M., *et al.*, ‘11 kV polymeric insulated triplex cable’, *IEE Second Int. Conf. on Power Cables and Accessories 10 kV to 180 kV* (*IEE Conference Publication No. 270*), 41–45 (1986)

CNE cables

BURTON, J. M., ‘Consac cable system development in the Midlands Electricity Board’, *IEE Conf. on Distribution* (May 1976)

GEER, P. K. and SLOMAN, L. M., ‘Cables for PME distribution systems’, *IEE Conf. on Distribution* (May 1976)

HENDERSON, J. T. and SWARBRICK, P., ‘The Consac cable system’, *IEE/ERA Conf. on Distribution*, Edinburgh (1970)

HUGHES, O. I. and BROMLEY, G. E. A., 'Development and production of a PME elastomeric insulated MV cable', *IEE/ERA Conf. on Distribution*, Edinburgh (1970)

McALLISTER, D. and COX, E. H., 'Behaviour of MV power distribution cables when subjected to external damage', *Proc. IEE*, **119**(4), 479–486 (April 1972)

RADCLIFFE, W. S. and McALLISTER, D., 'Cables and joints for PME distribution systems', *Elec. Rev.* (24 March 1972)

ROCKCLIFFE, R. H., HILL, E. and BOOTH, D. H., 'Protective earthing practices in the UK and their associated underground cable systems', *Paper 71C 42-PWR*, *IEEE Conf. On Power Distribution* (1971)

Polyethylene and XLPE insulated cables

BAHDER, G., KATZ, C., LAWSON, J. H. and VAHLSTROM, W., 'Electrical and electrochemical treeing effect in polyethylene and cross-linked polyethylene cables', *IEEE Trans.*, **PAS-93**, 977–990 (May/June 1974)

BAHDER, E., KATZ, C., GARCIA, F. C., WALLDORF, F., CHAROY, M., FAVRIE, E. and JOCTEUR, R., 'Development of extruded cables for EHV applications in the range 138–400 kV', *CIGRE Paper 21–11*, Paris (1978)

BERNSTEIN, B. S., *Research to Determine the Acceptable Emergency Operating Temperatures for Extruded Dielectric Cables* (*EPRI Report EL-938*) (November 1978)

DENSLEY, R. J., 'An investigation into the growth of electrical trees in XLPE insulation', *IEEE Trans.*, **EI-14**(3), 148–158 (June 1979)

EICHHORN, R. M., 'Treeing in solid extruded electrical insulation', *IEEE Trans.*, **EI-12**(1) (February 1976)

FERREN, J. and PINET, A., 'Development of a new 20 kV cable with synthetic insulation and of its fittings', *CIRED Paper 31*, Liege (1979)

HYDE, H. B. *et al.*, 'Earth fault spiking tests at system voltage on 11 kV polymeric cables', *IEE Conf. on Distribution* (May 1976)

JACOBSON, C. T., ATTERMO, R. and DELLBY, B., 'Experience of dry-cured XLPE-insulated high voltage cables', *CIGRE Paper 21–06*, Paris (1978)

JORGENSON, J. and NEILSON, O. K., 'Straight-through joints for extruded solid dielectric insulated cables, 12–170 kV', *CIRED Paper 33*, Liege (1979)

LACOSTE, A., ROYERE, A., LEPERS, J. and BENART, P., 'Experimental construction prospects for the use of 225 kV, 600 MVA links, using polyethylene insulated cables with forced external water cooling', *CIGRE Paper 21-12*, Paris (1974)

LANCTOA, T. P., LAWSON, J. H. and McVEY, W. L., 'Investigation of insulation deterioration in 15 kV and 22 kV polyethylene cables removed from service—Part 3', *IEEE Trans.*, **PAS-98**(3), 912–925 (May/June 1979)

LAWSON, J. H. and THUE, W. A., 'Summary of service failure of high voltage extruded dielectric insulated cables in the USA', *IEEE Symp. on Electrical Insulation* (June 1980)

LAWSON, J. H. and VAHLSTROM, W., 'Investigation of insulation deterioration in 15 kV and 22 kV cables removed from service—Part 2', *IEEE Trans.*, **PAS-92**, 824–835 (March/April 1973)

MATSUBA, H. and KAWAI, E., 'Water tree mechanism in electrical insulation', *IEEE Paper F 75* (1975)

McKEAN *et al.*, 'Investigation of mechanism of breakdown in XLPE cables', *EPRI Report 7809-1* (1976)

NAYBOUR, R. D., 'The growth of water trees in XLPE at operating stresses and their influence on cable life', *IEE 3rd Int. Conf. on Dielectric Materials* (10 September 1979)

NAYBOUR, R. D. and PAPADOPULOS, M. S., 'Water trees in polymeric insulated cables', *Distribution Developments*, 30–37 (1988)

NUNES, S. L. and SHAW, M. T., 'Water treeing in polyethylene—a review of mechanisms', *IEEE Trans.*, **EI-5**(6), 437–450 (December 1980)

PINET, A. and FERRAN, J., 'Operating experience with the 20 kV XLPE cable used on the French network', *IEE Second Int. Conf. on Power Cables and Accessories 10 kV to 180 kV* (*IEE Conference Publication No. 270*), pp. 37–40 (1986)

SHAW, M. T. and SHAW, S. H., 'Water treeing in solid dielectrics', *IEEE Trans.*, **EI-19**(5), 419–452 (1984)

SWARBRICK, P., 'Developments in XLPE Cables', *Elec. Rev.* (7 January 1977)

SWARBRICK, P., 'Developments in the manufacture of XLPE cables' *Elec. Rev.* (28 January 1977)

TABATA, T., NAGAI, H., FUKUDU, T. and IWATA, Z., 'Sulphide attack and treeing of polyethylene insulated cables—cause and prevention', *IEEE Summer Meeting, Paper 71-TP 551-PWR* (1971)

TANAKA, T., FUKUDU, S., SUZUKI, Y. and NITTA, H., 'Water trees in cross-linked polyethylene power cables', *IEEE Trans.*, **PAS-93**, 693–702 (March/April 1974)

VAHLSTROM, W., 'Investigation of insulation deterioration, in 15 kW and 22 kV polyethylene cables removed from service', *IEEE Trans.*, **PAS-91**, 1023–1035 (May/June 1972)

WHITE, T. M., GIBBS, J. W., PHILLIPS, R. B., HYDE, H. B., PHILBRICK, S. E. and BUNGAY, E. W. G., '11 kV polymeric insulated triplex cable', *IEE Second Int. Conf. on Power Cables and Accessories 10 kV to 180 kV* (*IEE Conference Publication No. 270*) (1986)

Transmission cables

AIHARA, M., FUJIKI, S., KATO, N., MAGASAKI, S., NAKAGAIRA, M. and YOSHIDA, N., 'Philosophy of design and experience on high voltage XLPE cables and accessories in Japan', *CIGRE Paper No. 21–01*, Paris (1988)

ALEXANDER, S. M. *et al.*, 'Rating aspects of the 400 kV West Ham–St John's Wood cable circuits', *2nd IEE Int. Conf. Progress Cables for 220 kV and Above* (September 1979)

ALLAM, E. M. and McKEAN, A. L., 'Design of an optimised ±600 kV d.c. cable system', *IEEE Trans.*, **PAS-99** (Sep. 1980)

ALLAM, E. M. and McKEAN, A. L., 'Laboratory experiments of a ±600 kV d.c. pipe type cable', *IEEE Trans.*, **PAS-100** (March 1981)

ARKELL, C. A., 'Self-contained oil-filled cable: installation and design techniques', *IEEE Underground and Transmission Conf.*, 497–502 (1976)

ARKELL, C. A. and PARSONS, A. F., 'Insulation design of self-contained oil-filled cables for D.C. operation', *IEEE Trans.*, **PAS-101**, 1805–1814 (1982)

ARKELL, C. A., ARNAUD, U. G. and SKIPPER, D. J., 'The thermomechanical design of high power, self-contained cable systems', *CIGRE Paper 21–05*, Paris (1974)

ARKELL, C. A., *et al.*, 'Design and construction of the 400 kV cable system for the Severn Tunnel', *Proc. IEE, 124 (3)*, 303–316 (March 1977)

ARKELL, C. A., GREGORY, B. and SMEE, G. J., 'Self-contained oil-filled cables for high power circuits', *Paper F 77 547-3*, *IEEE PES Summer Meeting*, Mexico City 1977

ARKELL, C. A., HUTSON, R. B. and NICHOLSON, J. A., 'Development of internally oil-cooled cable systems', *Proc. IEE*, **124**(3), 317–325 (March 1977)

ARKELL, C. A., *et al.*, 'The design and installation of cable systems with separate pipe water cooling', *CIGRE Paper 21–01*, Paris (1978)

ARKELL, C. A., *et al.*, 'Development of polypropylene paper laminate OF cable UHV systems', *CIGRE Paper 21-04*, Paris (1980)

ARKELL, C. A., GALLOWAY, S. J. and GREGORY, B., 'Supertension cable terminations for metalclad SF_6 insulated substations', *Eighth IEEE/PES Conf. and Exposition on Overhead and Underground Transmissions and Distribution* (September 1981)

ARKELL, C. A., BALL, E. H., HACKE, K. J. H., WATERHOUSE, N. H. and YATES, J. B., 'Design and installation of the U.K. part of the 270 kV d.c. cable connection between England and France, including reliability aspects', *CIGRE Paper No. 21-02*, Paris (1986)

ARNAUD, U. G., *et al.*, 'Development and trials of the integral pipe cooled e.h.v. cable system', *Proc. IEE*, **124**(3), 286–293 (March 1977)

BALL, E. H. and HOLDUP, H. W., *Development of Cross-linked Polyethylene Insulation for High Voltage Cables*, CIGRE, Paris (1984)

BALL, E. H., *et al.*,'Connecting Dinorwic pumped storage power station to the grid system by 400 kV underground cables', *Proc. IEE,* **126**(3) (March 1979)

BEALE, H. K., 'Underground cables for HV power transmission', *CEGB Res.*, 24-32 (June 1979)

BROOKS, E. J., GOSLING, C. H. and HOLDUP, W., 'Moisture control of cable environment with particular reference to surface troughs', *Proc. IEE*, **120**(1), 51–60 (Jan. 1973)

CRABTREE, I. M. and O'BRIAN, M. T., 'Performance of the Cook Strait ±250 kV d.c. submarine cables, 1964–1985', *CIGRE Paper No. 21–01*, Paris (1986)

DONAZZI, F., OCCHINI, E. and SEPPI, A., 'Soil thermal and hydrological characteristics in designing underground cables', *Proc. IEE*, **126**(6), 505–516 (June 1979)

EDWARDS, D. R., 'Supertension or superconducting cables. 20 years on—can the U.K. regain the lead', *Proc. IEE, Part C*, **135**(1), 9–23 (January 1988)

ENDACOTT, J. D., 'Underground power cables', *Phil. Trans. R. Soc. London, Series A*, **275**, 193–203 (1973)

ENDACOTT, J. D. *et al.*, 'Progress in the use of aluminium in duct and direct buried installations of power transmission cable', *IEEE Underground Transmission and Distribution Conf.*, 466–474 (1974)

FAVRIE, E. and AUCLAIR, H., '225 kV low density polyethylene insulated cables', *2nd IEE Int. Conf. Progress Cables 220 kV and Above* (September 1979)

GREGORY, B. and LINDSEY, G. P., 'Improved accessories for supertension cable', *CIGRE Paper No. 21-03*, Paris (August 1988)

GREGORY, B. and NICHOLLS, A. W., '66 kV and 132 kV XLPE supertension cable systems', *Sixth Conf. on Electric Power Supply Industry. Transmission and Distribution Systems and Equipment*, *Paper No. 3.08* (November 1986)

GREGORY, B. and VAIL, J., 'Accessories for 66 kV and 132 kV XLPE cables', *Second Int. Conf. on Power Cables and Accessories 10 kV to 180 kV* (*IEE Publication No. 270*), pp. 248–256 (November 1986)

HEAD, J. G., GALE, P. S. and LAWSON, W. G., 'Effects of high temperatures and electric stresses on the degradation of OF cable insulation', *IEE 3rd Int. Conf. On Dielectric Materials* (10 September 1979)

HEUMANN, H. *et al.*, '380 kV oil-filled cable for municipal power supply in Vienna', *2nd IEE Conf. Progress Cables for 220 kV and Above* (September 1979)

ITOH. H., NAKAGAWA, M. and ICHIMO, T., 'EHV self-contained oil-filled cable insulated with composite paper, DCLP', *2nd IEE Int. Conf. Progress Cables 220 kV and Above* (September 1979)

IWATA, Z., ICHIYANAGI, N. and KAWAI, E., 'Cryogenic cable insulated with oil-impregnated paper', *IEEE Winter Meeting* (1977)

KOJIMA, K. *et al.*, 'Development and commercial use of 275 kV XLPE insulated power cables', *IEEE Trans.*, **PAS-100** (January 1981)

KUSANO, T. *et al.*, 'Practical use of "Siolap" insulated OF cables', *IEE Paper WM 114-8* (1981)

LOOMS, J. S. T., Insulators for High Voltages, *IEE Power Engineering Series 7*, Peter Peregrinus, London (1988)

MAINKA, A. G., BRAKELMANN, H. and RASQUIN, W., 'High power transmission with conductor cooled cables', *CIGRE Paper 21-10*, Paris (1978)

MIRANDA, F. J. and GAZZANA PRIOROGGIA, P., 'Self-contained oil-filled cables. A review of progress', *Proc. IEE*, **123**(3), 229–238 (March 1976)

MIRANDA, F. J. and GAZZANA PRIOROGGIA, P., 'Recent advances in self-contained oil-filled cable systems', *IEE Electron. Power*, 136–140 (Feb. 1977)

MITZUKAMI, T. *et al.*, 'Prototype tests of EHV cryoresistive cable', *IEEE Trans.* **PAS-99** (March/April 1980)

OCCHINI, E. *et al.*, 'Self-contained oil-filled cable systems for 750 kV and 1100 kV. Design and tests', *CIGRE Paper 21-08*, Paris (1978)

PEORMAN, A. J. *et al.*, 'Preliminary ageing tests on a super-conducting cable dielectric', *IEEE Symp. on Electrical Insulation*, 132–135 (1980)

RAY, J. J., ARKELL, C. A. and FLACK, H. W., '525 kV self-contained oil-filled cable systems for Grand Coulee third powerplant—design and development', *IEEE Paper T 73*, 492–496 (1973)

ROGERS, E. C., SLAUGHTER, R. J. and SWIFT, D. A., 'Design for a superconducting a.c. power cable', *Proc. IEE*, **118**(10) (October 1971)

ROSEVEAR, R. D. and VECELLIO, B., 'Cables for 750/1100 kV transmission', *2nd IEE Int. Conf. Progress Cables for 200 kV and Above* (September 1979)

ROSEVEAR, R. D., WILLIAMS, G. and PARMIGIANI, 'High voltage XLPE cable and accessories', *Second Int. Conf. on Power Cables and Accessories 10 kV to 180 kV* (*IEE Publication No. 270*), pp. 232–237 (November 1986)

SHIMSHOCK, J. F., *Installed Cost Comparison for Self-contained and Pipe-type Cable* (*EPRI Report EL-935*) (Nov. 1978)

SPENCER, E. M., *et al.*, 'Research and development of a flexible 362 kV compressed gas insulated transmission cable', *CIGRE Paper 21-02*, Paris (1980)

SMEE, G. J. and WEST, R. S. V., 'Factors influencing the choice between paper and XLPE insulated cables in the voltage range 66 kV–132 kV', *Second Int. Conf. on Power Cables and Accessories 10 kV to 180 kV* (*IEE Publication No. 270*), pp. 193–197 (November 1986)

Ratings

BUCKINGHAM, G. S., 'Short-circuit ratings for mains cables', *Proc. IEE*, **108**(A), (June 1961)

BUCKINGHAM, G. S., BOWIE, G. A. and PARR, R. G., 'Current ratings of PVC and polythene insulated cables', *IEE/ERA Symp. on Plastic Insulated Cables* (1962)

GOSDEN, J. H. and KENDALL, P. G., 'Current ratings of 11 kV cables', *IEE Conf. on Distribution Cables and Jointing Techniques* (May 1975)

GOSLAND, L. and PARR, R. G., 'A basis for short-circuit ratings for paper insulated cables up to 11 kV', *Proc. IEE*, **108**(A) (June 1961)

MOCHLINSKI, K., 'Assessment of the influence of soil thermal resistivity on the ratings of distribution cables', *Proc. IEE*, **123**(1) (January 1976)

PARR, R. G., *Bursting Currents of 11 kV 3-core Screened Cables* (*Paper-insulated Lead-sheathed*) (*ERA Report F/T 202*) (1962)

Jointing and accessories

CROSSLAND, J., 'Joints on 3-core 11 kV paper insulated cables', *IEE Conf. on Distribution* (May 1976)
JORGENSEN, J. and NIELSEN, O. J., *Straight Joints for Solid Dielectric Insulated Cables of 12–170 kV*, CIRED, Liege (1979)
McALLISTER, D. and RADCLIFFE, W. S., 'Joints incorporating mechanical connectors and cast resin filling for 600/1000 V cables', *IEE Conf. on Distribution* (May 1976)
RADCLIFFE, W. S. and ROBERTS, B. E., 'Resin-filled joints for 11 kV paper insulated cables', *IEE Conf. on Distribution*, (May 1976)
ROSS, A., 'Jointing trials and tests on 11 kV aluminium sheathed cables', *IEE Conf. on Distribution* (May 1976)
WHYTE, D. H., 'A resin jointing system for all voltages', *Elec. Equip.* (November 1979)

32

HVDC

A Gavrilović OBE

The author thanks his colleagues for permission to make extensive use of their material: J D Ainsworth, B R Anderson, M H Baker, R Banks, H Gibson, F G Goodrich, C J B Martin, B A Rowe, H L Thanawala, M L Woodhouse (ALSTOM)

Contents

32.1 Introduction 32/3

32.2 Applications of HVDC 32/4
  32.2.1 Introduction 32/4
  32.2.2 Types of d.c. interconnection 32/4
  32.2.3 Purposes of transmission interconnections 32/4
  32.2.4 Reasons for choosing HVDC 32/4
  32.2.5 Application of HVDC to developing systems 32/5

32.3 Principles of HVDC converters 32/5
  32.3.1 Converter operation: simplified case of zero commutating inductance 32/5
  32.3.2 Converter operation: practical case of finite commutating inductance 32/6
  32.3.3 Converter operation: converter acting as an inverter 32/6
  32.3.4 Twelve-pulse converters 32/7
  32.3.5 Basic d.c. voltage/d.c. current characteristics 32/7
  32.3.6 Basic principles of control of HVDC transmission 32/7
  32.3.7 Starting and stopping an HVDC link 32/8
  32.3.8 Power reversal 32/9
  32.3.9 Isolating a valve group 32/9
  32.3.10 Numerical example 32/9

32.4 Transmission arrangements 32/9
  32.4.1 Bipolar lines 32/9
  32.4.2 Two monopolar lines 32/10
  32.4.3 Cable schemes with sea return 32/10
  32.4.4 Back-to-back arrangement 32/10
  32.4.5 Ground electrodes 32/10
  32.4.6 Sea electrodes 32/11
  32.4.7 Staged construction of HVDC 32/11

32.5 Converter station design 32/11
  32.5.1 Valve group arrangements 32/11
  32.5.2 Converter valves 32/13
  32.5.3 Converter transformers 32/13
  32.5.4 A.c. filters 32/14
  32.5.5 D.c. smoothing reactor 32/14
  32.5.6 D.c. isolators 32/15
  32.5.7 Protection 32/15
  32.5.8 Converter station losses 32/15
  32.5.9 Converter stations' prices 32/15
  32.5.10 Reliability 32/16

32.6 Insulation co-ordination of HVDC converter stations 32/16
  32.6.1 Introduction 32/16
  32.6.2 Sources of overvoltages 32/16
  32.6.3 Surge arresters 32/17
  32.6.4 Surge arrester arrangement 32/17
  32.6.5 Safety margins 32/17
  32.6.6 Creepage and clearance 32/18
  32.6.7 Application examples 32/18

32.7 HVDC thyristor valves 32/19
  32.7.1 Introduction 32/19
  32.7.2 Thyristor level circuits 32/20
  32.7.3 Voltage rating 32/20
  32.7.4 Current rating 32/21
  32.7.5 Turn-on behaviour 32/22
  32.7.6 Turn-off behaviour 32/22
  32.7.7 Valve arrangements 32/23
  32.7.8 Valve tests 32/24

32.8 Design of harmonic filters for HVDC converters 32/24
  32.8.1 Introduction 32/24
  32.8.2 A.c. harmonic current generation 32/24
  32.8.3 Filtering 32/26
  32.8.4 Harmonic performance evaluation 32/27
  32.8.5 D.c. filtering 32/28

32.9 Reactive power considerations 32/29
  32.9.1 Introduction 32/29
  32.9.2 Reactive power requirements of HVDC converters 32/29
  32.9.3 Steady-state voltage control and total ratings of reactive equipment 32/29
  32.9.4 Voltage disturbances caused by switching operations and requirements for smooth reactive control 32/29
  32.9.5 Control of temporary overvoltages caused by faults resulting in partial or total loss of d.c. power flow 32/29

32.10 Control of HVDC 32/30
  32.10.1 Summary of HVDC controls 32/30
  32.10.2 Pole controls 32/31
  32.10.3 The phase-locked oscillator control system 32/31
  32.10.4 Tap-changer controls 32/32
  32.10.5 Master control 32/32
  32.10.6 Telecommunication 32/32
  32.10.7 Performance examples 32/33

32.11 A.c. system damping controls 32/34
  32.11.1 D.c. link supplies power from dedicated generators or from a very strong system to a small system 32/34
  32.11.2 D.c. link connecting two systems which are not synchronised but are of similar size 32/35
  32.11.3 D.c. link connecting two parts of an a.c. system or two separate systems having also a parallel a.c. link 32/35

32.12 Interaction between a.c. and d.c. systems 32/35
  32.12.1 Study of HVDC systems 32/35
  32.12.2 A.c./d.c. system strength 32/36
  32.12.3 Short-circuit ratios 32/36
  32.12.4 Voltage/power curve 32/37
  32.12.5 Maximum-power curve 32/38
  32.12.6 Maximum available power 32/38
  32.12.7 Classification of the a.c./d.c. system strength 32/38
  32.12.8 Critical short-circuit ratios 32/40
  32.12.9 Short-circuit ratio as a guide to system planning 32/40
  32.12.10 ‘Island’ receiving system 32/41
  32.12.11 System interaction when the a.c. system impedance is high relative to d.c. power in-feed (low short-circuit ratio) 32/41

32.13 Multiterminal HVDC systems 32/42
  32.13.1 Series connection 32/42
  32.13.2 Parallel connection 32/43
  32.13.3 D.c. circuit-breakers 32/43

32.14 Future trends 32/44

32.1 Introduction

The first commercial generators were direct current (d.c.) and therefore so were the early distribution systems. As distribution was at relatively low voltages, transmission distances were by necessity very short. The potential benefits of electrical energy were fully recognised and work to improve existing transmission systems was undertaken both in Europe and in the USA.

In 1883, Nikola Tesla was granted patents for the inventions on which he had worked during the previous 10 years relating to polyphase alternating current (a.c.) systems. In May of that year he delivered his classic lecture to the American Institute of Electrical Engineers: ‘A New System of Alternating Current Motors and Transformers’. Although, today we cannot visualise life without a.c. electrical systems, they were not immediately or universally accepted. Edison, who was working on a comprehensive d.c. distribution system, wrote in 1889 in the *Scientific American*: ‘My personal desire would be to prohibit entirely the use of alternating currents. They are as unnecessary as they are dangerous. I can therefore see no justification for the introduction of a system which has no element of permanency and every element of danger to life and property’.

Elimination of commutators made generators simpler and transformers allowed voltage to be changed easily; the use of higher voltages became practical and transmission over longer distances feasible. Widespread use of a.c. generation and transmission followed the exploitation of the Niagara Falls energy in 1895. Yet engineers continued to seek means of transmitting d.c. at high voltages, because they realised that the cost of overhead lines and cables for high voltage direct current (h.v.d.c.) could be considerably lower than for a.c. at the same power. An a.c. line has three conductors each insulated for the crest value of the alternating phase voltage, but the power transmitted is related to the r.m.s. value; the design of a.c. lines has also to take into account the flow of reactive current. HVDC transmission lines require only two conductors and the normal working voltage equals the rated voltage of the line. However, it was necessary to develop an adequate a.c./d.c./a.c. converter in order to benefit from lower cost of d.c. lines and cables in an a.c. system environment.

The first commercial h.v.d.c. scheme connecting two a.c. systems was a submarine cable link between the Swedish mainland and the island of Gotland. The scheme, rated for 20 MW at 100 kV was commissioned in 1953. Mercury arc valves each rated at 50 kV, 200 A were used as the converting device. Eleven mercury arc valve schemes totalling 6400 MW have since been commissioned. The last scheme to use mercury arc valves was Nelson River Bipole 1 in Canada, rated for 1620/1800 MW at ±450 kV. It uses the world’s largest mercury arc valves, made in the UK, each rated at 155 kV, 1800/2000 A (*Figure 32.1*).

In the late 1960s experiments using thyristor valves in mercury arc schemes were carried out in Sweden and in England. In 1970 the Gotland scheme was uprated to 30 MW at 150 kV by the addition of a 50 kV, 200 A thyristor valve bridge. In 1972 the first thyristor scheme, 320 MW back-to-back, was commissioned at Eel River in New Brunswick, Canada. At present there are some 70 schemes totalling over 60 000 MW of installed capacity in service or under construction. The recent increase in the utilisation of h.v.d.c. can be explained partly by its technical attributes and partly by the advantages gained from the interconnection of power systems which it facilitates.

The largest long distance overhead line transmission h.v.d.c. system is the 6300 MW Itaipu scheme in Brazil consisting of two bipoles each rated at ±600 kV. The largest h.v.d.c. submarine cable scheme is the 2000 MW link

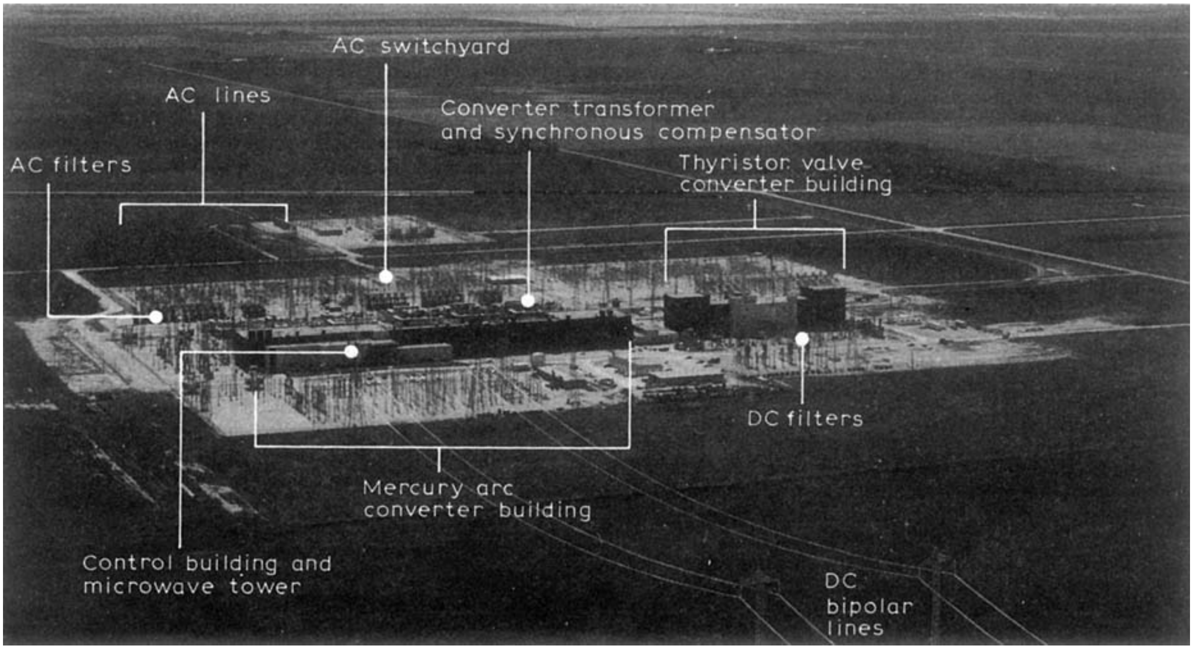

AC switchyard
AC lines
Converter transformer and synchronous compensator
AC filters
Thyristor valve converter building
DC filters
Mercury arc converter building
Control building and microwave tower
DC bipolar lines

Figure 32.1 Dorsey Converter Station of the Nelson River h.v.d.c. scheme. The photograph shows valve halls and outdoor equipment of Bipole I rated for 1800 A at ±465 kV and Bipole II rated for 1800 A at ±500 kV. In an emergency the two bipoles can be paralleled on one bipolar line transmitting 3600 A at ±465 kV. (Courtesy of Manitobe Hydro)

between France and England, consisting of two bipoles, each rated for 1000 MW at ±270 kV. The largest zero distance or 'back-to-back' schemes are rated for 1000 MW, Chateauguay in Canada and Chandrapur in India.

32.2 Applications of HVDC

32.2.1 Introduction

The answer to the question 'Why h.v.d.c.?' was, historically, that h.v.d.c. lines and cables are cheaper than those for a.c. for the same power transmission capability and, provided the transmission distance is more than a critical value (the 'break-even distance'), the savings from using h.v.d.c. lines or cables would more than pay for the a.c. to d.c. converters. The length of submarine cable made a.c. impracticable for some schemes due to the large charging current.

Today, only a few h.v.d.c. links are justified by such simple economics. They are not relevant to back-to-back schemes in which the distance between adjacent a.c. systems is zero, and in most other transmission schemes other attributes of h.v.d.c., its asynchronous nature or its ability to control power, play an equally important part in the choice of transmission.

32.2.2 Types of d.c. interconnection

An h.v.d.c. link is itself asynchronous but it may connect two asynchronous or synchronous a.c. systems and can be further sub-classified by distance according to whether or not there is a h.v.d.c. line or cable between the two converter terminals, as follows.

Asynchronous: where h.v.d.c. is the only interconnection between two systems with different frequencies, e.g. 50 and 60 Hz, or two systems at nominally the same frequency but with uncontrolled phase relationships.

Synchronous: where h.v.d.c. link is used within an a.c. system or in parallel with an a.c. interconnection.

Long distance point-to-point interconnections by: (a) overhead line, (b) undersea cable, (c) underground cable, or (d) combination of overhead lines and cables.

Zero distance back-to-back interconnections.

32.2.3 Purposes of transmission interconnections

A.c. or d.c. interconnections can be classified as follows.

(1) Power transfer exclusively or largely in one direction, normally characteristic of point-to-point applications:
  (a) from remote hydro, thermal or nuclear generation to load areas; or
  (b) from a strong to a weak a.c. system.
(2) Power transfer in either direction, normally characteristic of interconnections between neighbouring a.c. systems, typically of similar strength. Such system interconnections, where a.c. or d.c., offer one or more of the following benefits:
  (a) include the link's capacity in the spinning reserve for each system, minimising the generating capacity allocated in each system for such duties,
  (b) increase the security of supply by offering mutual support,
  (c) take advantage of seasonal generation and load pattern differences between the two systems,
  (d) take advantage of timing differences between daily load peaks of the two systems, and
  (e) take advantage of different types of generating plant with different base to peak cost ratios in the two systems.

32.2.4 Reasons for choosing HVDC

Of the many reasons which may contribute to the choice of the h.v.d.c. as a means of interconnecting two power systems (or elements of power systems), two stand out as being the most important, namely:

(1) Frequency or phase angle variation between the two terminals of the interconnection may render an a.c. link impractical. In an extreme case, the a.c. bus-bars at the terminals of the link may operate at different frequencies. Even if they are synchronised, it does not follow that reliable a.c. transmission can be established, because variations in relative phase angle between the two bus-bars, caused either by variation in load or by network disturbances, may result in unacceptable power flow severe enough to cause frequent tripping. Thus, it may prove economic to use h.v.d.c. for a zero length (back-to-back) transmission, or in parallel with an existing a.c. transmission path.
(2) The transmission distance may be so long that the cost savings arising from the use of relatively cheaper h.v.d.c. conductor systems is more than sufficient to outweigh the costs of the extra terminal equipment required for h.v.d.c.

Combinations of these two factors constitute more powerful economic pressures than either by itself.

Benefits which h.v.d.c. may provide beyond those provided by an a.c. interconnection can be summarised as follows:

(i) Provide the facility to interconnect two systems which have different operational procedures for frequency or voltage control.
(ii) Provide predetermined and controlled power transfer. Power flow in an a.c. interconnection is controlled by phase relationships which, being relatively uncontrolled, can cause inadvertent overloading or underutilisation during normal or disturbed operation. In the case of h.v.d.c., two utilities can pre-set the limits of power by which at any time they can assist each other and power will change automatically up to those limits in response to predetermined conditions, such as a frequency change.
(iii) Improve transient stability of the interconnected systems by modulating synchronising or damping power to reduce intermachine swings.
(iv) Avoid excitation of subsynchronous resonance as might occur in the case of series capacitor applications in an equivalent a.c. interconnection.
(v) Distribute the available power more effectively and thus delay the introduction of new power stations and major transmission reinforcements.
(vi) Permit staged development of a country's overall power system in a more controlled and hence less expensive manner by providing the means to utilise generation in geographically separate systems, compared to what could be done by a purely a.c. transmission development.
(vii) HVDC does not contribute to the a.c. system fault current. The contribution to the system fault current

by an a.c. interconnection may necessitate the replacement of the existing switchgear.

Therefore, in addition to the use of h.v.d.c. to connect two systems which cannot be synchronised, it should also be considered as one of the possible alternatives whenever enhancements are needed to make an a.c. interconnection attractive: the use of series compensation, variable shunt reactive compensation, phase shift boosters, etc.

32.2.5 Application of HVDC to developing systems

In developing countries or regions experiencing load growth the integration of h.v.d.c. should be considered at every stage of system development, not merely as an appendage to an otherwise fully designed system.

Two kinds of interconnection are often required: long distance links for bulk power transfer from remote generation, and shorter links (perhaps even of the back-to-back type) to interconnect adjoining relatively large regional systems. For both kinds of interconnection, but particularly in the latter case, to provide a sufficiently secure link using a.c. may require a large installation, typically of multiple e.h.v. lines, which may not be justifiable economically until a much later stage of load growth. The immunity of an h.v.d.c. link from problems arising from variations in the relative phase of the two networks may permit the benefits of interconnection to be realised economically at a much earlier stage of system and load growth than that at which a.c. becomes justifiable. If at some future date the natural system growth justifies an e.h.v. or u.h.v. overlay of a.c. transmission lines, the d.c. interconnection would readily become an integrated part of the combined a.c./d.c. system and, by virtue of its rapid controllability, improve the overall system stability and dynamic performance. Thus the economic and technical advantages of both d.c. and a.c. interconnections can contribute both to the intermediate and to the long-term transmission planning.

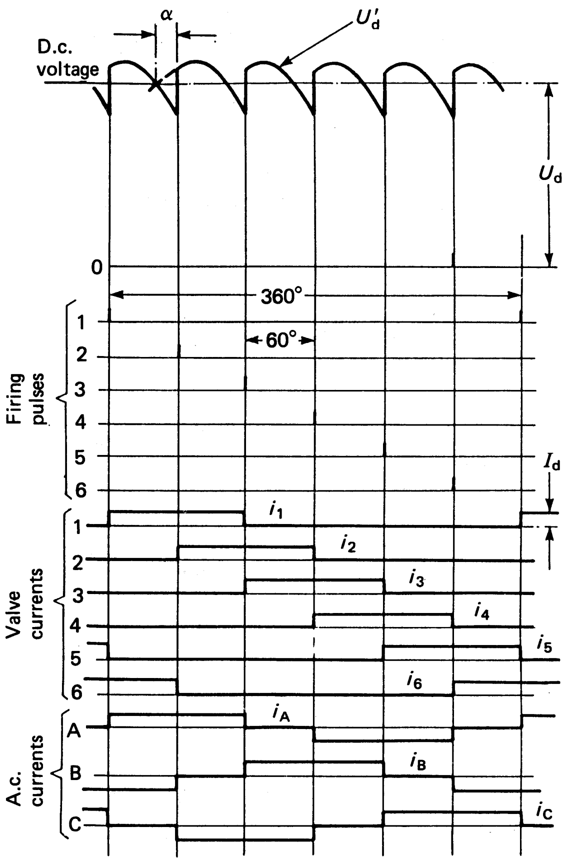


Figure 32.3 Idealised waveform for a six-pulse converter, neglecting commutation inductance

32.3 Principles of HVDC converters

32.3.1 Converter operation: simplified case of zero commutating inductance

The standard ‘building-block’ for h.v.d.c. converters is the three-phase full-wave bridge using six controlled (thyristor) valves, as shown in *Figure 32.2*. This is known as a ‘six-pulse’ converter group or bridge, because there are six valve firing pulses, and six pulses per power frequency cycle in the output, at the d.c. terminals. *Figure 32.3* shows the ‘idealised’ current and voltage waveforms, neglecting commutation inductance L in *Figure 32.2* and assuming acceptably smooth direct current output I_d, achieved by the action of the relatively large d.c. smoothing reactor L_d. For this case valve current pulses are 120° long, and their flat-tops have a magnitude equal to I_d. The time at which uncontrolled (diode) valves would commence conduction is used as a reference, and the ‘firing delay angle’ is defined to be zero at this point on the wave. *Figure 32.3* is drawn for the case where the firing time for each valve is delayed by the ‘firing delay angle’ α relative to diode operation, i.e. $\alpha = 0°$.

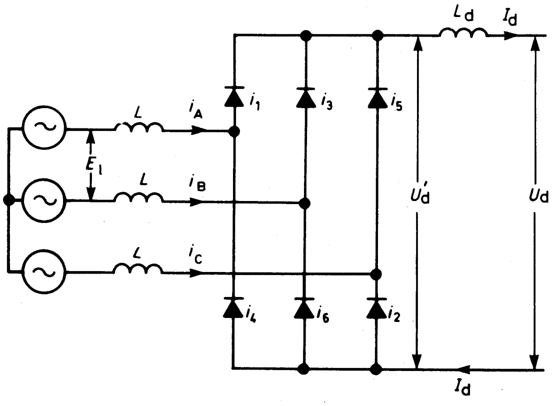


Figure 32.2 Basic six-pulse converter bridge

All conventional treatments of converter theory make the assumption that the e.m.f. E_l in *Figure 32.2* is sinusoidal; an assumption which is substantially true in practice because of the a.c. harmonic filters which are usually connected at the a.c. terminals of converter stations, preventing the non-sinusoidal converter current from appreciably disturbing the shape of the power frequency voltage. The analysis which follows neglects the effects of the current-dependent losses of converters. This is because they are both small and non-linear, making it unreasonably laborious to take them into account unless digital computers are used to carry out the calculations.

Some numerical relationships for this simplified case are

$$U_d = E_l(3\sqrt{2}/\pi)\cos\alpha = 1.35\ E_l\cos\alpha \qquad (32.1)$$

$$I_l = (\sqrt{2}/\sqrt{3})I_d = 0.816 I_d \qquad (32.2)$$

$$I = (\sqrt{6}/\pi)I_d = 0.780 I_d \qquad (32.3)$$

where U_d is the d.c. voltage of the six-pulse bridge, E_l is the commutation e.m.f. (r.m.s. line–line), I_d is the d.c. current, I_l is the r.m.s. a.c. current per phase, and I is the fundamental component of the a.c. current.

Equation (32.1) describes the principal control action of a converter, i.e. by change of firing angle α the d.c. voltage can be changed from maximum positive (rectification) at $\alpha = 0°$ through zero at $\alpha = 90°$ to negative (inversion) for α approaching 180°. The d.c. voltage U_d, at $\alpha = 0$ and $I_d = 0$ is termed ideal no-load voltage, U_{dio}; this is a fictitious quantity but it is often used as the basis for further calculations.

$$U_{dio} = 1.35E_l \qquad (32.4)$$

In practice the a.c. connection is via a transformer (not shown in *Figure 32.2*). The transformer rating is defined as $E_l I_l \sqrt{3}$. Although the choice of this definition is arbitrary from the viewpoint of converter operation, it offers the convenience that it would be equally applicable if the transformer were to be utilised for a.c. transmission. With the combined simplifications of zero commutation (leakage) reactance and assuming diode operation ($\alpha = 0$) from equations (32.1) and (32.2) this exhibits its minimum value of 1.047 times d.c. power.

32.3.2 Converter operation: practical case of finite commutating inductance

In a six-pulse bridge circuit (*Figure 32.2*) the valves 1, 3 and 5 commutate the outgoing direct current I_d between themselves, while the valves 2, 4 and 6 commutate the incoming direct current I_d; the two three-pulse conversion processes form the six-pulse bridge conversion. For clarity the *Figure 32.4* is drawn for one-half of the six-pulse bridge, i.e. the commutations between valves 1, 3 and 5 only are shown.

In practice, the converter transformer will have a finite leakage inductance (L in *Figure 32.2*). This causes current waveforms to exhibit more gradual transitions as shown in *Figure 32.4*, i.e. the current requires a finite time to commutate from one valve to the next valve in sequence in that particular row of three valves of the 6-pulse bridge. This is known as commutation overlap time, usually expressed as an angle u in electrical degrees. The value of u increases with increasing d.c. current, reaching typically 25° at rated current.

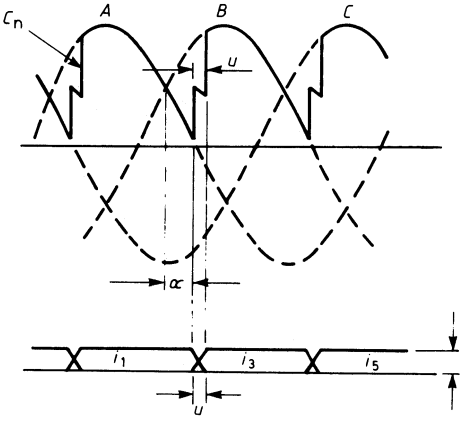


Figure 32.4 Rectifier operation with finite commutating reactance

The d.c. voltage U_d is reduced by the value dx due to the commutation notch C_n on *Figure 32.4*. (The derivation of the equations used is well documented in text books given as the general references at the end of the chapter.)

Equation (32.1) becomes

$$U_d = 1.35E_l\cos\alpha - \mathrm{d}x \qquad (32.5)$$

where

$$\mathrm{d}x = \frac{3}{\pi} I_d X_c \qquad (32.6)$$

where X_c is the commutating reactance. Usually the commutating (i.e. converter transformer) reactance is expressed in per unit of the converter transformer rating. The equation (32.5) becomes

$$U_d = 1.35E\left(\cos\alpha - 0.5\frac{I_d}{I_{d1}}x_c\right)\Bigg] \qquad (32.7)$$

or using equation (32.4)

$$U_d = U_{dio}\left(\cos\alpha - 0.5\frac{I_d}{I_{d1}}x_c\right)\Bigg] \qquad (32.8)$$

where I_{d1} is rated direct current.

32.3.3 Converter operation: converter acting as an inverter

This occurs when the firing angle α exceeds 90°. If current flow is to continue, this can only occur as a result of an external power source supporting the direct voltage. An inverter connected to an external circuit composed only of passive components does not conduct, being essentially a provider of back-e.m.f., to be overcome by the d.c. line voltage. The waveforms are generally similar to those above, but d.c. voltage U_d is negative. Thus to reverse power flow in a converter, although d.c. current cannot be reversed, d.c. voltage can be reversed by control action.

Figure 32.5 shows the inversion process for valves 1, 2 and 3 of *Figure 32.2*. When valve 3 fires, its current rises to I_d, and valve 1 current falls to zero, in a time u degrees, similarly as for rectifier operation. However, the current is now flowing due to d.c. line voltage and against the (negative) inverter transformer voltage, which acts as 'back-e.m.f.'. The commutation process must be completed before phase A voltage becomes more positive than phase B voltage, point D on *Figure 32.5*. If valve 1 is still conducting at that point it will continue to conduct driven by the sum of d.c. line and the phase A voltages. The inverter is operated so that the commutation process is completed well before point D. The quantity γ is the 'extinction angle' and is the time available for the valve to turn off, i.e. become capable of withstanding the subsequent forward voltage. Valve performance is discussed in Section 32.7.

Small working values of γ (i.e. values of α approaching 180°) lead to low capital cost of valves and transformers,

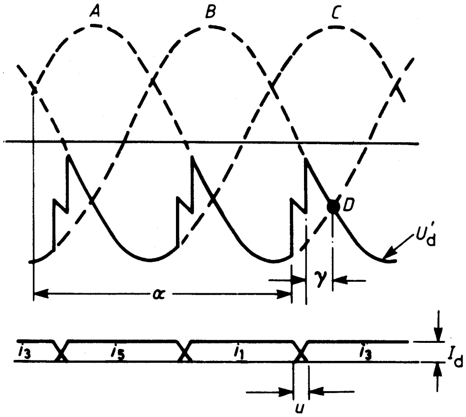


Figure 32.5 Inverter operation

low harmonic generation, low reactive power consumption and low station losses. However, too small a value of γ causes commutation failure. This is usually initiated by disturbances arriving from the a.c. system which distort the waveform at the a.c. terminals of the converter station, resulting in temporarily reduced γ for one or more commutations. A reduction of γ to less than 10° (12°) at 50 Hz (60 Hz) is usually needed before commutation failure becomes likely, but once it has occurred, it may temporarily collapse inverter operation, requiring 100 ms or so before the control system succeeds in restoring normal operation.

A typical running value of γ, which gives reasonable immunity to commutation failure, is 15° to 18° (for a 50 Hz system). The corresponding value of α to produce this is typically about 140°. It is important that inverters are provided with constant-extinction angle (γ) control to prevent commutation failures in normal steady-state operation, as discussed in Section 32.10.

For constant γ operation, equation (32.8) becomes

$$U_d = U_{dio}\left(\cos\gamma - 0.5\frac{I_d}{I_{dl}}x_c\right)\Bigg] \qquad (32.9)$$

32.3.4 Twelve-pulse converters

The harmonics produced by a six-pulse converter are large, requiring expensive filters. They can be reduced by use of a 12-pulse converter as discussed in Section 32.8. The usual arrangement of this for h.v.d.c. uses two six-pulse bridges connected in series on the d.c. side, with their transformers respectively star–star and star–delta, connected in parallel to the a.c. bus-bar. As the cancellation of harmonics takes place at the a.c. side of the converter/transformers, the conversion process takes place independently in each six-pulse bridge.

32.3.5 Basic d.c. voltage/d.c. current characteristics

Figure 32.6 shows these for a converter operating on a zero-impedance a.c. system. Natural boundaries to this occur at zero d.c. current (because d.c. current cannot reverse) and at α = 0° (firing cannot occur for α negative because this would mean attempting to fire valves when their anode–cathode voltage is negative). Other boundaries are applied by control action:

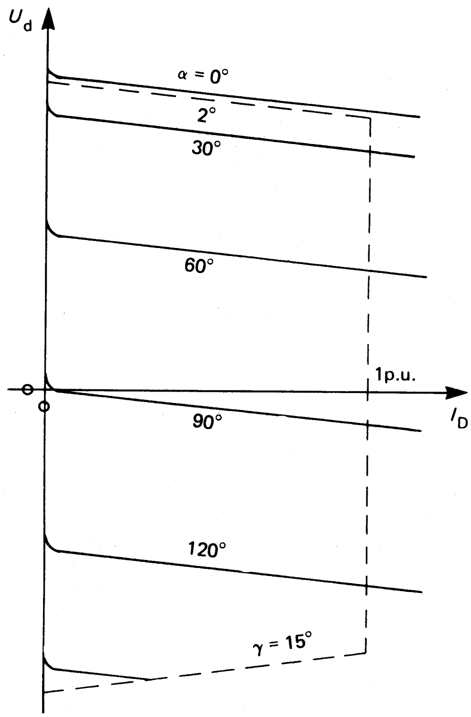


Figure 32.6 Basic firing angle control characteristics (a.c. voltage constant)

(1) a minimum limit of α = 2° is applied in practice to ensure reliable firing of each valve;
(2) a minimum γ limit (say at γ_l = 15°) prevents commutation failure in normal operation as described above; or
(3) a d.c. current limit is applied at the thermal current limit of valves and other components.

Within these boundaries, any desired shape of d.c. voltage/d.c. current characteristics can be obtained by control action, i.e. by change of α, as described later.

32.3.6 Basic principles of control of HVDC transmission

Figure 32.7 shows a simplified diagram for a two-terminal h.v.d.c. link, with elementary controls.

At the rectifier a closed-loop current control is provided, which adjusts firing angle α in response to the difference between measured d.c. current I_d, measured by means of a d.c. current transformer, and a current order signal I_o, assumed fixed for the present.

At the inverter, closed-loop γ control is provided, operating similarly but from measured γ, with a fixed reference demanding γ_l of typically 15° to 18°. A current control loop is also provided, similar to that at the rectifier, supplied with the same current order, but with a ‘current margin’ signal I_m subtracted from it. I_m is typically 0.1 of rated d.c. current, I_{dl}.

Figure 32.8 shows the resulting d.c. voltage/d.c current characteristics. The rectifier current loop generates the constant-current characteristic BCD. This has a natural transition at B to the α = 0° line AB. The inverter has a constant-γ characteristic FCE, with a transition at F to a constant-current characteristic FG at I_m below BCD.

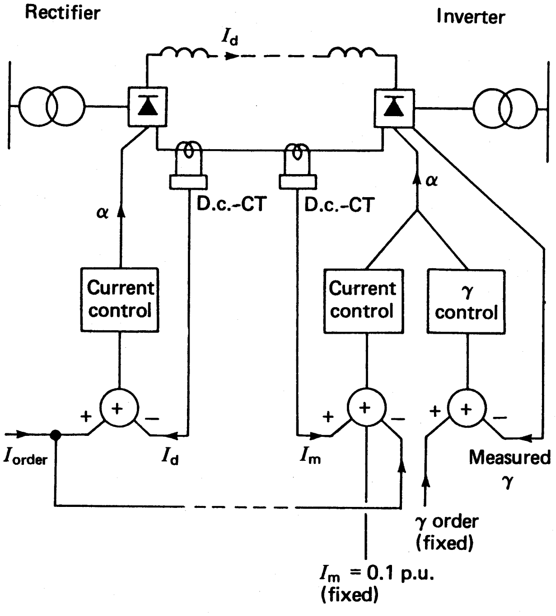


Figure 32.7 Elementary controls for a two-terminal h.d.v.c. link

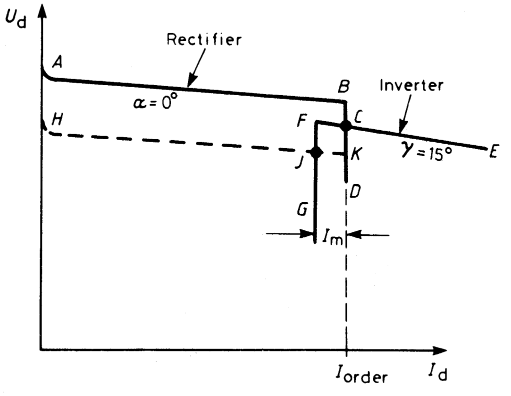


Figure 32.8 D.c. voltage/d.c. current characteristics for *Figure 32.7*

D.c. line resistance may be included with either characteristic for constructing the U_d/I_d diagram. The steady-state working point is at the cross-over, i.e. point C.

Thus in normal operation the rectifier controls current and the inverter controls voltage.

Tapchangers on each converter transformer are often used. These do not have any major control functions; their duty is to optimise working conditions for each converter. The inverter tapchanger is usually arranged to effectively move FCE up or down to obtain rated d.c. voltage; whereas the rectifier tapchanger adjusts AB up or down so that measured α lies within a range of about 5° to 20°.

Changes of a.c. voltage experienced by the converters are effectively corrected in the long term (many seconds) by the tapchangers, but can temporarily shift the characteristics. The only important case is that in which the rectifier a.c. voltage falls significantly (or inverter a.c. voltage rises), so that for example AB moves down to HJK. In the absence of the current loop at the inverter this would cause a complete loss of d.c. current. When this loop exists, the working point moves only to point J, at a d.c. current lower than I_{dl} by the current margin I_m. (The change of working point from C to J, even for a slow a.c. voltage change, would be too abrupt as shown in *Figure 32.8*. More sophisticated characteristics are described in Section 32.10.)

32.3.7 Starting and stopping an HVDC link

32.3.7.1 Starting

HVDC converters can be started and stopped very rapidly if required. However, in normal operation this is done relatively slowly to avoid shocks to the a.c. systems.

The normal starting procedure is to first de-block (i.e. initiate firing pulses) at the inverter, with a firing angle of about 160°; as the d.c. voltage is zero, this causes no current.

The rectifier is then de-blocked, initially at a similar firing angle, which is then slowly reduced over a few hundred milliseconds, raising d.c. voltage until the inverter current rises and the system settles at normal firing angles, with a low current order (0.1 per-unit (p.u.) or less). Current (or power) is then increased slowly over, say, 10 seconds to 10 minutes to the desired final value.

32.3.7.2 Stopping

Whilst in a.c. practice a circuit is invariably taken out of service by opening a switch, on the d.c. side of a converter station the technique for shutting down a two-terminal scheme is normally to reduce power by control action over a period to suit the needs of the a.c. system, and then to block all valves.

When a converter is shut down, a bypass path is often provided on the d.c. side. For example, normal firing pulses may be blocked and a pair of series connected valves fired to provide a bypass path. This collapses the d.c. voltage and the change in voltage can be detected at the other station and used to initiate its shut down sequence.

If converters at both ends of a link are bypassed whilst high current is flowing in the link, because the resistance of the line circuit is low and the inductance of the d.c. reactors is large, current may take a long time to decay. The d.c. line can be discharged faster by ensuring that one or both of the converter stations remains in inverter mode until current has stopped.

In an emergency, stopping is achieved much more rapidly. A typical method is to separate fault signals according to their origin, into non-urgent stop (from relays detecting persistent commutation failure, asymmetry or misfire, undervoltage, abnormal firing angle, etc.) or emergency stop (from relays detecting overcurrent, or flash-overs from differential measurement). Non-urgent stop signals are usually allowed to persist for about 300 ms, and at a rectifier cause forced-retard, i.e. firing angle is forced into inversion at, say, 150°, which will normally stop d.c. current in about 10 ms, at an inverter bypass operation is caused by blocking normal firing pulses and instead firing a pair of series connected valves in each bridge. The latter does not directly stop d.c. current, but causes zero d.c. voltage, which is detected by the rectifier which then stops current by forced retard after 300 ms.

For an emergency stop signal, full blocking, i.e. suppression of all firing signals is applied within about 2 ms, and the converter group circuit breaker is tripped. Zero a.c. side current usually occurs in less than 10 ms, except for some types of flashover within the station, which the circuit breakers will clear in, say, three cycles.

32.3.8 Power reversal

As described earlier, in h.v.d.c. schemes, due to the unidirectional property of the thyristors, the current cannot reverse. Power reversal is achieved by reversing d.c. voltage. This is carried out by changing over the operation characteristics of the converters in the two stations. The effect of this is to change the phase angle of the current at the a.c. bus-bars by reversing its power component; its reactive component always remains negative.

As for start and stop, power reversal can be done rapidly if desired, within typically 200 ms. However, reversal is normally carried out by reducing power order slowly to zero or a low value, and then reversing the power flow at low current. The power order is then increased to the new desired value.

32.3.9 Isolating a valve group

In some schemes, several valve groups at a converter station are connected in series and it may be desirable to be able to block or deblock, or isolate, one group whilst another remains in service. In such cases it is necessary to provide a bypass path for the current and means to transfer the current between the bypass path and the path through the converters.

The ultimate bypass path is normally a metallic switch in a substantially conventional form, and transfer of current from the converter valves to the bypass switch presents no difficulty as the volt drop in conducting valves will be greater than in the switch. The converter can first either be controlled just into the inverter region, or put into its bypass mode. When the contacts of the switch close, current will transfer and valve firing pulses can be blocked. The valve group may then be isolated if desired.

To transfer current from the bypass switch to the converter valves a special technique is necessary to extinguish the current in the switch. One alternative is to provide a means to increase the voltage drop in the switch to exceed that in the valve bypass path, to force current to transfer.

32.3.10 Numerical example

Calculations performed for the purpose of equipment steady state ratings are normally carried out by means of computer programs using equations similar to those described previously but taking into account the losses of the converters. For nominal conditions, simplified equations can provide reasonably accurate results. The following takes as an example a 1500 MW bipolar h.v.d.c. scheme with a rated voltage of ±500 kV at the rectifier d.c. line terminals. The nominal d.c. resistance of the overhead line is 10.0 ohm. The d.c. voltage (U_{di}) at the inverter d.c. line terminal, for rated d.c. current $I_{\mathrm{d}} = 1500\,\mathrm{A}$, is

$$U_{\mathrm{di}} = 500 - 0.5 \times 1500 \times 10 = 492.5\,\mathrm{kV/pole}$$

and the bipolar powers at the rectifier and inverter d.c. line terminals are

$$P_{\mathrm{dr}} = 2 \times U_{\mathrm{dr}} \times I_{\mathrm{d}} = 2 \times 500 \times 1500 = 1500\,\mathrm{MW}$$
$$P_{\mathrm{di}} = 2 \times U_{\mathrm{di}} \times I_{\mathrm{d}} = 2 \times 492.5 \times 1500 = 1477.5\,\mathrm{MW}$$

Each pole consists of two series connected six-pulse valve groups.

Therefore, U_{dr} and U_{di} of each six-pulse bridge is

$$\tfrac{1}{2}U_{\mathrm{dr}}\ \text{pole} = 500/2 = 250\,\mathrm{kV\ d.c.}$$
$$\tfrac{1}{2}U_{\mathrm{di}}\ \text{pole} = 492.5/2 = 246.52\,\mathrm{kV\ d.c.}$$

To limit the fault surge current in the thyristor valves to an acceptable level a transformer reactance of 15% is specified. Rated α is specified as 12° and rated γ as 15°.

The value winding e.m.f. required at the nominal operating point can be calculated using equation (32.7) for the rectifier and equation (32.9) for the converter.

$$E_{\mathrm{lr}} = \frac{U_{\mathrm{dr}}/2}{1.35 \times \left(\cos\alpha - 0.5\dfrac{I_{\mathrm{d}}}{I_{\mathrm{dl}}}x_{\mathrm{cr}}\right)}$$
$$= \frac{250}{1.35 \times (\cos 12^{\circ} - 0.5 \times 0.15)} = 205\ \mathrm{kV\ r.m.s.}$$
$$E_{\mathrm{li}} = \frac{U_{\mathrm{di}}/2}{1.35 \times \left(\cos\gamma - 0.5\dfrac{I_{\mathrm{d}}}{I_{\mathrm{dl}}}x_{\mathrm{ci}}\right)}$$
$$= \frac{246.25}{1.35 \times (\cos 15 - 0.5 \times 0.15)} = 204.7\,\mathrm{kV\ r.m.s.}$$

The transformer rating can be calculated as

$$\text{Rectifier: } \sqrt{2} \times E_{\mathrm{lr}} \times I_{\mathrm{dl}} = 1.41 \times 205 \times 1.5 = 433.6\,\mathrm{MVA}$$

$$\text{Incerter: } \sqrt{2} \times \mathrm{E_{li}} \times \mathrm{I_{dl}} = 1.41 \times 204.7 \times 1.5 = 432.9\,\mathrm{MVA}$$

The reactive power absorbed by the six-pulse group can be calculated as

$$Q_{\mathrm{r}} = U_{\mathrm{dr}} \times I_{\mathrm{dr}}\sqrt{\left(\frac{1.35E_{\mathrm{lr}}}{U_{\mathrm{dr}}}\right)^2 - 1}$$
$$= 250 \times 1.5 \times \sqrt{\left(\frac{1.35 \times 205}{250}\right)^2 - 1}$$
$$= 178\,\text{MV-Ar for the rectifier}$$

and

$$Q_{\mathrm{i}} = U_{\mathrm{di}} \times I_{\mathrm{di}}\sqrt{\left(\frac{1.35E_{\mathrm{li}}}{U_{\mathrm{di}}}\right)^2 - 1}$$
$$= 246.25 \times 1.5\sqrt{\left(\frac{1.35 \times 204.1}{246.25}\right)^2 - 1}$$
$$= 188\,\text{MVAr for the inverter}$$

32.4 Transmission arrangements

32.4.1 Bipolar lines

The bipole is the most commonly used arrangement. It consists of one line at positive potential with respect to

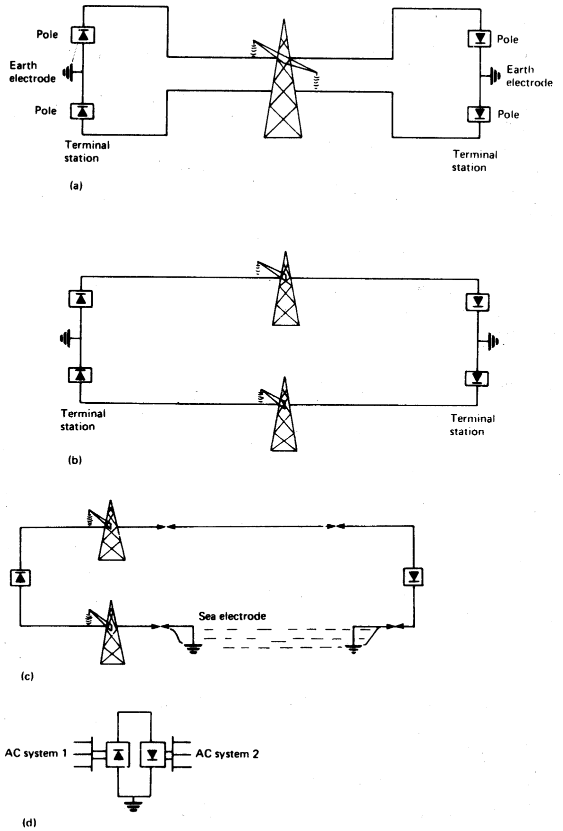


Figure 32.9 Transmission arrangements: (a) bipolar transmission line; (b) two monopolar transmission lines; (c) cable monopolar transmission with sea return; (d) back-to-back connection

earth and the other of negative potential, the neutral being solidly grounded in the converter station. *Figure 32.9*(*a*) indicates an overhead line, but the same arrangement is used for bipolar cable schemes.

Control is so arranged that during normal operation the currents in the two poles are balanced so that the current flowing out from the positive pole returns via the negative line, and the current flowing in the earth is negligible. For a fault on one pole controls will reduce direct current and voltage of the affected pole to zero in an attempt to clear the fault. In the meantime transmission by the unfaulted pole continues using earth as the temporary return path. Clearly this technique offers the prospect of increased reliability of h.v.d.c. transmission compares with three-phase a.c. transmission, in which phase faults may cause loss of the complete circuit.

32.4.2 Two monopolar lines

In circumstances where the probability of line failure arising from environmental conditions is high, two monopolar towers (*Figure 32.9*(*b*)) have been used on separate rights-of-way although the system operates as a bipole.

A monopolar line can be operated using earth return with the connection to earth made via a ground electrode. However, ground electrodes have so far only been constructed for use in emergency conditions, being designed to operate only for a matter of hours, or in some cases for a few months.

32.4.3 Cable schemes with sea return

Sea electrodes have been used successfully (*Figure 32.9*(*c*)) on several schemes for continuous monopolar operation, low resistivity sea water providing a permanent return path. The resistivity of sea water is in the order of 0.2 Ω-m compared with 10 Ω-m for earth at an ideal land site or 100 Ω-m for fresh water.

32.4.4 Back-to-back arrangement

If there is a need to connect two nearby systems by h.v.d.c., economies can be achieved by combining two converter stations in a back-to-back arrangement (*Figure 32.9*(*d*))

32.4.5 Ground electrodes

The earth provides a readily available medium for the return of direct current. While in certain countries permanent use of the ground in a power circuit is not permitted, its use under emergency conditions such as a line or terminal outage is accepted. A bipolar circuit with ground return thus provides two independent transmission paths.

A typical design of a ground electrode consists of a 3 m annular trench, 250–400 m in diameter, containing coke (*Figure 32.10*). A steel conductor embedded in the coke is connected to the electrode line by four or more radial insulated conductors. The coke acts as the electrode and the ring diameter is chosen so that the maximum voltage gradient at the ground surface does not exceed $(5 + 0.03\rho)$ V/m, where ρ is the surface material resistivity in ohm-metres. This is a safe value for humans or animals.

A ground electrode requires a damp site of low resistivity both at the surface and in the underlying strata. Very fine soils or sands may be unsuitable because thermal or electrical osmosis could remove the water from the soil–coke surface. A careful survey of neighbouring installations is necessary to identify any electrical conductor which by intercepting the equipotential lines near electrodes would carry a residual d.c. current. This current may cause corrosion where the (anodic) current enters, unless protected by sacrificial anodes or an applied reverse d.c. voltage. Pipelines, railway line, telephone and power cable sheaths, and power distribution systems using multiple grounding of

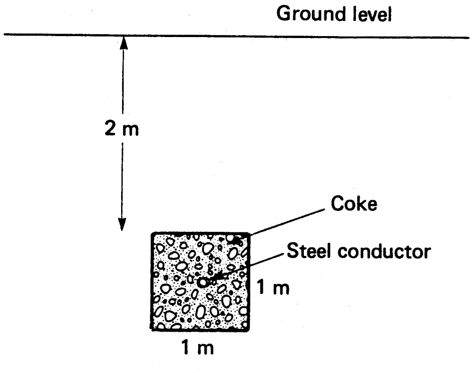


Figure 32.10 Typical ground electrode

neutrals may require additional corrosion protection or segmentation into insulated sections.

32.4.6 Sea electrodes

A sea electrode can be relatively simple. One design uses 24 concrete boxes each containing two 1.5 m high silicon cast iron electrodes.[1] The boxes prevent contact with marine life, protect electrodes from silt and damage and restrict the voltage gradient outside to 2.5 V/m. The resistance is $0.02\,\Omega$, well below that of the connecting cables and 35 km overhead line to the converter station ($1.1\,\Omega$). An alternative electrode material having good resistance to corrosion is platinised titanium.[2]

Electrodes can be built on shore, but assurance of contact with sea water is likely to require a pump installation and consequent maintenance. In both cases annual inspection of the anode electrode for deterioration is necessary. The cathode electrode consists of a simple single rod.

32.4.7 Staged construction of HVDC

HVDC systems can be built in stages to suit generation development. For example, the Nelson River Bipole I is made up of three six-pulse groups in series in each pole. Stage 1 was rated at 810 MW at +320 kV and −160 kV (*Figure 32.11*(*a*)) compared to final rating of ±460 kV (*Figure 32.11*(*b*)). The low initial voltage implies higher losses until the final voltage was reached. Similarly, Nelson River Bipole II was first used at ±250 kV and later at ±500 KV.

If the time between stages of generation development is very long, the cost of operation at low voltage may prove to be too high. In such a case it is possible to start with converters at full voltage but low current, subsequently adding another converter in parallel. The 'parallel' build-up is more expensive than the 'series' build-up from the converter station point of view, but it is more economic from the line loss point of view.

The Pacific Coast Inertie is an interesting example of both series and parallel extensions, illustrating the flexibility of h.v.d.c., *Figure 32.12*. The original scheme was rated for 1420 MW at ±400 kV using three mercury arc six-pulse groups in series per pole. The scheme proved to be capable of higher current rating, giving 1600 MW capability. By restringing parts of the h.v.d.c. line and by adding a 100 kV thyristor six-pulse bridge to each pole, it was possible to achieve a rating of 2000 MW at ±500 kV. One thyristor converter rated for 500 kV has now been added in parallel with each pole (labelled 'proposed extension II' on *Figure 32.12*) to increase the h.v.d.c. line current to 3100 A. This gives a total external rating of 3100 MW at ±500 kV, more than twice the original rating of 1420 MW.

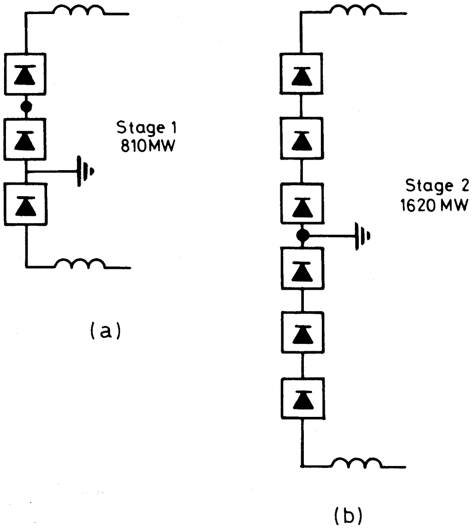


Figure 32.11 Development stages of Nelson River Bipole I scheme

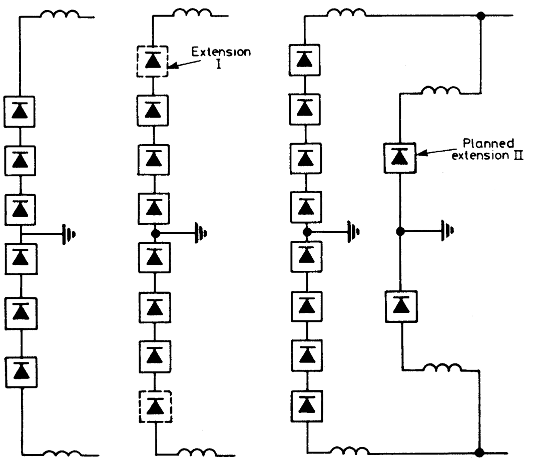


Figure 32.12 Development stages of Pacific Coast Inertie h.v.d.c. scheme

32.5 Converter station design

32.5.1 Valve group arrangements

The size and number of converter groups will generally be dictated by the firm power requirements and will in turn dictate the complexity of the d.c. switchgear requirements. Firm power requirements will also influence the selection of which equipment should be switched simultaneously or independently.

Figure 32.13(*a*) shows the main equipment of a typical converter station arrangement: converter transformers,

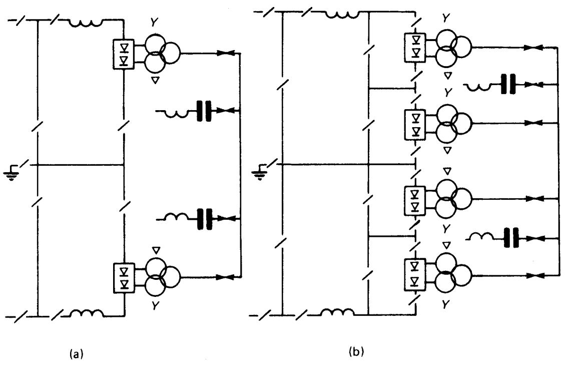


Figure 32.13 Alternative bipolar h.v.d.c. converter connections: (a) single 12-pulse groups; (b) series connected 12-pulse groups

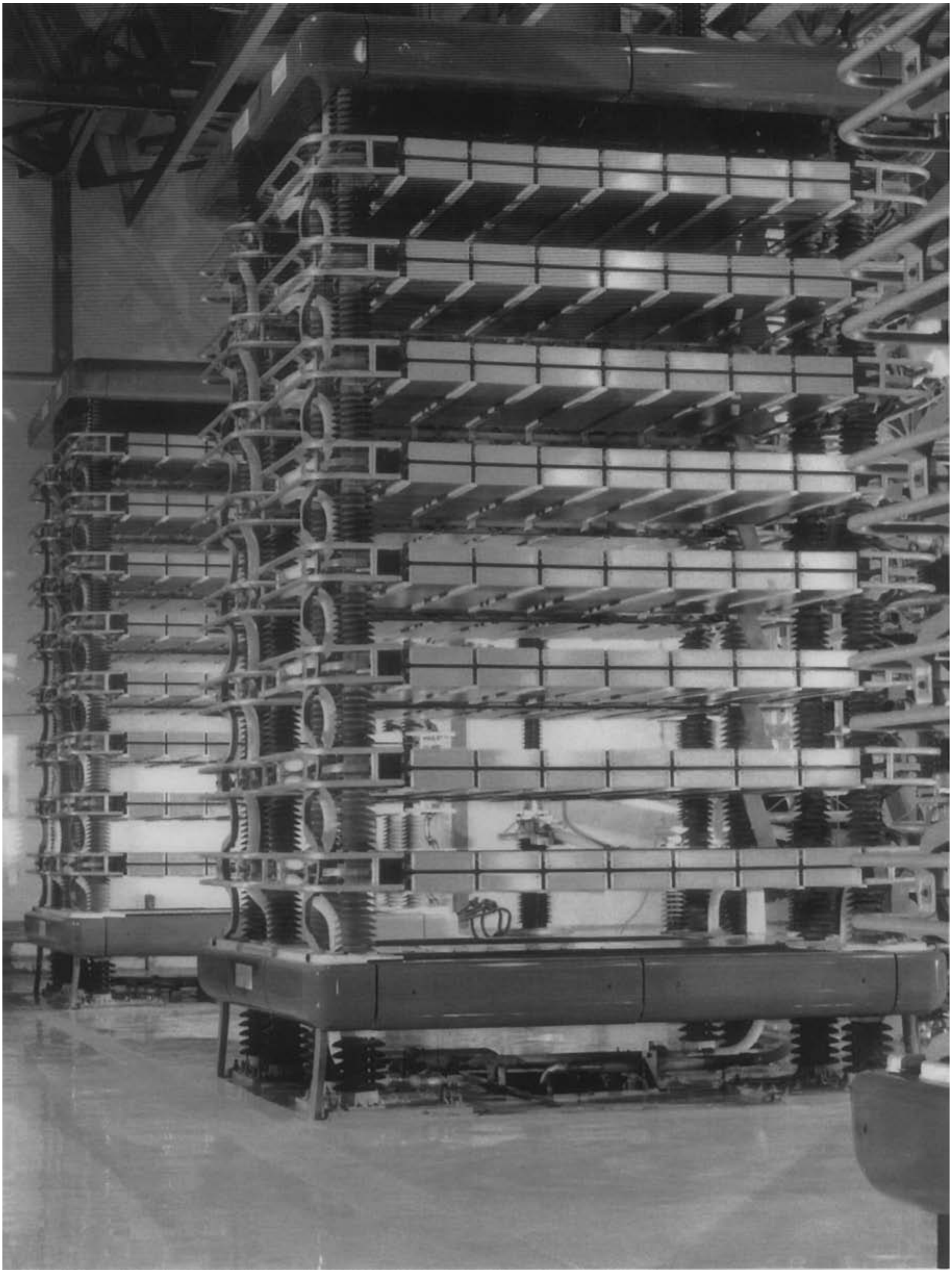


Figure 32.14 Air-cooled, air-insulated quadrivalves (1000 MW back-to-back scheme, Chandrapur, India)

a.c. filters, valves, d.c. reactors and associated isolators form two separate and independent poles of a bipole. Surge arresters are essential components and are dealt with in Section 32.6. D.c. filters may be added in overhead transmission line schemes.

In *Figure 32.13*(*a*), the loss of any major component, say a transformer, would lead to the loss of the 12-pulse pole. If power requirements are such that 75% of rated power must be firm then the arrangements in *Figure 32.13*(*b*) should be considered. Firm power of 75% can be achieved by dividing the pole into two 12-pulse groups. There is no penalty in the thyristor valve arrangement, as the same number of thyristors is required to withstand the desired voltage. However, failure of a d.c. reactor would constitute a loss of 50% power.

In mercury are schemes it was economic to use a six-pulse bridge as the operating unit. In thyristor schemes in most cases, it is more economic to use a 12-pulse converter group as the operating unit, and avoid the need for large filter, for fifth and seventh harmonic currents (*Figure 32.14*).

32.5.2 Converter valves

The valves are arranged in three-phase bridge circuit as discussed in Section 32.3. The rating of the thyristor valve is flexible and the operating voltage of the valve can be varied by choosing a different number of series connected thyristors. Thyristors connected in parallel have been used, but current ratings in excess of 5000 A d.c. bridge current can be accommodated by a single power thyristor of the type now available.

Physically the thyristor valves of a 12-pulse group are usually arranged in stacks with four valves mounted on top of each other to form a quadrivalve, as shown on *Figure 32.14*. This arrangement has the advantage of enabling insulation to ground for the valves operating at the highest voltage potential to be provided by the other valves in the stack (*Figure 32.16*). *Figure 32.15* shows a section through such a valve hall. For the electrical power circuit of the equipment shown refer to *Figure 32.21* in Section 32.6. An alternative design from floor supported valves is to hang them from the valve hall ceiling.

32.5.3 Converter transformers

The following transformer arrangements can be used to supply a 12-pulse converter group:

(1) *One* three-phase transformer having one line (primary) and two valve (secondary) windings;
(2) *two* three-phase transformers each having one line and one valve winding;
(3) *three* single-phase transformers each with one line and two valve windings; and
(4) *six* single-phase transformers each with one line and one valve winding.

Figure 32.17 shows a layout using transformer arrangement (1) in which transformer bushings protrude into the valve hall. This layout has the advantage of eliminating outdoor connections between the transformers and the valve hall which can be a source of radio interference.

Figure 32.16 shows a layout using transformer arrangement (3). It may not be economic to arrange for transformer bushings to protrude into the building as this would necessitate a very long valve hall. Gas insulated busbars could be considered for such a layout.

For very large ratings, arrangement (4) may be used for a 12-pulse group, requiring a much larger area to provide the six a.c. connections between transformers and valve hall.

Because the converter groups are connected in series on the d.c. side, the windings of the outer converter transformers will be biased at a d.c. potential with respect to earth, which is equal to the sum of the voltage of the inner groups. D.c. potential has a different distribution between oil and paper insulation from a.c. which has to be considered in the design and test. The valve winding line-to-line voltage has a sinusoidal waveshape, but the design must take into account the significant harmonic current content due to the converter action. A major factor in the choice of transformer reactance is the prospective fault current in the thyristor valves during worst case fault conditions. This may dictate the use of a transformer reactance greater than the most economic value for a given transformer design.

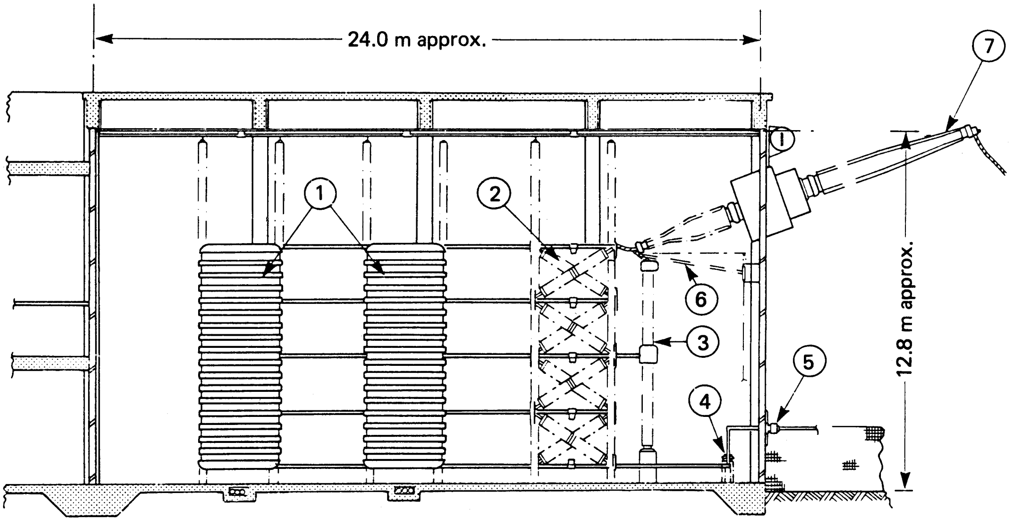


Figure 32.15 Section of valve hall: 1, thyristor quadrivalves; 2, surge arrester-valves; 3, surge arrester-valve group; 4, surge arrester-neutral; 5, through-wall bushing neutral d.c.; 6, earth switch; 7, through-wall bushing h.d.v.c.

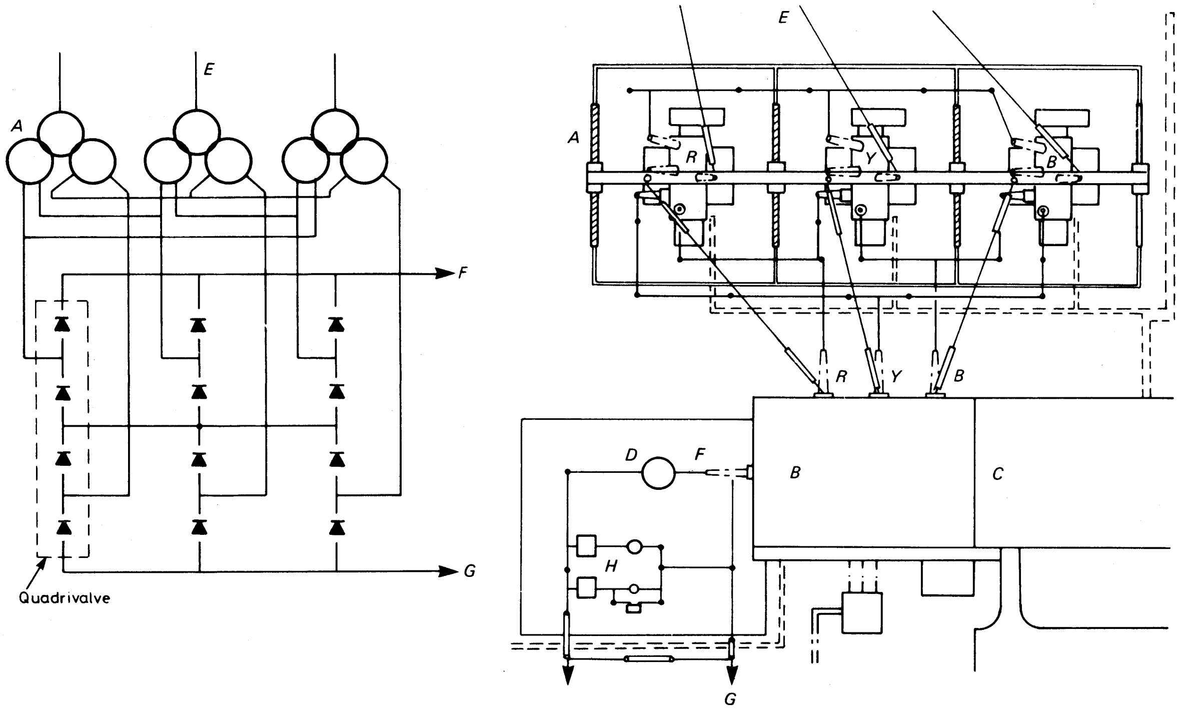


Figure 32.16 Single-phase three-winding transformer arrangement: A, converter transformers; B, valve hall; C, control building; D, d.c. smoothing reactor; E, a.c. connections; F, h.v.d.c. connection; G, neutral connection; H, d.c. filters

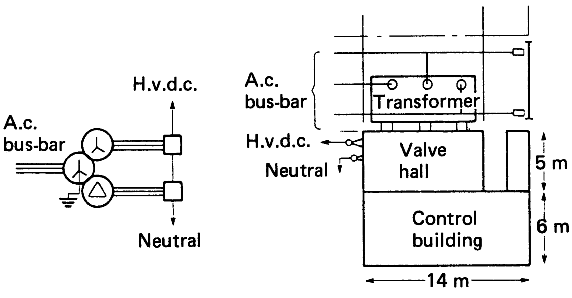


Figure 32.17 Three-phase three-winding transformer arrangement

32.5.4 A.c. filters

Filters, to absorb harmonic currents and to provide reactive power, are connected to the same a.c. bus-bar as the converter transformers. It will normally be necessary to split filters into several banks, both for separate maintenance (e.g. capacitor replacement) and to restrict the voltage step at switching. Filter design and reactive power are considered in Sections 32.8 and 32.9.

32.5.5 D.c. smoothing reactor

The converter valve groups are connected to the d.c. transmission system via a smoothing reactor. In deciding the inductance of this reactor several factors have to be considered. The reactor ensures that the overcurrent transient occurring during an inverter commutation failure or a d.c. line fault is kept within limits acceptable to the valves.

The smoothing reactor exhibits a very low resistance to direct current but provides a high impedance to the characteristic 12-pulse harmonic voltage resulting from converter operation. In the case of transmission schemes employing overhead lines the smoothing reactor acts to filter the harmonics appearing on the d.c. side of the converters in conjunction with shunt connected capacitors or filters. Unattenuated, these harmonics may cause telephone interference in the area surrounding the d.c. line.

Another important feature of the d.c. smoothing reactor, arising from its high impedance to high frequencies, is that it shields the remaining converter station equipment from being directly exposed to fast voltage transients which can occur on the d.c. line.

The reactor can be placed either in the h.v. or the l.v. connection. By placing the reactor in the l.v. connection savings can be made on reactor insulation costs and this arrangement is feasible for back-to-back schemes. The insulation to ground of the inner valve group would have to be increased however and for this reason, and because protection from fast voltage wavefronts is required for schemes with high voltage lines, it is usual to place the reactors in the h.v. end of transmission schemes.

In a back-to-back d.c. scheme large currents due to d.c. line faults and fast voltage transients which could occur on d.c. lines are not present. The first d.c. link at McNeill connecting Canadian eastern and western a.c. systems, which cannot be synchronised, was commissioned in October 1989. D.c. smoothing reactor is not used in this back-to-back scheme.[3]

There are some operational advantages when the d.c. reactor is not used. On the other hand, in some instances the generation of harmonics and the transfer of harmonics between the two a.c. systems could marginally increase. It should be noted that the commutating (mainly converter transformer) reactance provides an inherent reactance in the d.c. loop.

The d.c. smoothing reactor was considered an essential requirement for h.v.d.c. operation. It was possible to omit it in the case of McNeill station thanks to modern controls and analytical study techniques.

32.5.6 D.c. isolators

Most isolators in a two-terminal scheme will be of the conventional slow type. In the few cases where fast operation is required, high-speed isolators which do not have a d.c. current interruption capability will normally be sufficient to provide the switching of lines and valve groups while zero d.c. current is temporarily imposed by the converter action.

When series connected 12-pulse groups are used in a pole it is usually necessary to incorporate across each group high speed bypass switches, to assist the blocking and deblocking sequences, and to allow independent operation of the groups.

32.5.7 Protection

The protective functions required of a converter terminal can be divided broadly into three groups.

(1) *Conventional protection*: this group covers the standard forms of protection applied to transformers and reactors and would include differential, overcurrent, earth fault, Buchholz, etc.

(2) *Special power equipment protection*: this group covers special forms of protection which have been developed for converter plant. High-speed systems based on fibre optic coupling used with circuit breakers having two-cycle interruption time, can provide tripping in less than 50 ms. For the capacitor banks used in a.c. filters and static compensators, capacitor unbalance in protection is utilised to detect fuse operations and will have alarm, delayed shutdown, and immediate trip settings.

(3) *Protection control equipment*: special forms of protection for the converter equipment are incorporated as part of the electronic controls for the poles and valve groups. This protection will cover commutation failure, asymmetry, d.c. line or cable fault, d.c. undervoltage, etc. Transient occurrences do not cause shutdown, but if the condition persists for longer than say 300 ms shutdown would be initiated. Asymmetry protection would operate as a result of a converter valve misfire resulting in the generation of disturbed d.c. voltage waveforms. Again, this condition would cause shutdown if it persisted for more than 300 ms. Fault currents on the d.c. line or cable can be limited very quickly (within about 20 ms) by exerting control on the triggering of the valves. For cable schemes this action would be followed by shutdown, but for overhead line schemes, where recovery from the fault may be achieved by temporary reduction of the d.c. current to zero, one or more restarts can be attempted before shutdown is initiated.

These three groups of protection are co-ordinated where appropriate, and are used as back-up to each other to provide a comprehensive protection system. The functions of control and protection are increasingly being coordinated and carried out by microprocessors.

32.5.8 Converter station losses

The high cost of losses can appreciably influence the equipment design (thyristor size, and transformer copper size). Larger thyristors than required for the current rating of the schemer incur a lower current dependent loss. Also, by its ability to withstand a higher short-circuit current it permits a lower transformer reactance to be used.[4] This in turn favours lower transformer copper loss. Such a choice gives as by-products some overload capability and a reduction in converter var consumption. The specified filter performance and var control requirements also influence the converter station's losses. For all these reasons, losses for different schemes may vary greatly. The following loss figures are for a scheme having an average value of losses: at rated load the loss for the two converter stations is 1.6% of rated power and the standby loss is 1/12 of this figure. However, in general, losses for different converter stations may vary by 10% each side of this example. *Table 32.1* gives the distribution of losses for this case between the major items, at full load and at standby. It can be seen that transformers and valves account for over 80% of losses.

Table 32.1 Distribution of converter stations' losses

	Losses at standby	*Losses at 1 p.u. power*
1 Valves	0.1	0.411
2 Converter transformers	0.755	0.407
3 Filters	0	0.044
4 Smoothing reactors	0	0.095
5 Auxiliaries*	0.145	0.043

* Excludes building services

32.5.9 Converter stations' prices

As discussed in 32.2.1, h.v.d.c. converters plus h.v.d.c. overhead lines or cables were initially being proposed as a straight economic alternative for h.v.a.c. overhead lines or cables, provided the transmission distance was sufficiently long. To enable utilities to make a preliminary estimate of the 'break-even' distance, beyond which it was worthwhile considering h.v.d.c., curves of converter station prices were being produced. These price curves of \$/kW were of the shape given in *Figure 32.18*, but included a margin to allow for scheme differences. However, it has become almost impossible to give a meaningful value to the margin of this price curve.

The spare capacity of h.v.a.c. transmission systems and the spinning reserves have been greatly reduced, compared to the time when the \$/kW curves were introduced. The integration

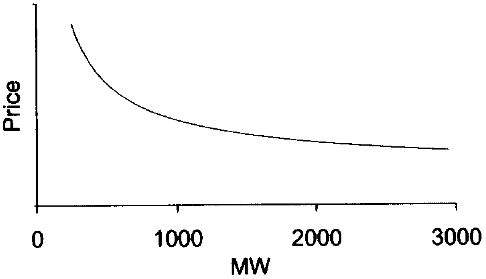


Figure 32.18 Variation of converter station price with the rated MW

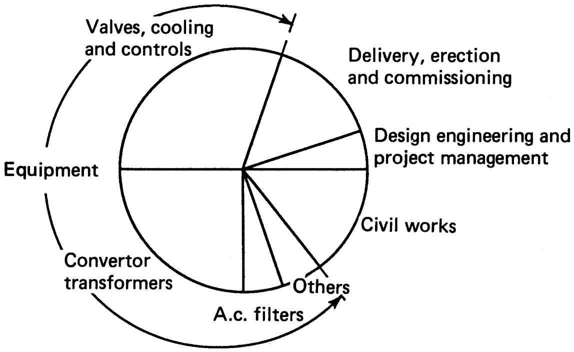


Figure 32.19 Typical cost division for h.v.d.c. converter stations

of power electronics, h.v.d.c. or large a.c. FACTS (see Chapter 41), into a.c. systems has become a complex process involving a substantial amount of engineering and the use of specialised analytical tools. Each power system is different with respect to voltage, system strength, harmonic limits and reactive power limits and each utility has different operating requirements concerning overloads, availability and reliability etc. Each h.v.d.c. scheme is therefore unique and hence caution must be exercised when making cost comparisons between different schemes or indeed between indicative prices from different manufacturers for the same scheme. Appropriate division of costs between the various components of a station is illustrated in *Figure 32.19*.

32.5.10 Reliability

The term 'reliability' is often used to describe the overall operating performance of an h.v.d.c. scheme which is quantified by its average frequency of failure and average energy availability. The desired performance criteria are usually specified for a scheme at the early planning stages to meet the overall requirements of power transmission strategy. The inclusion of financial penalty clauses in scheme contracts has led to great emphasis being placed on system reliability and availability by system designers.

A quantitative analysis is used to assess the effect of the basic elements in the scheme on the overall performance. Whenever possible equipment with proven reliability is used, but the use of redundant capacity is also extensive. Equipment such as thyristor valves, cooling plant, auxiliary power supplies and control systems usually include redundancy.

The provision of adequate spares to minimise maintenance and repair times contributes to system availability. Spare converter transformers and smoothing reactors may be considered essential due to the long lead time for repair or replacement even though in practice both have proved to be highly reliable items.

The energy availability of a bipolar transmission scheme can be maximised if the scheme has the capability to transmit power during forced outages of one pole by operating its remaining pole as a monopole. This can be achieved in the case of converter station pole equipment failures by using its conductors as a metallic return path for the remaining operational pole. The availability can be further increased by utilising an earth return system rated at full current to permit power transfer even during outages of a transmission line conductor.

Emphasis is placed on providing independence between the poles of bipolar schemes such that the number of possible common failure modes is kept to an economic minimum, while still sharing the transmission line and associated d.c. switchgear. Overhead transmission lines usually have both pole conductors on common towers. Experience has shown that common mode failures in this arrangement are unlikely. Even in areas of high lightning activity failures are usually restricted to one pole.

Typical performance targets would be a frequency of failure per pole of a converter station of 1 per year, an availability at full rated power of 98% and scheme total energy availability of 99.25%.

32.6 Insulation co-ordination of HVDC converter stations[5]

32.6.1 Introduction

Insulation co-ordination is the selection of the electric strength of equipment in relation to the voltages to which it may be exposed. Protective devices are chosen to reduce the voltage stresses imposed on the equipment to an economically and operationally acceptable level. The main object of insulation co-ordination for any system, whether a.c. or d.c., is to ensure reliable operation of the scheme at minimum cost.

Generally, an a.c. system is considered to consist of parallel-connected equipments which all have identical insulation levels. A d.c. converter station consists of both series and parallel connected equipment. However, when examining the insulating characteristic of individual a.c. equipment in detail, it is often found that the equipment has been designed and manufactured in discrete units which are connected in series. Obvious examples of this technique are found in shunt capacitor banks, insulators and a.c. circuit breakers which may use several interruptors in series. Less obvious examples include transformers and reactors.

Insulation levels on a.c. systems are relatively higher the lower the a.c. system voltage (facilitating co-ordination between different voltage levels). In a d.c. converter it is generally economical to have relatively lower insulation levels than on the adjacent a.c. system. These lower levels are made possible by close control of the voltages applied during both normal and transient conditions, but this does mean that a.c. system overvoltages can cause significant energy absorption in the converter surge arresters.

32.6.2 Sources of overvoltages

The magnitude and slope of overvoltages arriving at the converter station from the a.c. system will be attenuated by the action of a.c. filters and converter transformer reactance so that overvoltages with fast front times (less than, say, 10 µs) do not penetrate the converter from the a.c. system.

Similarly, lightning or other impulsive overvoltages travelling along d.c. overhead line towards the h.v.d.c. converter station will be almost fully reflected at the d.c. smoothing reactor.

Thus, thyristor valves are protected from fast transient overvoltages arising from the a.c. system by the converter transformer, and from the d.c. system by the smoothing reactor. However, in the event of an insulation breakdown within the boundaries established by these protecting inductances, a fast transient overvoltage may occur across the thyristor valves. The most onerous overvoltage occurs if a flashover from an outer (highest d.c. voltage) converter transformer bushing to ground takes place when the converter station has been charged by a switching surge originating in the a.c. system. During such an event the prospective valve overvoltage can be up to twice the switching surge protective level of the valve arrester. It is important to ensure that this surge arrester is able to limit the overvoltage to a level which is safe for the thyristor valve and that its energy absorption capability is adequate for this duty.

32.6.3 Surge arresters

The zinc-oxide non-linear resistor material used in modern surge arresters exhibits a very high impedance at normal applied voltage whilst at a voltage only some 50% higher a very low impedance is provided. The extremely non-linear relationship between voltage and current shown in *Figure 32.20*, has rendered obsolete the spark gaps which were a feature of previous arresters based on silicon carbide.

Gapless metal oxide arresters are simple in construction, consisting merely of enough resistor blocks connected in series to ensure that the current during normal operating conditions is very small, typically less than one milliampere. The zinc oxide resistor becomes unstable if the continuously applied voltage appreciably exceeds this level, which defines the *maximum continuous operating voltage* (MCOV). It may also suffer thermal runaway if the surge energy absorbed is too high (unless the applied voltage is removed after the transient). Its capacity for energy absorption is limited either by the accumulation of the consequences of these two major thermal considerations or in some cases by instantaneous thermal shock.

The protective characteristic of the arrester is the envelope of the discharge voltage (being the maximum voltage developed across the arrester during the passage of a specified current impulse waveform) for various waveshapes. Parallel connected surge arresters can be made to share the total energy to be absorbed in limiting switching surge overvoltages if their voltage–current characteristics are properly matched during manufacture.

32.6.4 Surge arrester arrangement

A typical arrangement of surge arresters is shown in *Figure 32.21*. Arresters are normally connected across each individual thyristor valve, and also across each six-pulse group. In schemes employing more than one 12-pulse group in series per pole, the point of interconnection between groups is also protected by a surge arrester connected to ground. The d.c. line (and d.c. reactor) is similarly protected by a surge arrester.

The surge arrester protective level applicable to each item of equipment is obtained by examination of the circuit to find the path giving the lowest discharge voltage. Thus, the protective level between the outer converter transformer bus-bar and ground is determined by the series connection of the valve arrester and the six-pulse bridge arrester (in the case of two 12-pulse groups in series the protective level of the 12-pulse group arrester must also be added).

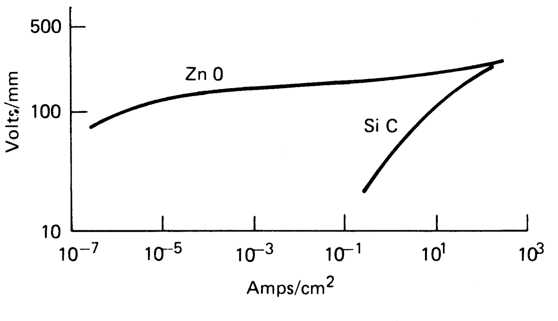


Figure 32.20 Voltage–current characteristic of non-linear resistor material

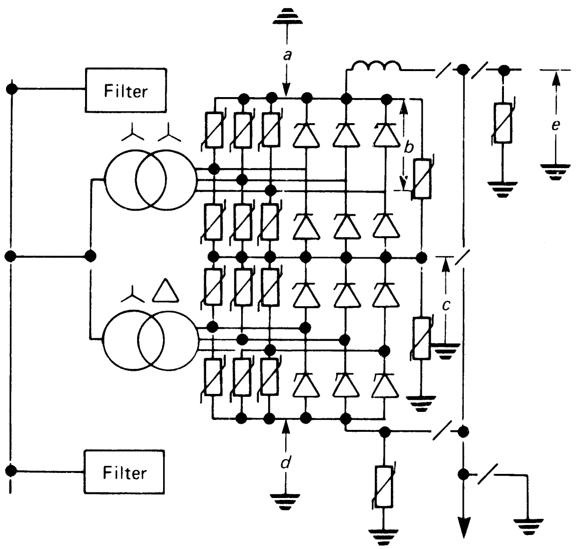


Figure 32.21 Arrangement of surge arresters in one pole of a 500-kV converter

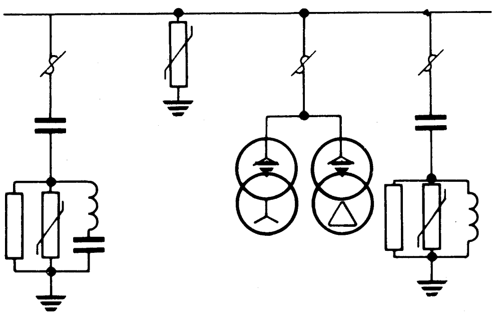


Figure 32.22 Arrangement of surge arresters on the a.c. side of a converter

Surge arresters are also used on the a.c. system to protect the converter station in the manner shown in *Figure 32.22*. The phase-to-earth insulation is protected by surge arresters which are often placed close to the transformer terminals. Additional surge arresters may be applied within the a.c. harmonic filters specifically for the protection of filter reactors and resistors.

32.6.5 Safety margins

The source and magnitude of each credible overvoltage can be calculated, knowing the impedance of the circuit elements and the characteristics of arresters. Then the insulation level of the equipment in the converter station can be determined. For most items of conventional equipment the voltage withstand increases as the front time of the applied impulse decreases (*Figure 32.23*). For such conventional

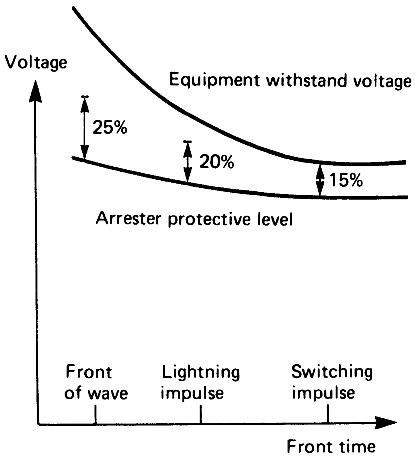


Figure 32.23 Insulation characteristics of conventional equipment

plant the safety margin between the protective level of the arrester and the withstand voltage of the equipment is 15% for switching surges. A minimum safety margin of 20% at lightning impulse waveforms and 25% at front of wave (FOW) waveforms is usually applied. The relatively flat protective characteristic of zinc oxide means that the actual margin between the protective level and the equipment withstand at lightning and FOW for most equipment is substantially higher than the minimum recommended. In practice, standard insulation levels (e.g. the IEC 71 series) are used for conventional eqipment.

The safety margins applied to conventional equipment have evolved over many years and are intended to take into account both the measuring tolerances and the anticipated deterioration with age of the insulation of the protected equipment and of the surge arrester characteristics. A margin is also necessary to allow for the increase in voltage which may arise as the distance from the surge arrester to the protected equipment increases.

Arresters are placed immediately adjacent to valve terminals, so no allowance is needed for distance effects. The thyristor valves incorporate redundancy which can be restored at regular (e.g. annual) intervals by replacement of any failed thyristors ensuring that the insulating properties of the thyristor valve stay virtually constant throughout its life.

The withstand voltage of the thyristor valve in its off-state is dictated not only by the sum of the withstand voltages of all the thyristors (which is essentially independent of waveform), but also by the interaction between its distributed in-rush-limiting reactors and the thyristor grading and damping network. During slow wavefronts only a small proportion of the applied voltage will appear across the reactors. However, as the front time decreases an increasing proportion of the voltage will appear across the reactors, and as a result the valve withstand voltage increases. It is possible to match the valve voltage–time curve to the surge arrester characteristic, achieving the margins listed above for the various impulse waveforms.

32.6.6 Creepage and clearance

The selection of creepages and clearances for the converter station is an important part of insulation co-ordination. The creepage required for a given item of equipment will vary substantially with the environment in which it is required to operate. The creepage length required is proportional to the maximum continuous voltage. In the clean air-conditioned environment provided by a valve hall, a creepage of 14 mm/kV peak will provide satisfactory performance. Under polluted conditions, such as may be present on d.c. overhead lines or on outdoor converter equipment in industrial areas, a creepage distance of 40 mm/kV peak, or sometimes even more, may be required to give adequate performance. Clearances within the converter station are determined primarily by lightning impulse and switching impulse withstand requirements.

32.6.7 Application examples

The economic incentive for using a low protective level across the thyristor valves is very strong, since the number of thyristor levels required is directly proportional to this voltage. Therefore, the valve arresters normally exhibit a low protective level, which means that they may be required to absorb large amounts of energy during overvoltages.

With all a.c. harmonic filters connected, recovery from a local three-phase short circuit to ground in a weak a.c. system may cause high prospective overvoltages. These overvoltages will have a high content of low order harmonics, and will therefore appear relatively unattenuated on the valve winding side of the converter transformer. The a.c. system phase to ground arresters will limit the peak amplitudes of these recovery overvoltages, but the thyristor valve surge arrester may nevertheless be required to absorb a large amount of energy.

A flashover from the outer converter transformer valve winding bus-bar to ground, occurring when the d.c. line is charged to overvoltage can also lead to high energy absorption in the thyristor valve surge arrester. During this event the surge arrester absorbs a substantial part of the energy stored in the capacitance of the d.c. line and d.c. filters (where applicable). Although this event is not very likely, the surge arresters across the top three thyristor valves are sometimes specified to be capable of higher energy absorption than the other valve arresters to accommodate it.

Figure 32.24 shows typical lightning and switching impulse levels in a 500 kV d.c. pole as applicable at various locations, (a) to (e), shown in *Figure 32.21*. It should be noted that the thyristor valves use non-standard insulation levels.

During normal operating conditions the inductor and resistor of an a.c. or d.c. harmonic filter experience only a small fraction of the total line-to-ground voltage. However, a major fraction of any transient overvoltage can appear across the inductor and/or resistor. Filter energisation is an example of a routine event causing an overvoltage across the filter components. If the a.c. system is strong or if several filters are already connected to the bus-bar, the voltage across the inductor when energised at peak voltage can easily approach the full line-to-ground voltage. By connecting a surge arrester in parallel with the inductor or resistor as shown in *Figure 32.22* it is possible to utilise components with an insulation level significantly below that applicable to the rest of the a.c. system. However, if it is fitted, such an arrester may be subject to very fast-rising wavefronts,

	a	b	c	d	e
BIL	1300	603	650	60	1175
BSL	1172	586	586	35	992

Figure 32.24 Typical lightning and switching impulse levels in a 500-kV d.c. pole

associated with high energy discharge duties. For example, if a flashover occurs from the a.c. bus-bar to ground, most of the energy stored in the main capacitor will be discharged into the surge arrester. Such an event can lead not only to high energy absorption in the arrester, but also to very high discharge current amplitudes. For example for a filter connected to a 400 kV system the arrester protecting the inductor of the filter may need a co-ordination current of 80 000 A. By using the arrangement shown in *Figure 32.22* it becomes possible to specify for inductors the Basic Insulation Level (BIL) of 650 kV, whereas the a.c. system BIL is 1425 kV.

32.7 HVDC thyristor valves

32.7.1 Introduction

Together with the central control system, the valves and their auxiliary cooling and overvoltage protection equipment account for approximately one-third of the equipment cost of the converter terminal. In addition, between 30% and 40% of the total station loss is incurred by the valves. At typical capitalised values of US $3000–8000 per kilowatt, the evaluated cost of losses can approach and sometimes exceed the capital cost of the valves. In such cases, it is economical to invest more in the hardware to reduce the level of losses.

The single most significant variable that influences both the capital cost and the level of losses is the number of series connected thyristors in each valve. The most economic solution is one that uses the minimum number of series levels consistent with reliable long term operation.

A valve is required to act as a switch. It should switch on (turn-on) and switch off (turn-off) efficiently. When off, it should withstand the applied forward and reverse voltages and when on, it should have low resistance.

Unfortunately, thyristors are not perfect switches. At turn-on, they initially have reduced current carrying capacity. At turn-off, the current reverses for a brief period while the thyristor stored charge is extracted and the thyristor's ability to withstand forward voltage is severely limited for some time after negative recovery starts.[6,7] *Table 32.2* summarises how the principal thyristor characteristics are influenced by design parameters. The thyristor designer can trade-off one characteristic against another to achieve the most economic solution for a given application.[6] Close collaboration between the converter valve designer and the thyristor designer is essential in order to achieve the best economic solution.

When properly applied, the thyristor does not suffer from any 'ageing' effects which would cause deterioration of its characteristics with time and it has proved to be a very reliable device provided it is used within its rating. *Figure 32.25* indicates the significant advances which have been achieved in recent years in both voltage and current ratings of thyristors.

Table 32.2 Interaction between thyristor parameters

Thyristor parameter	*Desired magnitude*	*Increased by*	*Reduced by*
Current rating	High	Increased area, increased carrier lifetime, packaging (improved cooling), reduced thickness	Reduced area, reduced carrier lifetime Increased thickness
Voltage rating	High	Increased resistivity, increased thickness, shallow impurity gradients, uniformity of purity distribution (NTD silicon), edge profiling (reduced peak surface field), edge passivation (stability)	Lower resistivity, reduced thickness, steep impurity gradients, less uniform purity distribution, steeper edge bevels
Turn-off time	Low	Increased carrier lifetime, increased thickness, emitter geometry	Reduced carrier lifetime, thinner slices
Stored charge	Low	Increased carrier lifetime, increased thickness	Reduced carrier lifetime, thinner slices
Surge current	High	Increased area, reduced thickness, increased carrier lifetime	Reduced area, increased thickness, reduced carrier lifetime
dV/dt withstand	High	Shorted emitter pattern density, vertical diffusion geometry	Trigger sensitivity
In-rush dI/dt capability	High	Fast rising gate pulse, amplifying gate, interdigitated gate	Heavy shorted emitter pattern, minimal gate area
V_{gt}, I_{gt}	Low	Design for high dV/dt	Favourable lateral patterns, favourable diffusion profiles
Turn-on delay time	Low	Carrier lifetime control	Favourable vertical diffusion profile, vertical geometry
Forward voltage drop	Low	Thicker slices, shorter emitter pattern density	Thinner slices, less dense shorter emitter pattern

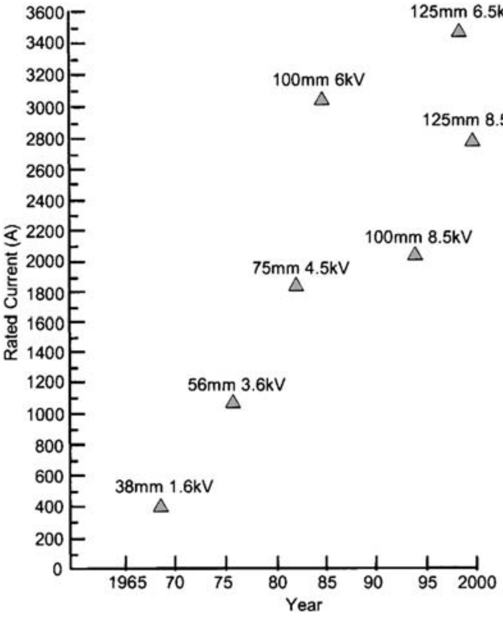


Figure 32.25 Advances in thyristor ratings

32.7.2 Thyristor level circuits

To cater for high voltages for h.v.d.c., each valve is made up of many thyristor levels connected in series.

Figure 32.26 shows the basic electrical circuit of one thyristor level as it may typically be implemented. Many variants are possible but all key features are shown. In some designs one series reactor serves several thyristor levels.

The power thyristor is electrically triggered via its gate and logic unit in response to an optical command received via a fibre optic waveguide from earth potential. Electrical power for energising the electronics is derived from the main damping circuit, local to the thyristor.

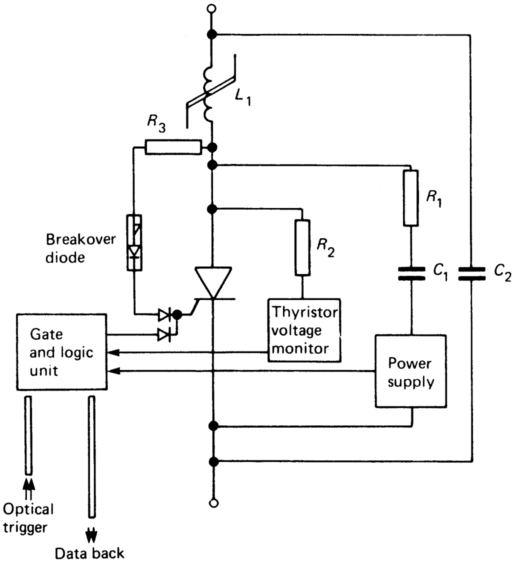

L_1
R_3
R_1
R_2
Breakover diode
C_1
C_2
Thyristor voltage monitor
Gate and logic unit
Power supply
Optical trigger
Data back

Figure 32.26 Basic circuit of one thyristor level

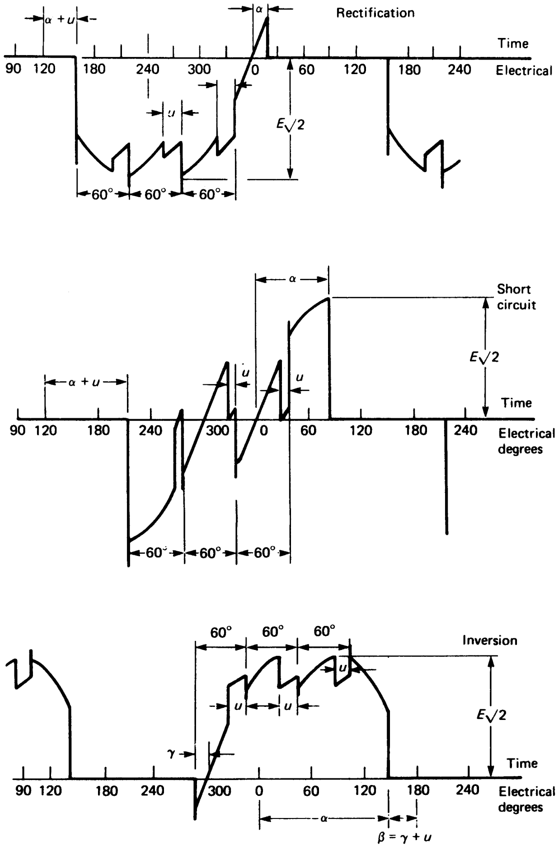
The gate and logic unit and the associated power supply are designed to ensure that full and adequate control and protection of the thyristor is afforded, not only under normal steady-state operation, but also under abnormal and disturbed conditions, such as operation with discontinuous current or the loss of a.c. system voltage for up to the maximum a.c. system fault clearance time (typically 200–300 ms).

In series with the thyristor is a saturable reactor to control inrush current and across each thyristor the usual capacitor and capacitor resistor circuits used for voltage grading and damping purposes.

32.7.3 Voltage rating

The required d.c. voltages can vary from 25 kV for a 50 MW back-to-back scheme to 600 kV to ground (1200 kV −ve to +ve line) for a 3000 MW, long transmission scheme. Thyristors for h.v.d.c. having a withstand capability in excess of 5 kV are already available and the economic pressure to reduce the number of series levels will continue to force ratings upward.

In the case of long distance schemes, the need to optimise the transmission line costs and losses usually results in a defined optimum transmission voltage for a particular power transfer requirement.


$\alpha + u$
α
Rectification
Time
90
120
180
240
300
0
60
120
180
240
Electrical
u
$E\sqrt{2}$
60°
60°
60°
α
Short circuit
$E\sqrt{2}$
$\alpha + u$
u
u
Time
90
120
180
240
300
0
60
120
180
240
Electrical degrees
60°
60°
60°
60°
60°
60°
Inversion
u
u
u
$E\sqrt{2}$
γ
Time
90
120
180
240
300
0
60
120
180
240
Electrical degrees
α
$\beta = \gamma + u$

α = Firing delay angle
u = Commutation overlap angle
γ = Safety angle
E = Commutating EMF (RMS line-line)

Figure 32.27 Voltages across a valve operating in a six-pulse bridge for rectification, zero voltage and inversion

For back-to-back interconnections the designer can choose the operating voltage and current to suit the capability of available thyristors. This usually results in a scheme with low operating voltage and high current. Whatever the application, the general form of the normal valve terminal-to-terminal voltage is similar. The voltage waveshape is complex (*Figure 32.27*), containing fundamental frequency components, fast transient components and a d.c. off-set.

Each valve is protected against overvoltages by a zinc oxide surge arrester, connected directly across its terminals. For maximum economy, the protective level of this arrester should be as low as possible. The achievable protective level depends on the performance requirements of the arrester, which in turn is determined by the system voltage conditions. While the nature of transient overvoltages is important because it determines the energy rating of the arrester, it is the ability of the arrester to withstand the peak of the fundamental frequency voltage after a maximum energy surge, which is normally the determining factor. Once the voltage rating of the arrester is chosen, the protective level arises naturally from the zinc oxide voltage–current characteristic and the value of co-ordinating current.

The protective level of the arrester is not a single value but is somewhat dependent on the front time of the incident voltage wave, the value being higher for faster fronted impulses. This must be taken into account in the valve design. In addition, the valve must contain enough thyristors in series to be capable of being tested at values sufficiently above the arrester protective level to give confidence that long service life will be achieved.

In accordance with international standard IEC 700, valves are normally designed for the following test margins with respect to the surge arrester protective levels:

15% for switching impulse,
15% for lightning impulse, and
20% for front-of-wave.

Having established the test voltages to be applied, the number of thyristors may be determined. This number is a function of the reverse voltage capability of the thyristors and on the achievable accuracy of the grading circuit.

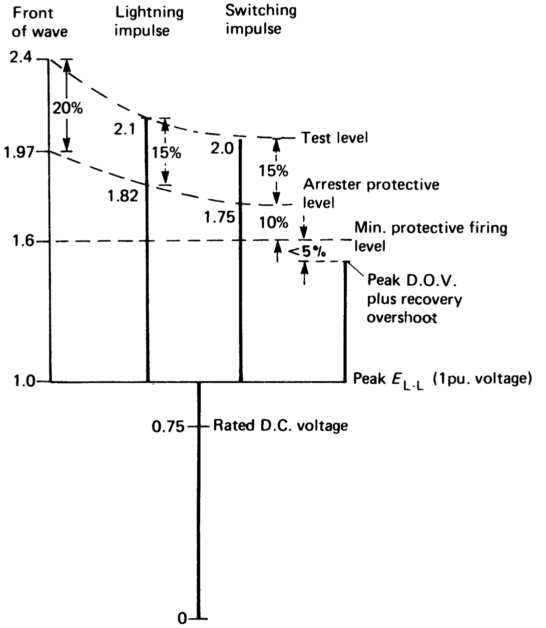

Front of wave
Lightning impulse
Switching impulse
2.4
20%
1.97
2.1
15%
1.82
2.0
15%
1.75
10%
<5%
1.6
1.0
0.75
0
Test level
Arrester protective level
Min. protective firing level
Peak D.O.V. plus recovery overshoot
Peak E_{L-L} (1pu. voltage)
Rated D.C. voltage

Figure 32.28 Typical insulation co-ordination for a thyristor valve

Thyristors can be damaged if they are exposed to excessive voltage in the forward direction. Further, the forward voltage capability is usually lower than that in the reverse direction. For this reason, protective firing of the thyristors is often employed for overvoltages in the forward direction. *Figure 32.26* shows this feature implemented by a breakover diode (see Section 32.7.5.). The valve voltage, above which protective self firing may occur, depends on the detailed application but, as a general rule, it will not be below 90% of the surge arrester protective level for switching surge wavefronts (see *Figure 32.28*).

32.7.4 Current rating

For an h.v.d.c. application it is desirable that any credible fault current, arising from insulation failure, can be blocked by the thyristors at the first current zero. It is therefore necessary to ensure that the post-fault thyristor junction temperature is below the critical value, above which the thyristor may not be able to block the ensuing recovery voltage.

The problem is a thermal one, bound by two limits: one limit is the temperature above which the thyristors may be

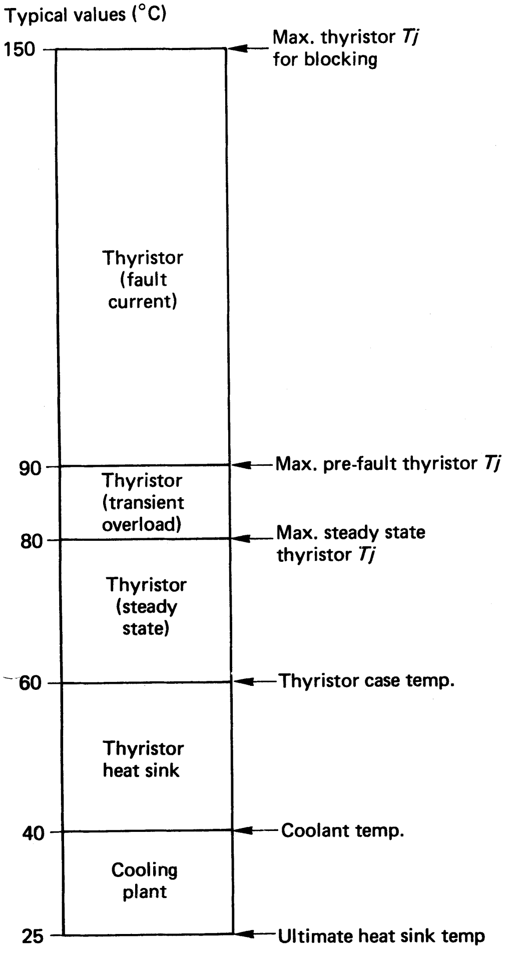

Typical values (°C)
150
90
80
60
40
25
Max. thyristor Tj for blocking
Thyristor (fault current)
Max. pre-fault thyristor Tj
Thyristor (transient overload)
Max. steady state thyristor Tj
Thyristor (steady state)
Thyristor case temp.
Thyristor heat sink
Coolant temp.
Cooling plant
Ultimate heat sink temp

Figure 32.29 Typical breakdown of temperature rise between the coolant–thyristor junction

unable to block the post-fault recovery voltage, the other limit is the ultimate heat sink temperature to which all losses are finally dissipated (e.g. ambient air, river water, etc.).

Figure 32.29 shows a typical breakdown of temperature rise between the two limits. In practice, the designer has direct or indirect control over each component of the temperature rise and, for any given application, aims to achieve an optimum economic balance. As can be seen, nearly 60% of the available temperature rise has to be kept in reserve as a margin for transient overloads and fault currents. It is most important to be able to determine the temperature rise resulting from any fault current.

Figure 32.30 indicates how junction temperature follows the current. A short circuit caused by a d.c. side flashover must not be allowed to raise junction temperature to a value that would prevent forward blocking after current zero.

There are sophisticated techniques for producing an accurate mathematical model from which the worst case thyristor junction temperature excursion, for any applied current waveform, can be derived.[6,8] Using such models, the sensitivity of junction temperature to changes in fault current produced by changes in transformer leakage reactance can be determined and an optimum value of reactance chosen.

Technically, the method of cooling employed is irrelevant to the performance of the thyristor as it is only a means of achieving a defined steady-state condition from which junction temperature excursions due to transients and faults can commence. Because heat generated at the thyristor junction does not reach the heat sink in times less than 1 s, the design of heat sink and the method of cooling have no influence on the transient excursions due to faults and disturbances, which rarely last more than a few hundred milliseconds.

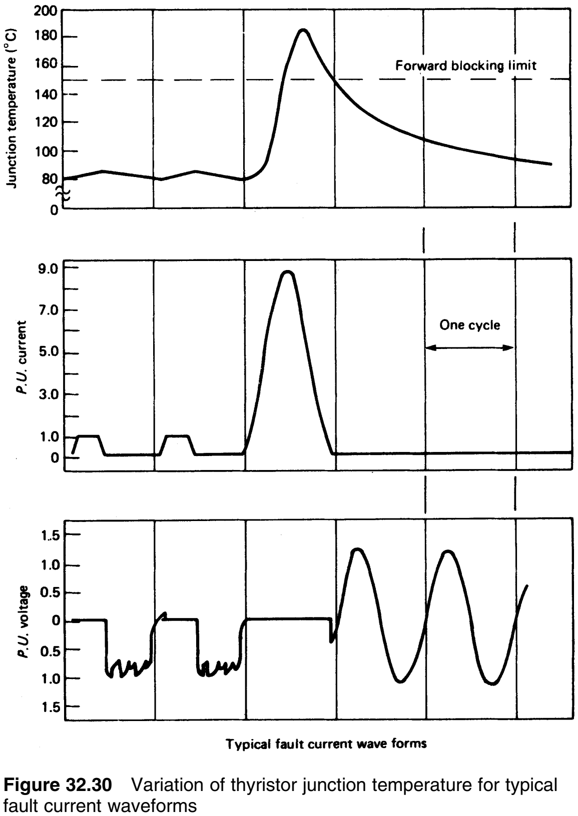


Figure 32.30 Variation of thyristor junction temperature for typical fault current waveforms

Increasingly, the cost of losses is such that it is often justified to select a larger thyristor than could technically satisfy the current requirements of a scheme. Under these conditions, considerations other than overcurrent rating may become limiting (e.g. maximum coolant temperature).

32.7.5 Turn-on behaviour

When thyristors are gated, there is a short (1 to 2 μs) delay before any significant change in impedances takes place. After this initial delay, the impedance of each thyristor collapses rapidly, but the steady-state impedance is not reached until several hundred microseconds later. During this turn-on phase, the thyristors have reduced current carrying capacity and it is necessary to protect them from the prospectively high rates of rise of current arising from the discharge of the circuit stray capacitances. Series connected saturable reactors are used which are active during the initial stages of turn-on but exhibit a low inductance value after conduction is established.

The rate of rise and amplitude of voltage applied to the thyristors, when fast fronted voltage waves appear at the valve terminals are controlled by the reactor and capacitor grading circuit.

Normally, valve turn-on is initiated by the coherent triggering of all thyristors in response to a firing command originating in the central control system. It is clear that if a thyristor is late turning on, it could be destroyed by excessive voltage. Protective self-firing described earlier can be achieved by electronic means or, if very high reliability is required, by use of a break-over diode (BOD) at each thyristor level. The BOD is a voltage-sensitive semiconductor switch that operates in response to an overvoltage. It gates the main power thyristor directly and, unlike the electronic alternative, is fully independent of the normal firing system. Because of tolerances, overvoltage firing of the valve is noncoherent and a cascading type of turn-on may take place, and the last level to turn-on may experience increased stress. Being independent, the BOD firing circuit acts as a back up to the normal firing circuits. If normal firing on any one levels fails, the BOD responds to the rapid rise in voltage arising from firing of other thyristors in the valve and safely initiates conduction before damage to the thyristor can occur. Efforts to integrate the BOD into the main thyristor, making it self-protecting against overvoltages will eventually remove the need for such circuits.

32.7.6 Turn-off behaviour

At turn-off, the switching action of the valve excites an oscillation between the transformer leakage reactance and the circuit stray capacitances. This oscillation cannot be critically damped but the degree of voltage overshoot can be minimised by the correct choice of values for the components of the valve grading network. The thyristors themselves aggravate the recovery overshoot as they do not cease conduction immediately the current reaches zero, but shortly after. The reverse current flow that is established before turn-off occurs allows energy to be stored in the magnetic fields of the inductive components, and this energy has to be dissipated in the damping circuits.

Immediately after turn-off has occurred, thyristors are unable to support any forward voltage without spontaneously re-conducting. Gradually the thyristors acquire some forward hold-off capability, but it is typically more than one millisecond before their full off-state capability is attained.

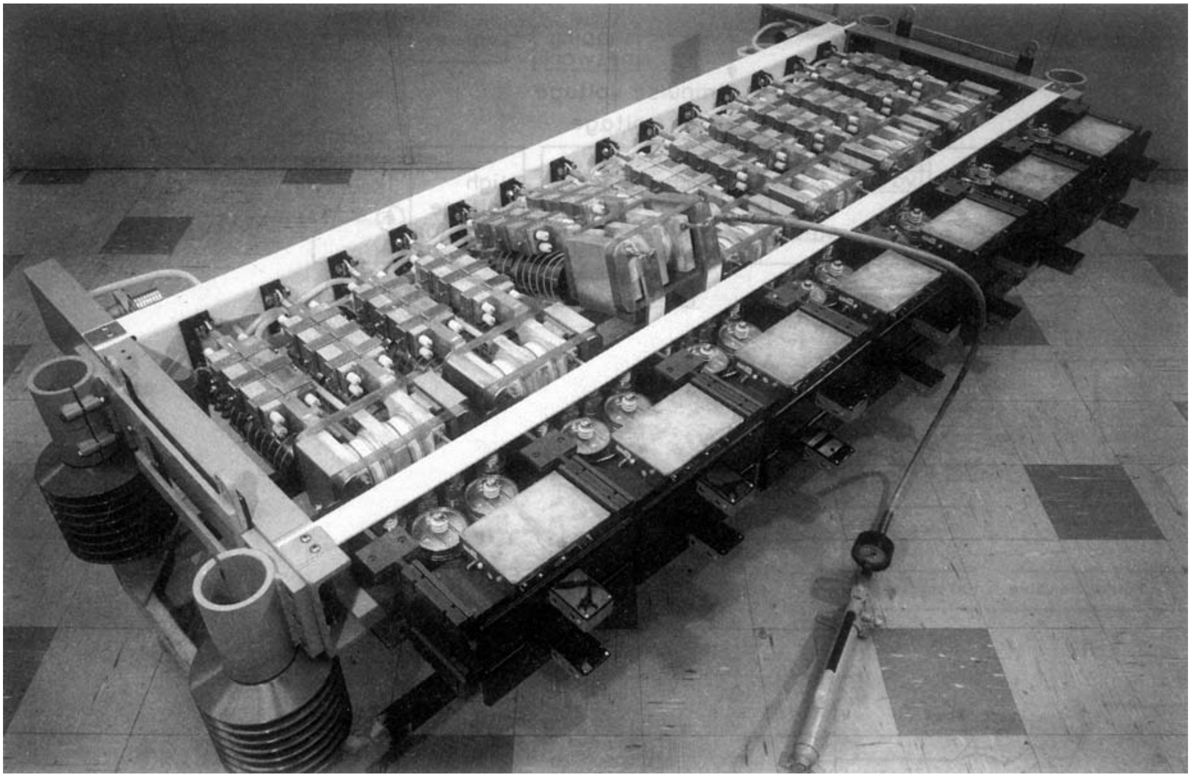


Figure 32.31 Water-cooled assembly with 14 thyristors, levels in series. (Courtesy of GEC Alsthom Transmission and Distribution Projects Ltd)

The economics of converter design dictate that inverter operation be achieved with the smallest practicable extinction angle margin (γ) but the value of γςchosen must not only allow time for the thyristors to recover sufficient forward voltage capability, but must also include additional margin so that temporary minor distortion of the a.c. system voltage waveforms does not result in an unacceptably high incidence of commutation failure. Severe transients, such as those arising from nearby a.c. line faults, will result in loss of margin angle such that commutation failure becomes inevitable (see Section 32.10).

Two aspects relating to commutation failure are worthy of mention. First, a thyristor may be exposed to positive voltage (arising from, for example, disturbed a.c. network voltage) before its recovery process is complete. Forward recovery failure of thyristors may be in itself potentially destructive to the thyristors, particularly when re-application of forward voltage is rapid. The mechanisms leading to failure are complex and are discussed in more detail in references 6 and 7. It is essential for the security and reliability of the system that proper protection is afforded to the thyristors. This usually has to be carried out at each thyristor level.

Second, in marginal cases, some thyristors may re-conduct while others do not. This leaves the valve in a partially blocked state with only a portion of the thyristors supporting the applied voltage. Those thyristors which have successfully blocked are now exposed to a prospective overvoltage, even though the valve terminal voltage may not exceed 1 p.u. The use of individual overvoltage protection, at each thyristor level, and/or whole valve protective firing in response to a data-back signal, ensures that no damaging overvoltage can occur.

32.7.7 Valve arrangements

Oil-cooled, oil insulated and air-cooled, air-insulated valves have been used in the past. However, most economic arrangements at present use air-insulation and water cooling.

The following description of a valve is given as an example of how a 1500 MW ±500 kV d.c. converter is built up. Two six-pulse bridges connected in series are used to form a 500 kV 12-pulse valve group. Therefore each valve has to be capable of operating in a 250 kV six-pulse bridge. To achieve this voltage between 70 and 80 thyristor levels (*Figure 32.26*) would have to be used, including two to three levels for redundancy.

Figure 32.31 shows a water-cooled assembly (valve section) comprising 14 series-connected thyristors arranged into seven subassemblies (banded pairs), though the number of thyristors in a valve section will vary depending on the application. Each banded pair has interleaved heat sinks and is held in compression by glass fibre composite bands and internally mounted spring washers. Replacement of a thyristor is possible without disturbance of the water circuit and the main electrical connections. The banded pair is raised into the position shown and the bands are stretched by means of an integral hydraulic ram to enable packing pieces and then a thyristor to be removed. Two counter-flow water circuits provide uniform temperature through the valve.

The valve sections which form one tier are mounted in a structure with support insulators as shown diagrammatically

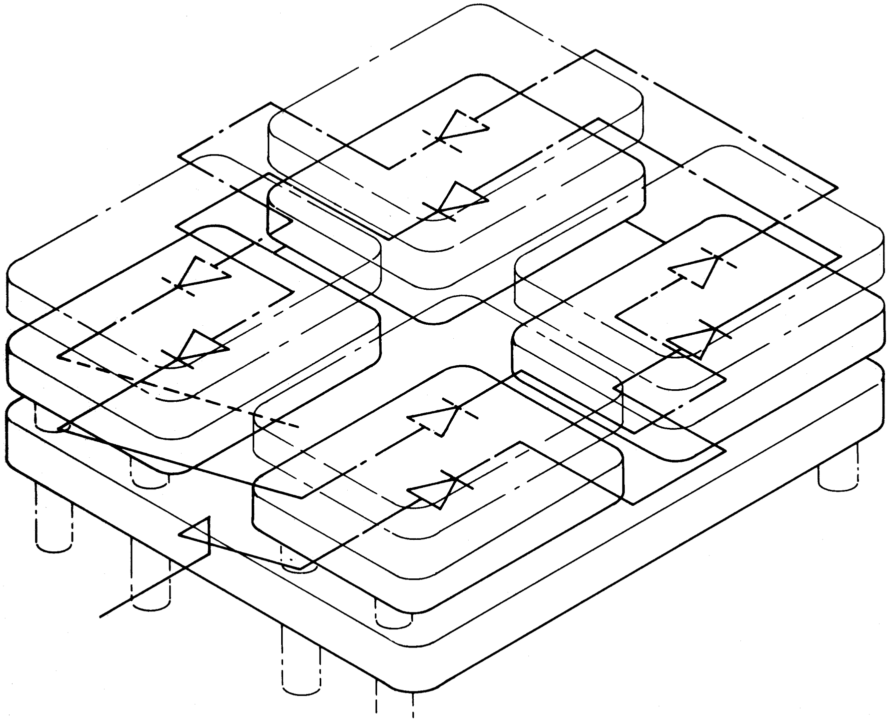


Figure 32.32 Diagrammatic illustration of part of a valve structure showing four valve sections connected in series at each tier

in *Figure 32.32*. Four such tiers are required to form one valve for the example chosen. Four valves are stacked on top of each other to form a quadrivalve. Electrical connections form a spiral from ground level (neutral voltage) through four series valves forming a quadrivalve (i.e. four valves associated with one phase of a 12-pulse valve group). Three such quadrivalves form a complete 12-pulse pole contained in one valve hall, rated 500 kV, 1500 A, 750 MW (see *Figure 32.15*).

32.7.8 Valve tests

The testing of thyristor valves for h.v.d.c. is covered by IEC 60700 (1998) and IEEE 857 (1996). Broadly speaking, these two standards have similar requirements.

Dielectric tests are performed on compete, single, valves and, when required, on multiple valve units and their supporting (or suspending) structures.

The testing of the normal operating duty of a valve is difficult because of the power limitations of practical test circuits. As a result, operational tests are normally performed on valve sections at realistic current but at proportionally reduced voltage. In order for 'realistic' stresses to be reproduced, both standards normally require the valve sections for operational tests to have five or more series-connected thyristor levels.

Two basic test methods for operational tests can be used: the first one uses 12 valve sections configured as two 6-pulse bridges in back-to-back connection and the second uses single valve sections in a synthetic test circuit with repetitive (50 Hz/60 Hz) capability. In a synthetic test circuit, various voltage and current sources are combined in an appropriate manner to reproduce representative stresses (see *Figure 32.33*). Synthetic circuits offer greater flexibility, are more accommodating of increases in thyristor rating and require lower test plant MVA than the back-to-back alternative.

32.8 Design of harmonic filters for HVDC converters

32.8.1 Introduction

The a.c./d.c. converter is a source of harmonics, and since excessive levels of harmonic distortion on an a.c. system can lead to a number of undesirable effects (overheating of induction motors, generators and capacitors, telephone interference, etc.) shunt harmonic filters are used on the a.c. terminals of all h.v.d.c. converter stations.

Harmonic distortion on the d.c. line may in some cases cause unacceptable interference in adjacent telecommunication circuits. In order to minimise this interference, harmonic filters are often fitted at the terminations of d.c. overhead lines.

32.8.2 A.c. harmonic current generation

Due to large inductance of the d.c. smoothing reactor the current conducted by each converter valve consists of a train of nearly flat topped current pulses. Thus, the current at the a.c. terminals of the converters is not sinusoidal.

Figure 32.34 superimposes the a.c. current pulses from two ideal six-pulse bridge converter circuits connected via star–star and star–delta transformers with zero commutating

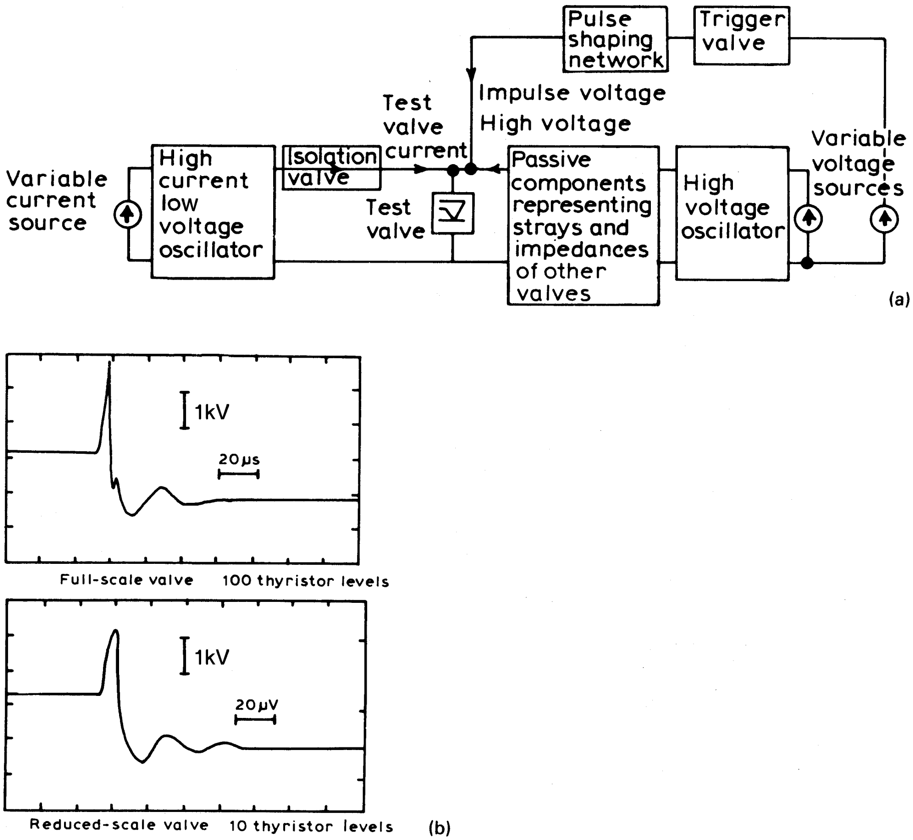

Pulse shaping network
Trigger valve
Impulse voltage
High voltage
Test valve current
Variable current source
High current low voltage oscillator
Isolation valve
Test valve
Passive components representing strays and impedances of other valves
High voltage oscillator
Variable voltage sources
(a)
1kV
20μs
Full-scale valve 100 thyristor levels
1kV
20μV
Reduced-scale valve 10 thyristor levels
(b)

Figure 32.33 (a) Synthetic circuit for valve testing (block diagram); (b) Voltage response of thyristor level with back-up triggering. (α⇐= 90°⇐firing conditions)

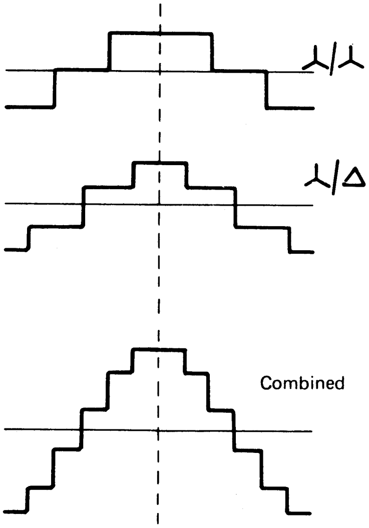


Figure 32.34 Idealised phase current on the a.c. side of a converter station

reactance but with infinite d.c. circuit inductance. A Fourier analysis gives, for a d.c. current of I_d, the following series for the star–star and the star–delta transformer, respectively:

$$i = \Leftarrow \frac{2\sqrt{3}}{\pi\varsigma} I_d[\cos(\omega t) - 1/5\cos(5\omega t)\Leftarrow$$

$$+ 1/7\cos(7\omega t) - 1/11\cos(11\omega t) + \cdots] \Leftarrow$$

$$i = \Leftarrow \frac{2\sqrt{3}}{\pi\varsigma} I_d[\cos(\omega t) + 1/5\cos(5\omega t)\Leftarrow$$

$$- 1/7\cos(7\omega t) - 1/11\cos(11\omega t) + \cdots] \Leftarrow$$

By addition of these two currents the fifth and seventh harmonics cancel and only harmonics of orders $12\,\mathrm{k} \pm 1$ will enter the a.c. system for a 12-pulse group.

In practice, because the converter transformer reactance is not zero, the current takes a finite time to transfer from one valve to the next. As shown in *Figure 32.35* the resulting current waveform is smoother than that shown in *Figure 32.34*. The amplitude of each harmonic component of current depends on the value of overlap angle u which is a function of the commutating reactance, of the firing angle $\alpha\varsigma$(or γ) and of the load current. The effect of u can be judged from *Figure 32.36*. The 11th harmonic with $u = 0$ would be 1/11 (9.1%) of the fundamental, while in practice it is nearer to 4%. *Figure 32.37* shows the variation of characteristic harmonics with d.c. load. Text books should be consulted for a full analysis of harmonic currents.

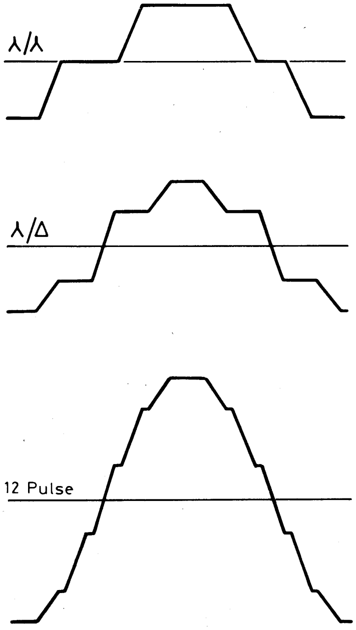


Figure 32.35 Phase current with firing and overlap delays

The theoretical analysis above is valid for balanced a.c. systems and converter operation. In practice, neither the a.c. system nor the converter circuit are perfect, and cancellation of harmonics will be incomplete. The major causes of harmonics other than $12k \pm 1$ are:

(1) a.c. system phase unbalance including non-linear loads;
(2) unbalance between six-pulse bridges; and
(3) unbalances within six-pulse bridges.

32.8.3 Filtering

Shunt harmonic filters connected to the converter a.c. bus-bars provide a low impedance into which most of the harmonic currents are diverted. Shunt filters also generate reactive power at fundamental frequency providing some or all of the reactive power required by the converters.

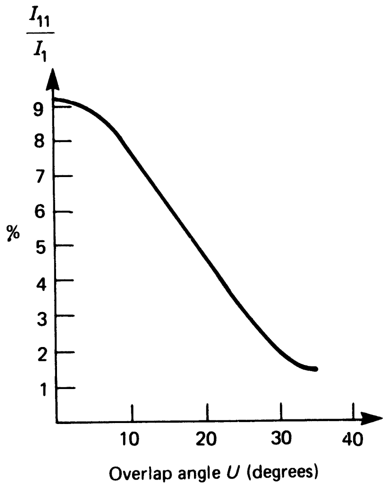


Figure 32.36 Eleventh harmonic as a percentage of the fundamental harmonic

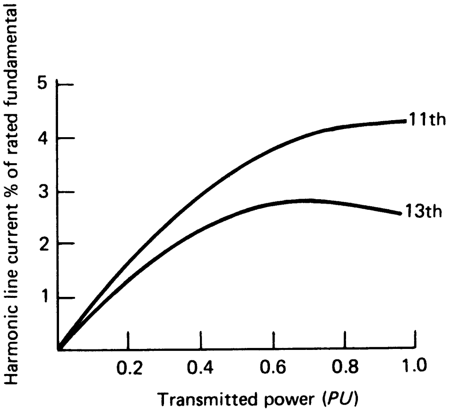


Figure 32.37 Harmonic generation

The most direct method of achieving a low impedance at a given frequency is by means of a tuned filter as shown in *Figure 32.38(a)*. The admittance of a tuned filter varies sharply around the resonant frequency as demonstrated in *Figure 32.40*. The sharpness of tuning and the var rating of the filter must be chosen to achieve the specified performance over the required range of system frequency, temperature and component tolerances.

Each sharply tuned filter is capable of providing significant attenuation at one frequency, but gives virtually no damping at other frequencies. This means that the a.c. bus-bar voltage during transient phenomena may exhibit ringing at low order non-characteristic frequencies, and that this ringing may persist for a long time. The bus-bar voltage shown in *Figure 32.41* is typical of the transient occurring during the energisation of a combination of tuned filters and a single high pass damped filter.

It is possible to combine two separate tuned filters in a single filter as shown in *Figure 32.39*. The characteristics

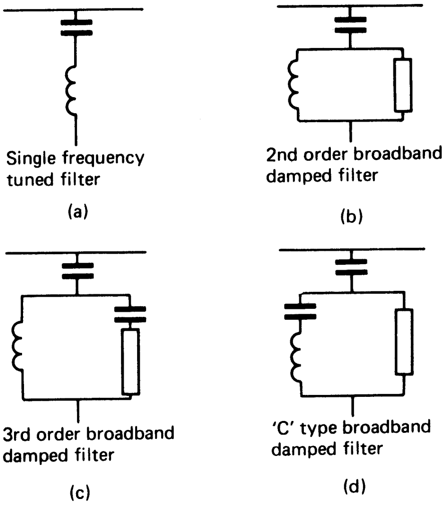

Single frequency tuned filter
(a)
2nd order broadband damped filter
(b)
3rd order broadband damped filter
(c)
'C' type broadband damped filter
(d)

Figure 32.38 Alternative types of harmonic filter

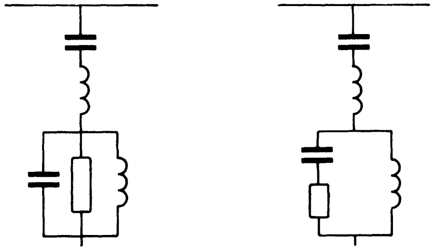


Figure 32.39 Alternative types of double-tuned harmonic filters

can be altered by increasing the damping thus reducing the sharpness of 'tuning'. Combining the main capacitors of two individual tuned filters to create one double-tuned filter can often give significant cost savings.

The damped broadband filter shown in *Figure 32.38*(*b*) requires a significantly higher var rating than a corresponding sharply tuned filter to provide the same harmonic absorption at a given frequency, but *Figure 32.40* shows that damping and filtering is provided over a range of harmonic frequences.

Damped filters must have a higher var rating than the corresponding tuned filters to achieve same filtering performance, and their losses are higher. The losses can be reduced significantly by the use of an additional capacitor as shown in *Figure 32.38*(*c*) and *32.38*(*d*). In the third-order broadband damped filter the fundamental frequency losses in the resistor are reduced by the use of a capacitor C2 in series with the resistor. In the 'C'-type damped filter a capacitor C2 is connected in series with the inductor L and the fundamental frequency losses are minimised by tuning C2 to resonate with the inductor at fundamental frequency. The 'C'-type filter attenuates low order non-characteristic harmonics, where the losses in a second-order filter would be uneconomically high.

Figure 32.42 shows the bus-bar voltage following energisation of a combination of a 'C'-type and second-order damped filter. The transient can be seen to be virtually damped out within 10–15 ms after energisation.

In general, a combination of tuned and high-pass damped filters will provide the lowest loss solution. However, a combination of damped filters can often provide substantially better service to the a.c. system by preventing resonances.

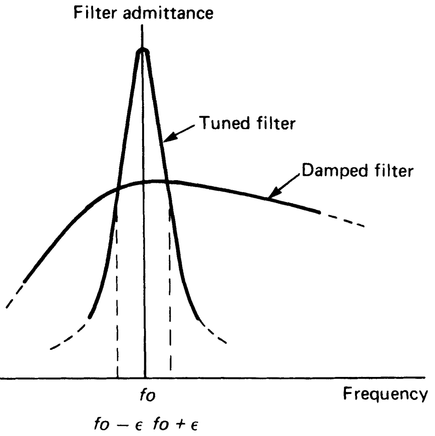


Figure 32.40 Filter performance

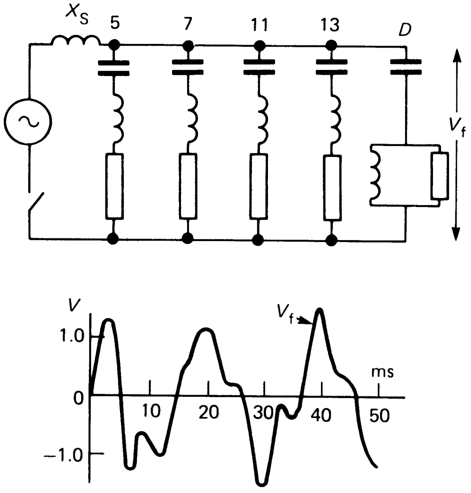


Figure 32.41 Switching transient tuned combination single- and high-pass damped filter

32.8.4 Harmonic performance evaluation

The harmonic current generated by the converter is injected into the parallel combination of the filter impedance and the a.c. network impedance as shown in *Figure 32.43*. Whilst the impedance of the filters can be determined the harmonic impedance of the a.c. system can vary substantially because of line switching operations and load/generation changes in the network. Variation of system impedance with harmonic number for a typical system condition is shown in the polar *R–X* diagram of *Figure 32.44*. Information presented in this manner is more informative than that provided in tabular form at integral harmonics because it gives a clear indication of where the resonant frequencies occur, i.e. when the locus crosses the *R* axis. When resonance occurs close to a particular harmonic under

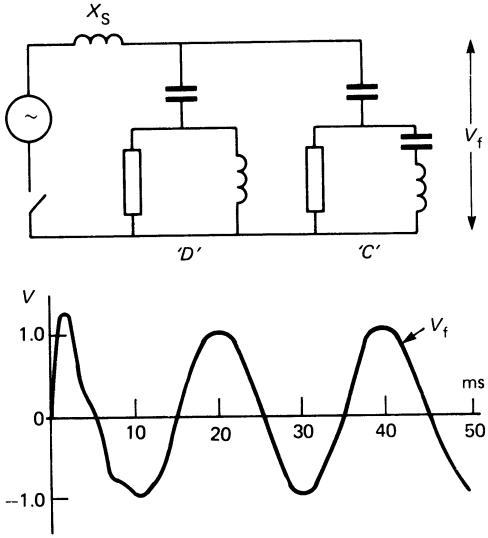


Figure 32.42 Switching transient C-type and second-order damped filter

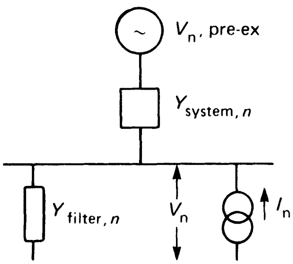


Figure 32.43 Circuit diagram for the calculation of harmonic distortion V_n at harmonic number R_n

consideration there will be a rapid change in impedance with frequency, making it difficult to assess performance accurately. Where the supply system has many different operating configurations it can be assumed that the impedance lies within a circle limited by the impedance angles ϕ, or within a segment of a circle of specified radius. The co-ordinates of such a circle encompass all the anticipated system conditions and enable the determination of the worst case of harmonic distortion to be carried out. A typical example of an harmonic impedance locus specified for filter performance evaluation purposes is given in *Figure 32.45*.

The a.c. network impedance locus is converted into an admittance locus and the worst case of harmonic voltage distortion V_n due to the converter harmonic current is determined by assuming an a.c. system admittance giving the minimum resultant admittance of the parallel combination of the a.c. harmonic filters and a.c. system.

Pre-existing distortion originating in the a.c. network must be added to the voltage distortion caused by converter harmonic currents. Permitted harmonic voltage distortion levels vary from country to country and according to the a.c. network voltage level, but typical values are 1% for odd harmonics, 0.5% for even ones, and 2% for the total root of the sum of the squares.

Filtering performance must normally be achieved for system voltage and frequency variations and ambient temperatures which can persist for long periods. For conditions which can only exist for short times, it is often acceptable to allow the specified harmonic distortion to be exceeded. However, the equipment must be rated for these more arduous conditions.

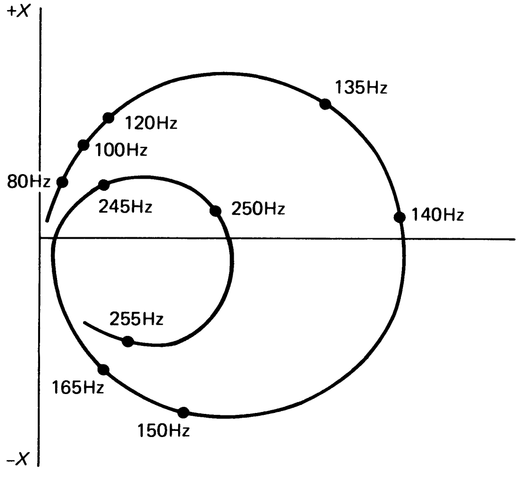


Figure 32.44 Typical supply network impedance diagram

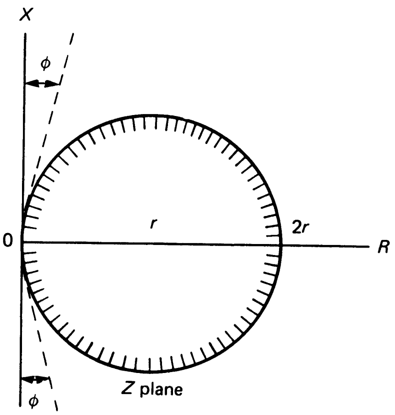


Figure 32.45 A.C. network impedance

32.8.5 D.c. filtering

The converters produce harmonic voltages between the d.c. terminals of the valve groups. In a 12-pulse scheme the lowest order characteristic harmonic is the twelfth, but because of inevitable imperfections in the a.c. system and the converter circuit, harmonic voltages of other orders will also be present.

The converter circuit contains a d.c. reactor which exhibits a large impedance at high harmonic orders, minimising the harmonic current flowing into the d.c. line. Nevertheless, the voltage and current profile along the line should be calculated and the possible induced noise in nearby communication circuits checked. The current and voltage profile is dependent on the earth resistivity and the line characteristics, and it will vary along the line.

If calculation suggests that unacceptable noise is likely to be induced in nearby circuits it may be necessary to provide filters at the converter station terminals. These filters interact closely with the d.c. reactor and the line, but often it is possible to achieve the required performance by means of a simple damped filter.

Figure 32.46 shows a typical arrangement of a d.c. filter on a converter pole.

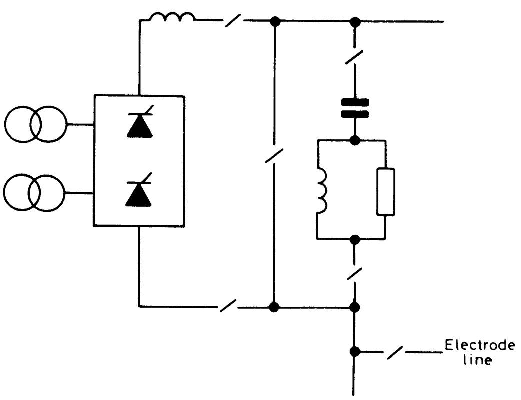


Figure 32.46 Second-order damped d.c. filter

32.9 Reactive power considerations

32.9.1 Introduction

Generation and absorption of reactive power constitute major consideration for long h.v.a.c. overhead lines and for much shorter a.c. cables. There is usually a surplus of reactive power at light load and a deficit at heavy load, and it often becomes necessary to provide fixed or variable, shunt and/or series compensation which also affects the stable power transmission limit. While h.v.d.c. lines do not consume, generate or transmit reactive power, converters do. At the a.c. terminals of the converter stations, the solution of reactive power requirements of the converters is combined with the reduction of harmonic distortion.

32.9.2 Reactive power requirements of HVDC converters

The converter absorbs reactive power irrespective of whether it is operating as a rectifier or an inverter as explained in Section 32.3.

At the rectifier end the a.c. system has normally some excess reactive power capability. Therefore, the a.c. filters are usually designed to achieve economically the filtering requirements. This may lead to the fact that the MVAr rating of filters is lower than the reactive power consumption of the rectifier, the a.c. system supplying the difference. At the receiving end the situation is normally different. Here, reactive power is required not only for inverter operation but also for loads supplied by the inverter. It is normal to specify that the converter station must as a minimum supply all inverter reactive power requirements. System requirements for reactive power can be supplied by additional shunt capacitors in the converter station or situated elsewhere in the system. In bi-directional schemes each end will in turn act as the receiving or the sending end stations.

Figure 32.47 shows the variation of the reactive power Q with changes in d.c. power for the converter of a 2000 MW scheme. The two continuous curves must allow for all the relevant design and operating tolerances. The nominal reactive power demand of the converters at full load is about 50% of the megawatt rating for this scheme.

In the example considered in *Figure 32.47* assuming that the filters are designed to supply all inverter var, Q_i at rated load, there would be an excess of reactive power supplied to the system at lower powers. This may not be acceptable. In such cases filters are designed so that parts can be switched out at lower loads, reducing Q_f, but always having sufficient filtering capacity. For the example in *Figure 32.47* the filter has been divided into four identical banks for filtering of lower order harmonics (indicated as 1C etc. in *Figure 32.47*) and four banks for filtering higher order harmonics (indicated as 1D etc.).

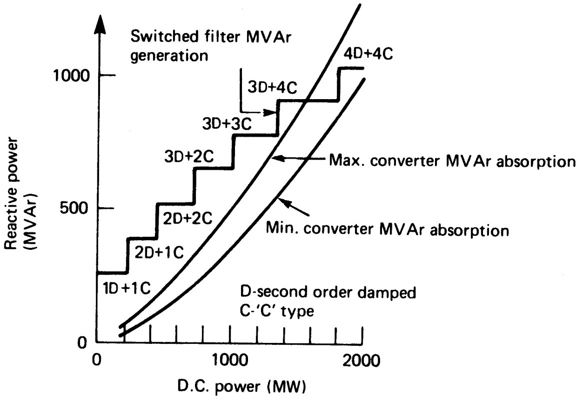


Figure 32.47 Reactive power of h.v.d.c. converters

32.9.3 Steady-state voltage control and total ratings of reactive equipment

A.c. system load flow studies including the effects of d.c. link P and Q should be carried out for all likely network situations in order to determine the limiting levels of Q that the a.c. system would be able to absorb or generate if the network and converter station voltages were to remain within the usual operating limits (e.g. ±5%).

The size of harmonic filters necessary from voltage distortion considerations at different power levels should first be determined. Further studies would indicate the required positive and negative range of additional reactive compensation taking account of the tolerances of all components. This additional reactive compensation equipment may be switched in steps or may have a proportion under smooth or continuous control, as determined by the following further considerations.

32.9.4 Voltage disturbances caused by switching operations and requirements for smooth reactive control

Step changes of reactive power caused by switching of shunt capacitors and filters have to be limited in order to minimise voltage disturbances to other consumers and to the h.v.d.c. converter itself. The permissible magnitude of voltage step may be appreciable for infrequent occurrences (e.g. 3%) but for frequent events the value may have to be smaller (e.g. 1% or 1.5%). Such step changes may be regularly imposed by filter switchings corresponding to the daily load cycle. If the minimum practical step change is not acceptable some form of smooth acting var compensator should be considered, such as a static var compensator as described in Chapter 41.

32.9.5 Control of temporary overvoltages caused by faults resulting in partial or total loss of d.c. power flow[15]

Changes in the system conditions, such as line and load switching, may cause sudden changes of reactive power; the balance between the demand and the supply of reactive power will be disturbed which will result in voltage changes. The overvoltage which may result from such a reactive power change is termed the temporary overvoltage (TOV). TOV refers to the total overvoltage waveform: the fundamental component of TOV plus any oscillatory superimposed component. $\mathrm{TOV_f}$, the fundamental component of TOV can be calculated from the following equation (in per unit) for complete loss of full d.c. load at rated a.c. bus voltage. ($\mathrm{TOV_f}$, replaces the previously used term DOV (dynamic overvoltage)):

$$\mathrm{TOV_f} = [1 + 2Z_e(P_d \cos\phi\varsigma + Q_d \sin\phi) + Z_e^2(P_d^2 + Q_d^2)]^{1/2}$$

where Z_e is the a.c. system effective impedance which includes station shunt and filter capacitors, $\phi\varsigma$is the damping angle of the a.c. system impedance, P_d is the d.c. power, which is positive for the rectifier and negative for the inverter, and Q_d is the reactive power consumed by the converter, which is positive for the rectifier and the inverter. All quantities (except the angle ϕ) are in per unit of P_d. From Section 32.12.3 it can be seen that the effective a.c. system impedance is $1/[Y_m + Y_c] = \text{SCR}$ for $P_d = 1.0$ p.u.

D.c. load rejection resulting, for example, from a d.c. line fault can lead to excessive temporary overvoltage at converter station bus-bars because the connected filters and capacitors represent a substantial surplus over the reduced var demand of the converter. The worst disturbances for a converter at one end of the d.c. link are usually due to faults in the a.c system at the other end. This normally results in reduced power flow, the exact nature of it being dependent on whether the converter is operating as a rectifier or an inverter.

Faults in the a.c. system, at the remote rectifier end, cause low d.c. line voltage giving low d.c. power flow. The d.c. link is usually designed to attempt to ride through such temporary system faults. Usually the rectifier-end fault situation is communicated to the receiving end by a telecommunication link, but, in order to discriminate against temporary, rapidly cleared rectifier-end faults, no protective switching action is initiated at the inverter terminal for a few hundred milliseconds. Subsequently the converter is blocked and the filter and other capacitors are switched off to reduce the overvoltage.

The transient load rejection effect on the a.c. system of such inverter blocking should be investigated for the a.c. network conditions of minimum fault level at which full power could be transmitted. As an example, a theoretical fundamental component of temporary overvoltage, TOV_f of about 1.35 p.u. at power frequency, may be expected at the inverter station a.c. bus-bars for a system short circuit level of three times the d.c. power level (short-circuit ratio (SCR) = 3, see Section 32.12), if the a.c. system impedance were a pure reactance. In practice a lower overvoltage will be more usual, due to the resistive effect of network loads. In many urban networks such an overvoltage, even for a fraction of a second, may not be acceptable.

For a rectifier, faults in the remote receiving system have the same effect as a d.c. short circuit. The reactive power demand of the rectifier is then determined by the low voltage current limit (l.v.c.l., see Section 32.10) setting of its current controller. A typical value of l.v.c.l. of 0.3 p.u. would result in reduced var demand compared with the levels provided by the connected filter capacitors, so that there would be some surplus var generation and the rectifier a.c. terminal voltage would rise.

The reduction of such temporary overvoltages due to load rejection cannot generally be achieved by d.c. converter controls themselves. Fast responsive equipment capable of absorbing a large part of the filter Mvar temporarily until the filters can be disconnected is required. Synchronous compensators can only influence this in accordance with their effect on the system SCR, on the basis of x''_d for two or three cycles and x'_d for somewhat longer. If a.c. system requirements demand faster control of temporary overvoltage, this can be achieved by switching actions, variable static var compensators may have to be used.

Figure 32.48 compares the calculated dynamic overvoltage for a converter station with and without the presence of static compensators following a fault leading to total d.c. load rejection. To achieve the desired control of the temporary overvoltage, the saturated reactor compensator was designed for reactive absorption overcurrents of three times its rated current. In *Figure 32.48* the main parameters of a thyristor controlled reactor (TCR) or of a saturated reactor (SR) compensator required to achieve this rating are summarised.

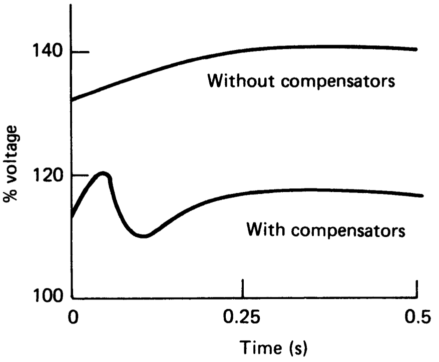


Figure 32.48 Overvoltages due to h.v.d.c. link blocking

A very large bank of non-linear zinc oxide type resistors could be considered as an alternative as a voltage limiter of high energy capability suitable for repetitive current surges lasting up to 0.5 s or more. At present this would be expensive. (See also Chapter 41.)

32.10 Control of HVDC

32.10.1 Summary of HVDC controls

Figure 32.50 shows the general arrangement of controls for a typical h.v.d.c. bipole. Progressing backwards from the converter valves these are as follows.

Valve firing circuits: these have some protective and monitoring functions, but in normal conditions they act only as an interface between the pole controls and the valves.

Pole controls: these are the main controls responsible for changing the firing angles of converters in response to various control loops, and are fast (response time typically 5 ms to 50 ms).

Tap-changer controls: these are relatively slow (about 5 s per step) and act only to optimise working conditions of the converters for minimum reactive power, losses, and harmonics.

Master control: this is at one station only and controls the whole bipole in response to a power order; it has a slower response than pole controls.

Telecommunication: this transmits current order digitally to the remote station, with a check-back signal on a return channel. It may also carry supervisory and other signals.

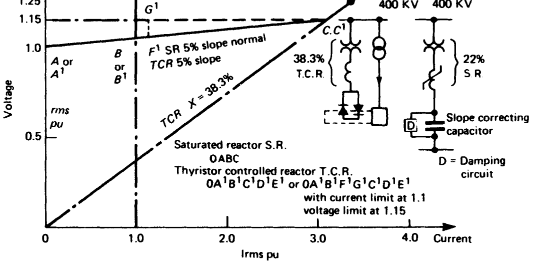


Figure 32.49 Voltage–current characteristics of static variable reators in the overload region

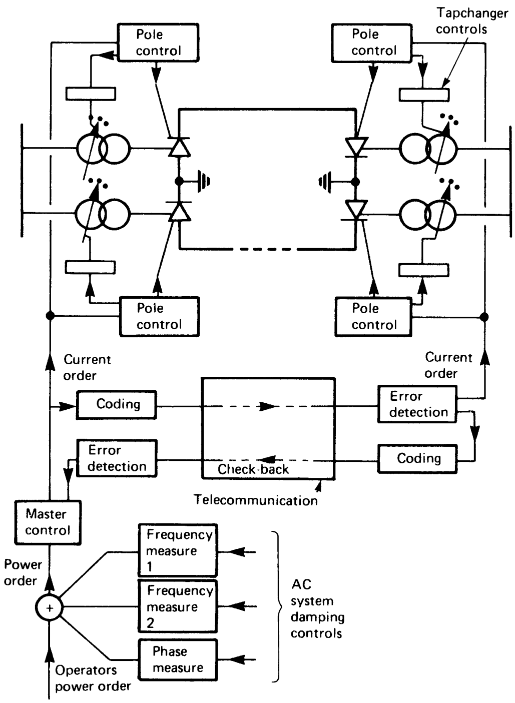

Tapchanger controls
Pole control
Pole control
Pole control
Pole control
Current order
Current order
Coding
Error detection
Error detection
Coding
Check-back
Telecommunication
Master control
Power order
Frequency measure 1
Frequency measure 2
Phase measure
AC system damping controls
Operators power order
+

Figure 32.50 Control hierarchy for a bipole h.v.d.c. link

A.c. system damping controls: measurement and feedback of various a.c. system quantities to provide damping to one or both a.c. systems, normally acting via the master control. These are discussed individually in Section 32.11.

32.10.2 Pole controls

Figure 32.51 shows typical d.c. voltage–current characteristics for the rectifier and inverter respectively as seen from the d.c. line. This diagram effectively summarises the various control loops used in the converters as follows.

Rectifier: AB is from a voltage limit loop. BC is from a minimum alpha limit at about $\alpha = 2^\circ$, CDE is from a constant current loop at a current equal to current order. FG is the low-voltage current clamp (l.v.c.c.) characteristic and also acts via the current loop, switched to a fixed current order of 0.3 p.u. if d.c. voltage falls below 0.3 p.u.

Inverter: HD is from a constant-extinction angle (γ) loop, typically at $\gamma = 15^\circ$ to 18°; KD is a 'current-error characteristic' to ensure stable operation near normal voltage.[16] KL is from a constant current loop, at a current lower than current order by the current margin (0.1 per unit). LM is a low-voltage current limit (l.v.c.l.) characteristic obtained by compounding a current loop with measured d.c. voltage.

In normal operation the working point is at the crossover point D of the two characteristics, corresponding to the inverter in constant gamma (γ) control determining d.c. voltage, and the rectifier in constant current control determining d.c. current at the ordered value.

If rectifier a.c. voltage falls relative to inverter a.c. voltage, then BC falls, sweeping the working point along DK and perhaps down KL, in a stable manner, i.e. the inverter takes overcurrent control at a current below order by up to 0.1 p.u. It is not satisfactory to omit slope KD because the

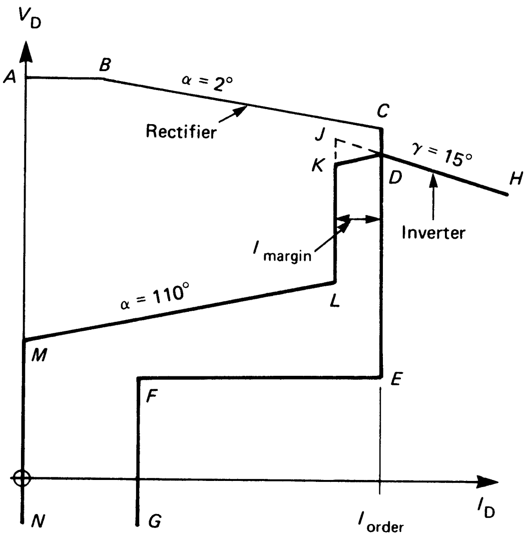


Figure 32.51 Typical V_d/I_d characteristics of rectifier and inverter stations

transition is then too abrupt. The master control will normally correct the resulting power error by increasing current orders.

32.10.3 The phase-locked oscillator control system

The type of converter control used in the pole controls of all modern h.v.d.c. schemes is the phase-locked oscillator control system[17] giving nominally equidistant pulse firing. The principal reason for this is its freedom from harmonic instability[18] when the converter is connected to a relatively weak a.c. system, as generally applies when the d.c. power forms a substantial part of the a.c. system infeed.

Figure 32.52 shows the principles of the phase-locked oscillator control in simplified form. Its basic components are a voltage-controlled oscillator, and a 12-stage ring counter (for a 12-pulse converter) which feeds the 12 pulses to the respective converter valves.

The oscillator comprises an integrator, comparator, and short pulse generator, and normally runs at 12 times supply

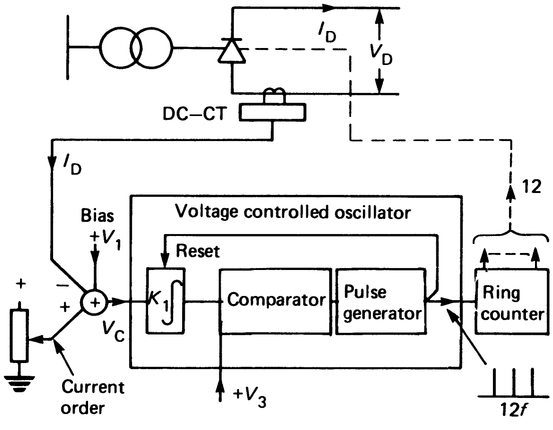

I_D
V_D
DC–CT
I_D
12
Bias
$+V_1$
Voltage controlled oscillator
Reset
K_1
Comparator
Pulse generator
Ring counter
+
–
+
+
V_C
Current order
$+V_3$
12f

Figure 32.52 Basic phase-locked oscillator control system

frequency, hence valve firing pulses are normally once per cycle per valve, at an accurate 30° spacing. The oscillator input has a fixed bias V_1, plus an error signal equal to the difference of the measured quantity for the particular loop (d.c. current as shown) and an order signal. The system settles as for an integral control system, with zero steady-state error, and with firing pulse times (firing angle α) at the correct values to obtain this.

In its recent form, the feedback signal is applied 'raw', i.e. without extra smoothing lags, giving fast response and good stability. This applies not only to approximately smooth signals such as d.c. voltage or current, but also to mark/space type signals such as α or γ. Because this method integrates the input signal, it correctly controls the *mean* value of the controlled quantity. Some methods (including some forms of digital control) do not have this property and tend to respond more to sampled values of the measured quantity near to the normal firing times; this gives excessive response to ripple and sudden disturbances, with poor stability, unless extra smoothing lags are added, which also degrades stability.

All practical control systems must be *multi-loop*, as seen from *Figure 32.51* since the operating mode must change according to system conditions at each break. As an example, operation changes between alpha control ($\alpha = 2°$) to constant-current control at point C of the rectifier characteristic.

Figure 32.53 shows a recent development of multi-loop control based on multiple oscillators.[19] As an example this is shown for three loops, respectively for d.c. current, γ and α. (A scheme may use ten or more loops.) Each loop has an individual integrator and comparator, and is coupled via OR and AND gates to a common pulse generator; the latter resets all integrators together, and also operates a common ring counter (not shown) as before. This gives extremely fast and precise handover between modes, without the extra smoothing lags and amplifier desaturation delays which characterise systems which use handover in the pre-oscillator portions of the controls.

The most modern controls include γ-*balancing* circuits, which equalise γ values even in conditions of a.c. system unbalance; this gives the control the combined advantages of the equal pulse-spacing method (stable operation on weak a.c. systems) and the individual phase control method (maximum real power and minimum reactive power during unbalanced a.c. system conditions).

Flux-control circuits are included to prevent core saturation instability in converter transformers, and also to reduce the effects of even harmonics from the a.c. system or of a small fundamental frequency component on the d.c. line.

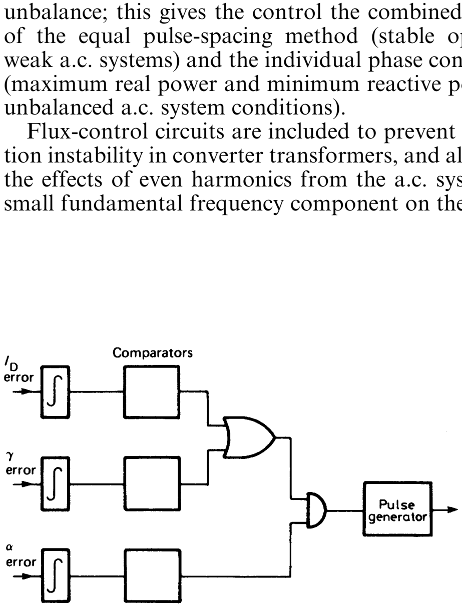


Figure 32.53 Loop control system

32.10.4 Tap-changer controls

At the inverter station the converter transformer tap-changers are automatically controlled from d.c. voltage to obtain rated d.c. voltage.

At the rectifier the tap-changers are controlled from measured firing angle to hold it within set limits, say 8–18°. The lower limit is as low as possible to give minimum reactive power, losses, and harmonics, while preserving a small control range in case of small a.c. voltage dips. The upper limit is also as small as possible without giving tap-changer control instability [i.e. (cos 8°–cos 18°) > tap voltage stepper unit].

The tap-changers are conventional, and rather slow, but any errors in power before they reach steady values are corrected by the master control.

32.10.5 Master control

The principal duty of the master control is to adjust the current orders of the two poles so that measured power is equal to a power order. This is done by direct measurement of total bipole power from the d.c. quantities, subtracting it from the power order signal, and integrating the resulting error signal to generate the 'master current order', which is then applied to both poles in both stations.[20] It is therefore a feedback loop, at a higher level than pole controls, but rather slow (settling time of say, 100–300 ms, depending on telecommunication rate); the telecommunication gives a direct delay when the remote station happens to be controlling current, and an indirect delay even when the local station is in control because it may then have to wait for the checkback signal (see Section 32.10.6 below). However, the master control contains various other functions including special forms of limit; an example of the latter is in recovering from a zero-voltage fault, when it would otherwise try to generate an infinite current order.

The master control will normally control to constant power in the presence of changes of a.c. voltages and other disturbances, subject to maximum current limits applied locally in pole controls. This applies even if the inverter temporarily takes over current control; the master current order will then be temporarily above the actual current.

Where the rectifier and inverter stations are a long distance apart, a telecommunication system (see below) is required to send current order to the remote station. Although schemes have been operated without telecommunications, optimum performance cannot be obtained without it. Generally the master control is at one end and the remote station acts only as a 'slave' without any master control function. It is not in principle important whether the master control is in the rectifier or the inverter station. Both have been used in practice, though a.c. system damping controls (see Section 32.11) may influence the choice.

The power order signal input will normally be set by an operator; either locally or in a remote control centre. It may be modified by a.c. damping controls.

Back-to-back schemes have both converters in one station, hence do not require a long-distance inter-station telecommunication, but their control and performance is otherwise similar to that for long-distance schemes (except that their response is much faster due to the omission of d.c. line capacitance and telecommunication delays).

32.10.6 Telecommunication

The main duty of this is to send a current order signal to the 'remote' station. Because practical telecommunication media are subject to noise and interference, signals are always sent in digital (binary) form, with error checking code.[21] The coding must generally be much more powerful than normally used for sending computer data, because of the serious effect of an undetected error on power flow.

Detected errors are less of a problem since they are arranged to 'freeze' pole current orders in both stations at the last correct value, using a check-back channel.[21] The latter must itself have error-detection coding; this is usually shared with other signals, e.g. supervisory signals or a.c. network damping signals. Since normal operation is steady state, or with very slow change, the effect of brief errors is then negligible. For permanent telecommunication failure the controls are transferred to a simpler manual mode.

Various media are possible for telecommunication. Direct wires via private land lines or the telephone system have been used, but usually only for emergency operation because of their high cost and poor reliability. Power-line carrier is probably the cheapest method. It would require two or three repeaters for say 900 km of overhead line. Interference generated by the converters, penetrating via the d.c. reactors, is greater than normally seen on a.c. systems, and requires special measures. Microwave radio has been used on several schemes. It requires rather short repeater intervals of about 40 km since line-of-sight transmission is required, but its bandwidth is very large and can accommodate many telephone and even television channels in addition to the h.v.d.c. control requirements.

Tropospheric scatter radio is another possible method, requiring repeater intervals between 200 and 500 km depending on transmitter power. It is subject to heavy fading, and requires multiple redundant channels and multiple antennae to give an acceptable basic signal error rate. Optical fibre is being used more and more for medium-distance telecommunications, and has been installed experimentally inside power conductors and overhead earth wires. It is almost immune from interference and has wide bandwidth but would require repeaters about every 100 km at present, though future improvements to this are expected.

The choice of telecommunication will depend on cost and the policy of the utility. Generally a baud rate of 2400 (block period about 50 ms) will be adequate to obtain sufficient bandwidth for damping a.c. system swings up to about 1 Hz.

32.10.7 Performance examples

Figures 32.54–32.57 are oscillograms taken from simulator tests showing the behaviour of a typical h.v.d.c. link for a.c. and d.c. faults. These are all for the relatively onerous case of a weak receiving system (effective short-circuit ratio 2.4, impedance angle 75°), and at rated power.

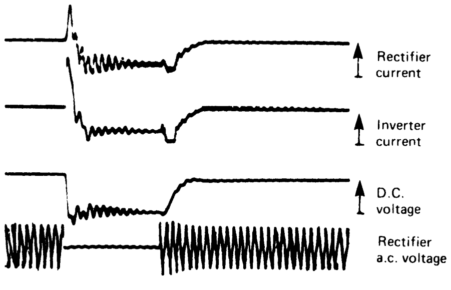


Figure 32.54 Three-phase 100% fault at the inverter

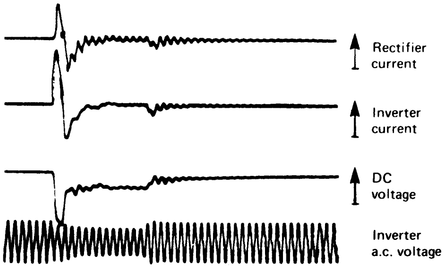


Figure 32.55 One-phase 40% fault at the inverter

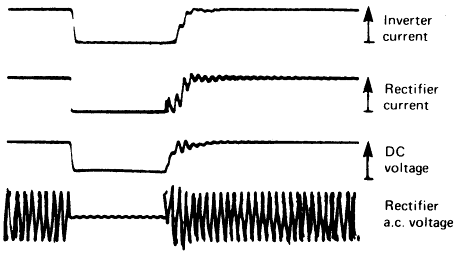


Figure 32.56 Three-phase 100% fault at the rectifier

Figure 32.54 is for a three-phase fault close to the inverter a.c. bus-bar. There is an initial current surge, then the current settles to 0.3 p.u. during the fault, at zero d.c. voltage. After fault removal, most of the recovery occurs in about 100 ms, but the last part of the recovery is relatively slow because of the effect of magnetising inrush current of converter transformers on a.c voltages.

Figure 32.55 is for a more remote fault in the a.c. system fed by the inverter, which reduces a.c. voltage by about 40%. There is an initial commutation failure, but commutation is then restored at reduced voltage until the fault is removed, with recovery as before.

Figure 32.56 shows a three-phase fault at the rectifier. D.c. current falls to zero during the fault, and recovery afterwards is similar to *Figure 32.54*.

Figure 32.57 shows a fault, e.g. due to lightning, about halfway along a d.c. overhead line. The fault reduces d.c. voltage to an average of zero, with line oscillations and a rectifier current surge. The rectifier starts to reduce current to 0.3 p.u., but a line fault detector relay operates at 30 ms after the fault, forcing complete (temporary) shut-down; by about 40 ms both rectifier and inverter currents are at zero.

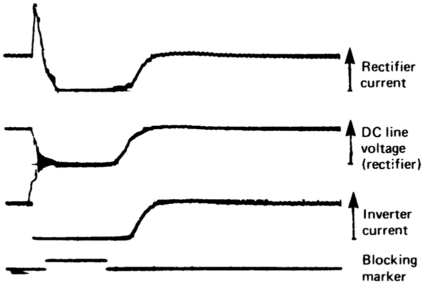


Figure 32.57 D.c. line fault

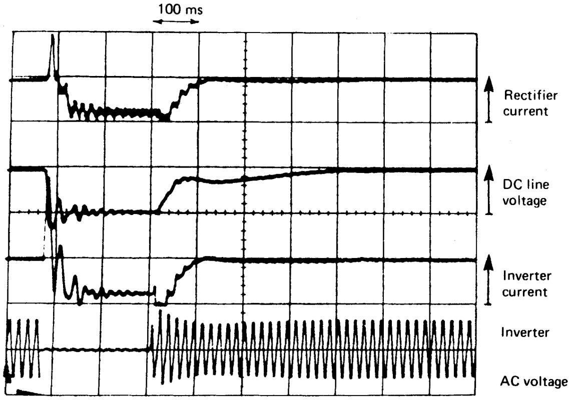


Figure 32.58 Three-phase fault at the inverter (ESCR = 1.5)

After waiting about 100 ms for transient line currents to decay, and the fault arc to de-ionise, power is then restored to normal. This simple action appears to be sufficient in most real cases, though if necessary operation can be more complicated, with several re-starts.

To illustrate several of the points discussed above, *Figure 32.58* shows an oscillogram from simulator tests showing the behaviour of an h.v.d.c. link for a three-phase fault at the inverter, similar to *Figure 32.54* but with a very low value of effective short circuit ratio (see Section 32.12) ESCR = 1.5 ∠75°. This gives good performance, only slightly worse than for *Figure 32.54* (SCR = 3), with fast recovery and good stability without using special controls. However, in practice, for an a.c. system so weak, the excessive a.c. voltage changes to which consumers may be exposed following, say, a d.c. line fault, may not be acceptable. In such a case some form of fast reactive control, such as a static var compensator, may have to be used.

32.11 A.c. system damping controls

The normal duty of an h.v.d.c. link is to transmit power at a preset level set, usually, by an operator. However, h.v.d.c. is a very powerful control device as it is capable of changing transmitted power in a controlled manner, rapidly and by a large amount and thereby influencing the a.c. system transient performance significantly. Three main system conditions are discussed below.

32.11.1 D.c. link supplies power from dedicated generators or from a very strong system to a small system[22–24]

Figure 32.59 shows a block diagram of the Nelson River h.v.d.c. transmission scheme in Manitoba, Canada. The scheme is essentially a 900 km long, 'asynchronous' interconnection from remote hydro generation to an urban industrial load area. The d.c. link controls are arranged to produce a current order in response to four principal signals:

(1) Steady-state power order from the system dispatch control.
(2) Transient power order in response to the receiving end load area frequency changes; this provides receiving frequency control by the d.c. link of the form usually carried out by governors with defined droop characteristics.

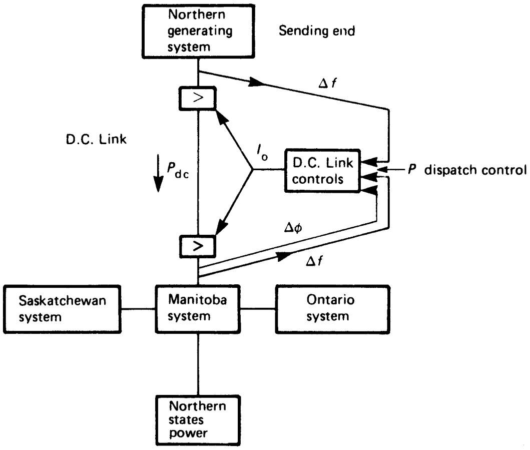


Figure 32.59 Block diagram of the Nelson River transmission scheme

(3) Transient power order in response to frequency changes at the sending end; this assists the hydrogenerator governors in controlling the sending end frequency.
(4) Transient power order in response to the inverter a.c. bus voltage phase angle changes; this signal modulates d.c. power to provide swing damping within the local network and also between the local (Manitoba) power system and the three neighbouring systems of Ontario. Saskatchewan and the USA Northern States Power Pool. The power change required for swing damping is usually quite small, typically less than 10% of rated h.v.d.c. power.

Figure 32.60 illustrates the damping effectiveness of the frequency controls on the speed (frequency) of the receiving end (Manitoba) system equivalent machine and on the sending end generator at the Kettle generating station. *Figure 32.61* shows site measurements of the effect on system frequency oscillations of an inadvertent loss of h.v.d.c. system damping control; before this event the system frequency is well damped, but after the loss of damping, sustained oscillations at about 0.3 Hz are to be observed on the nominal 60 Hz frequency trace. (These appear rather small on a frequency trace, but power oscillations in tie-lines are much larger in proportion.) In this scheme the damping

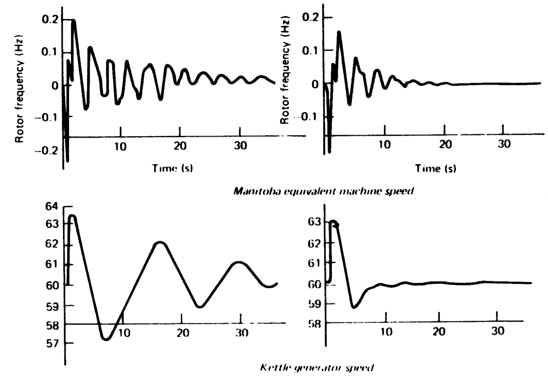


Figure 32.60 Nelson River transmission scheme: rotor frequencies following on disturbance without (left) and with (right) d.c. link system damping controls

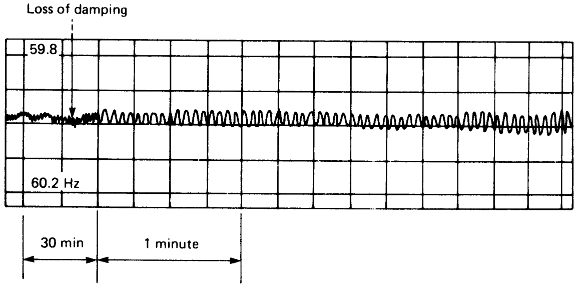


Figure 32.61 Site measurements showing the effect of loss of h.v.d.c. damping controls on a.c. system frequency

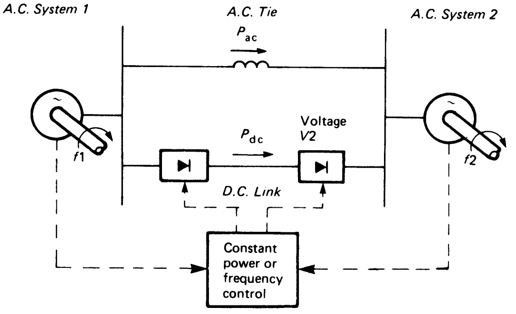


Figure 32.62 Block diagram of simple parallel a.c. and d.c. transmission

signal (4) is based on measurement of the rate of change of the absolute phase of the inverter a.c. bus voltage using a phase-locked oscillator. Direct derivation of a network frequency signal would require filter circuits which generally reduce the sensitivity and accuracy of the measurement.

Signal (4) is effectively 'a.c. coupled' having no effect in steady state; signals (2) and (3) may be either of this type or may be d.c. coupled to influence steady-state operation. Various other signals are also possible for particular systems, for example power in a tie-line.

The use of a damping signal such as (4) above (which is relatively fast) may influence the location of the master control. The average response to modulation at the master control is not much affected by its location (at rectifier or inverter stations) but if it is remote from the source of the fastest damping signal then an extra delay will occur in its loop because of the necessity for sending the damping signal via telecommunication; this is undesirable, hence the master control should be at the end where the highest natural frequency (e.g. 1 Hz) is to be damped. This may require the operator's power order to be sent via a telecommunication channel, but this is rather easy since a relatively slow channel can be used.

32.11.2 D.c. link connecting two systems which are not synchronised but are of similar size

Care must be taken that the requirements of one system do not adversely affect the other. For example, two systems of similar size may have the same dominant natural frequency, say, in the region of 0.5–1 Hz. In this case it will generally be necessary to provide damping signals from both a.c. systems, so that a disturbance in one system is shared between the two; this will slightly disturb the healthy system, but in a heavily damped manner. (Taking a damping signal from one a.c. system only will cause disturbance in the healthy system which may not be damped by the d.c. link.) It should be noted that even if the converter is operated at its maximum rating, a degree of damping can be achieved by power reduction modulation only.

32.11.3 D.c. link connecting two parts of an a.c. system or two separate systems having also a parallel a.c. link

Figure 32.62 shows a simple diagram illustrating parallel a.c. and d.c. transmission.

The Pacific Intertie on the USA West Coast is an example of an h.v.d.c. link between two a.c. systems which also have long interconnecting a.c. lines. Initially this scheme was operated with constant power control, but in the mid-1970s additional power modulation controls were provided which considerably enhance its use by arranging for the transient stability of the a.c. lines to be improved. In such applications, the amount of power modulation required to achieve beneficial transient stability and swing damping effects is generally small compared with the transmitted power.

There are several examples in Canada, USA and Europe of back-to-back h.v.d.c. links between two large systems where an a.c. interconnection would have been much less advantageous technically and economically. By providing controllable d.c. power flows in the interconnection, even in the event of a major disturbance in either a.c. system, the h.v.d.c. links can help reduce the instabilities, and prevent complete system collapse situations by avoiding the cascading of disturbances which could otherwise occur in uncontrolled a.c. ties.

32.12 Interaction between a.c. and d.c. systems[25–27]

32.12.1 Study of HVDC systems

The relationship between an h.v.d.c. link and the a.c. system to which it is connected can be considered in two categories.

(1) The line commutated converter depends for its operation on being supplied with a reasonable sinusoidal voltage. A distorted voltage, or a significant unbalance of the three-phase system voltage such as occurs during faults, will detract from the essentially symmetrical rectification and inversion processes.
(2) The rectifier takes from its a.c. system both power (*P*) and reactive power (*Q*). The inverter feeds power to its a.c. system but takes the reactive power from it. If the d.c. link is relatively large compared with the a.c. system to which it is connected, any large changes in *P* and *Q* could have significant effects on the system.

A.c. system disturbance can affect the operation of a small converter but maloperation of a small converter will have negligible effects on the a.c. system. However, it is not uncommon for an h.v.d.c. link to supply a large proportion of the a.c. system load so that the loss of d.c. power and associated reactive power changes can have a profound effect on the system.

The effects of h.v.d.c. operation and maloperation can be simulated accurately using loadflow, transient stability and other digital programmes as regards category (2) above and by the use of an h.v.d.c. simulator and digital programmes, such as electromagnetic transient programmes (EMTPs) as regards category (1) above.

An a.c./d.c. simulator is an important tool in the study of h.v.d.c. In particular, its use is important in the development

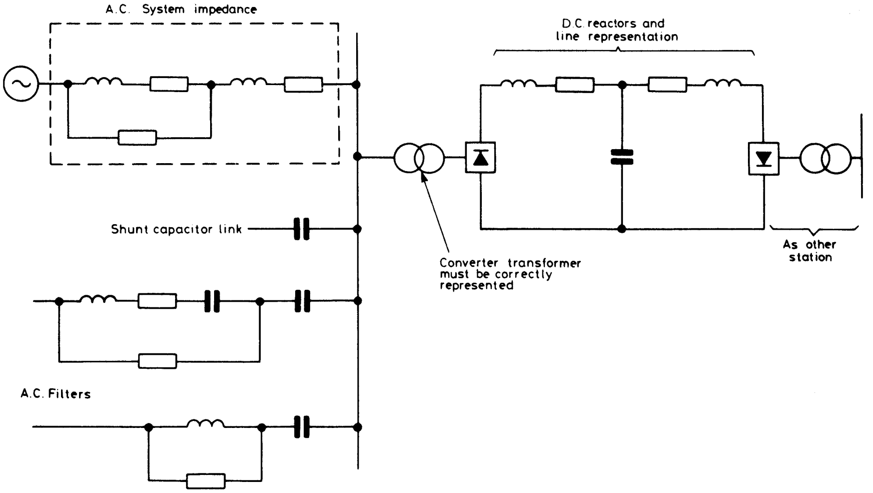

A.C. System impedance
D.C reactors and line representation
Shunt capacitor link
Converter transformer must be correctly represented
As other station
A.C. Filters

Figure 32.63 System representation for h.v.d.c. simulator

of h.v.d.c. controls and in the study of the inverter recovery after system faults, as illustrated in Section 32.10. It is not necessary to represent the a.c. network in detail for these studies. A.c. systems can be represented by a Thevenin equivalent of a constant e.m.f. behind an impedance, as shown in *Figure 32.63*. It is important to represent adequately both the value of the impedance and its damping. A.c. filters should be correctly represented as well as any shunt capacitor banks.

Simulator studies are usually concerned with a time-scale of up to 400 ms. Sufficiently accurate representation will generally be achieved by representing generators by their subtransient reactances, though more accurate representation is sometimes needed. As far as the converter transformer is concerned, it is important to represent it correctly as the major contribution to the distortion of the a.c. voltage is the transformer in-rush current. It is therefore important to represent not only the commutation reactance but also saturated self and mutual reactances.

32.12.2 A.c./d.c. system strength

The 'strength' of an a.c. system is represented by its impedance and by its mechanical (rotational) inertia. System strength expressed as an absolute numerical value, such as short-circuit MVA is useful only if compared to the power and reactive power values of its load. The short-circuit ratio (SCR), defined as the ratio of the system short-circuit level MVA to the a.c. power MW, has been used to indicate system strength.

A system consisting of a number of generators and transmission lines representing a network has more than one value of system strength, because loads connected at different locations in the same interconnected network will see different values of the system impedance and the loads themselves will have different values; changes by switching generators, lines, transformers, etc., will also occur from time to time.

32.12.3 Short-circuit ratios

32.12.3.1 Short-circuit ratio

The higher the system impedance and the lower the system damping for a given h.v.d.c. inverter, the greater the effect of the inverter maloperation on the a.c. system. It has become customary to refer to relative sizes of the a.c. system and the h.v.d.c. power by the term short-circuit ratio (SCR)

$$\mathrm{SCR} = \frac{S}{P_\mathrm{d}} \qquad (32.10)\Leftarrow$$

where S is the minimum a.c. system short circuit MVA at which the maximum d.c. power P_d is transmitted.

S is calculated similarly as in Section 32.12.1 using Thevenin equivalents. If synchronous compensators are used in the converter station, the effect of their reactance should be included in S, as shown in *Figure 32.64*.

32.12.3.2 Effective short-circuit ratio

A.c. harmonic filters must be used in all schemes. At fundamental frequency the filter acts practically as a shunt capacitor. Shunt capacitors increase the impedance at fundamental frequency of essentially inductive systems and to allow for this the term effective short-circuit ratio (ESCR) is used. In admittance form this is defined as for SCR but it is the admittance of the a.c. system plus all filters and capacitor banks additionally connected to the a.c. bars.

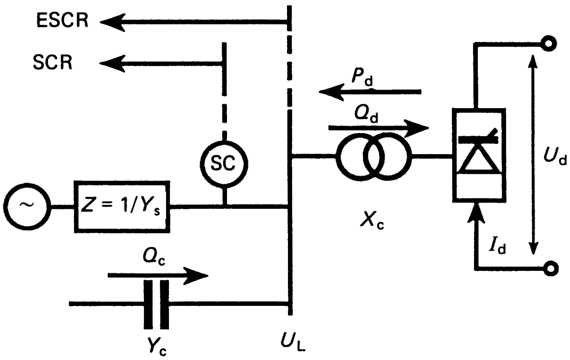


Figure 32.64 Simplified representation of a d.c. link feeding an a.c. system with shunt capacitors (C) and synchronous compensators (SC), if any, at converter station bus-bars

Note that both SCR and ESCR always have an angle as well as magnitude. Thus, for example, for a system with SCR = 3 ∠ 80°⇐ the addition of 0.6 p.u. of capacitors plus filters gives ESCR = 2.4 ∠ 78°. If the simplification of a system impedance angle of 90°⇐is assumed, then, in short- circuit MVA form

$$\mathrm{ESCR} = \frac{S - Q_c}{P_d} \qquad (32.11)\Leftarrow$$

where Q_c is equal to the sum of the fundamental frequency MVA of a.c. filters and of any additional shunt capacitors connected to converter station terminals.

Figure 32.64 indicates the definition of SCR and ESCR.

SCR: the value of the admittance Y at the fundamental frequency, on the base of the rated d.c. power at the rated a.c. voltage.

ESCR: is defined as SCR except that the admittance includes the admittance of the capacitor, $Y = Y_s + Y_c$

It should be noted that because a.c. systems are largely inductive, it is the change of reactive power which is mainly responsible for the effect of converter behaviour on the a.c. network voltage. Most schemes in the past were designed with transformer reactance of the order of 20% or more to limit the thyristor fault current. The availability of large thyristors and the pressure of cost of losses, are responsible for some schemes being designed using a larger thyristor than necessary in order to reduce the losses. An additional benefit is that it is possible to design the transformer for any suitable reactance down to, say, 11%, as the oversized thyristor has correspondingly larger surge current capability.[28]

Lower converter transformer reactance can bring a number of advantages, including:

(1) converter reactive power consumption (Q_d) will be reduced;
(2) a.c. filters and any additional shunt capacitors are normally designed to supply at least all converter reactive power (the amount of shunt-capacitor compensation can be reduced, reducing the cost and increasing ESCR);
(3) temporary overvoltages will be reduced, due to smaller shunt capacitors; and
(4) rating of equipment such as surge arresters may be reduced.

SCR and ESCR represent the a.c. system reasonably accurately for the short time-scale considered. However, the SCR concept should be used mainly to get a 'feel' of the system, which should be followed with adequate studies.

32.12.3.3 QESCR

Two converters may be rated for the same d.c. power, but their transformers may be designed to have different reactances and they may be operating at different values of αςand γςand, therefore, they would consume different amounts of reactive power. Both SCR and ESCR are referred to d.c. power and do not take into account the converter reactive power, Q_d. Q_d can be partly taken into account by referring the SCR to the sum of the power and the reactive power ($P_d + Q_d$). Thus,

$$\mathrm{QESCR} = (S - Q_c)/(P_d + Q_d) \qquad (32.12)\Leftarrow$$

32.12.3.4 Operational short-circuit ratios (OSCR, OESCR and OQESCR)

For operation at conditions other than at rated load, often at loads lower than rated, the corresponding operational short-circuit ratios (OSCR, OESCR and OQESCR) must be used: appropriate minimum short-circuit capacity of the a.c. system, actual values of shunt capacitors and power level, not the rated power, must be used.

32.12.4 Voltage/power curve

A simplified representation of two a.c. lines feeding a load is given in *Figure 32.65*(*a*) and the well known voltage/power curve of the simple a.c. system is drawn in *Figure 32.65*(*b*). As the load current increases, the power rises until it reaches a maximum. After that point, due to the increasing line voltage drop, a sharp decrease in power occurs. The operating point is normally at a power level sufficiently smaller than the maximum power point to avoid voltage collapse due to, say, a temporary load increase or trip of a line. The curve in *Figure 32.65*(*b*), for illustration purposes, was calculated assuming operation at the point of maximum power; it was further assumed that the load power factor is unity and the small a.c. line resistance was neglected. The fundamental component of the temporary overvoltage (TOV$_f$) for total load rejection for assumed conditions is equal to $\sqrt{2}$, as can be seen from *Figure 32.65*.[29]

A converter, rectifier or inverter, behaves as a static a.c. load. At the sending end, the rectifier can be represented as a P and Q load, with P positive and Q also positive (lagging). At the receiving end, the inverter delivers power, but consumes vars; P is negative, while Q is positive as for the rectifier. If, for approximate calculations, the small a.c. line resistance is neglected, the voltage/power curves of an a.c. load, a rectifier and an inverter are identical, for the same values of P and Q. Moreover, d.c. converters are normally compensated to operate near unity power factor and, therefore, the curve

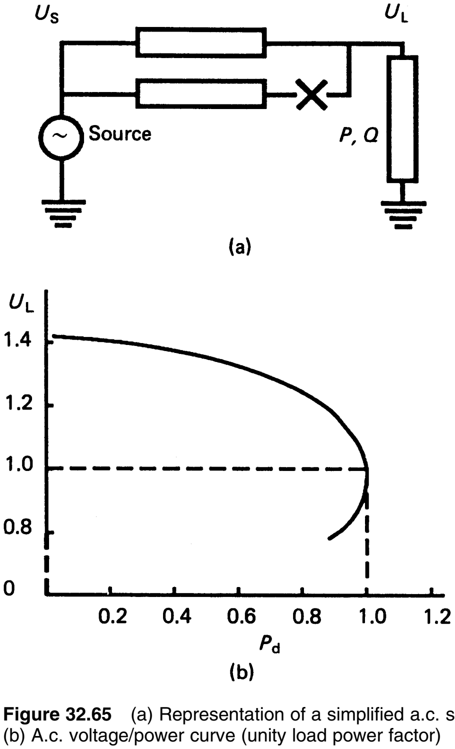


Figure 32.65 (a) Representation of a simplified a.c. system. (b) A.c. voltage/power curve (unity load power factor)

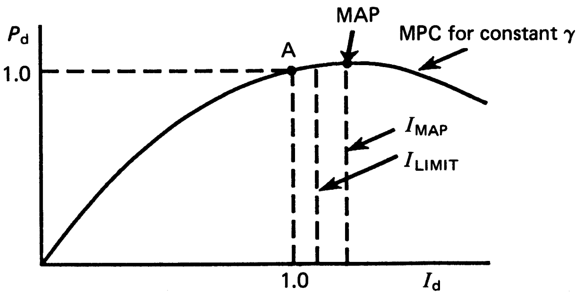


Figure 32.66 D.c. power/current curve for minimum $\gamma\varsigma$

of *Figure 32.65(b)* is in general relevant to h.v.d.c. operation near the point of maximum power, as discussed later.

32.12.5 Maximum-power curve

D.c. transmission is carried out at maximum designed d.c. voltage, for reasons of economy; the transmitted power is regulated by variation of the d.c. current. For this reason it is customary to plot d.c. power against d.c. current as shown in *Figure 32.66*. Maximum d.c. voltage is obtained by operating the inverter at the minimum permissible commutation margin angle (γ), as discussed in the Section 32.3.3. The power/current curve for operation at the constant minimum $\gamma\varsigma$ is termed the maximum power curve (MPC). This curve is generated by increasing and decreasing d.c. current from the initial conditions at the operating point A. Usually the starting point corresponds to the rated d.c. current and d.c. voltage at the rated a.c. voltage. MPC corresponds to transient conditions after the d.c. current was changed. The a.c. voltage is not controlled, but drops as the d.c. current increases; automatic voltage regulators, tap-changers, shunt reactors and capacitors are assumed fixed, which is what happens in practice for the first 100–300 ms after the onset of a transient disturbance. For definition of MPC, it is assumed that fast voltage control by thyristor-controlled rectifiers or saturated reactors, which would be comparable in speed to d.c. current changes, are not used.

No power greater than that corresponding to MPC can be obtained, unless the a.c. voltage feeding the converters is increased. On the other hand, any power can be obtained below the MPC by increasing γ. Similar MPC curves can be drawn for rectifier operation by replacing $\gamma\varsigma$ with α.

As indicated in *Figure 32.66*, a d.c. link is provided with a current limit, acting at the rectifier. This means that d.c. converters can operate closer to the maximum power point without risking a voltage collapse (see *Figure 32.65(b)*). In fact, a d.c. link can be made to act as its own thyristor-controlled rectifier (see Chapter 28) and respond to an a.c. voltage signal to reduce its power, and therefore its reactive power, and so prevent excessive a.c. voltage reductions.

32.12.6 Maximum available power

The maximum value of the MPC has been termed the maximum available power (MAP). The value of MAP, for a given a.c. system impedance (SCR), depends on the converter station rated reactive power and, therefore, it is a function of the commutating reactance x_c, usually equal to the converter transformer reactance, the value of minimum $\gamma\varsigma$ (or α) and the amount of shunt capacitance of the station.

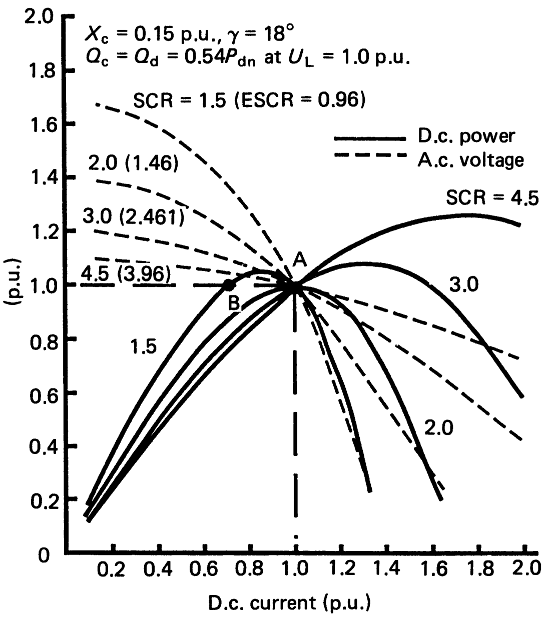


Figure 32.67 Variation of inverter a.c. terminal voltage and power with d.c. current

MPC curves of an inverter of given characteristics ($x_c = 15\%$, $\gamma\varsigma = 18°$; $Q_c = Q_d = 0.54P_d$ at $U_L = 1.0$ p.u.; see *Figure 32.64* for interpretation of the symbols) for four different values of SCR are plotted in *Figure 32.67*. These curves assume that the rectifier will not cause limitation of power demanded by the inverter. However, it should be noted that similar MPC curves apply for the rectifier operation. Depending on the strength of the sending-end a.c. system and on the design of its a.c. to d.c. converter system, i.e. on the voltage control and on the value of the operating α, the rectifier may transiently impose limitation on d.c. power. Therefore, the design of the sending and receiving a.c. to d.c. converter systems must be co-ordinated.

32.12.7 Classification of the a.c./d.c. system strength

An a.c./d.c. system has to be designed to be able to deliver the specified power for the specified a.c. system conditions, including outage situations. Loss of a part of a system or of a transmission line will increase the impedance of the a.c. system. The SCR of the simple system shown in *Figure 32.65* (*a*) may be assumed, for example, to reduce from 3 to 2 with one line tripped with consequent reduction of maximum power, as shown in *Figure 32.68*.

The relationship between MAP and the required power for specified system conditions is used to classify the a.c./d.c. system strength.

32.12.7.1 High SCR a.c./d.c. system

The strongest a.c./d.c. system of the four considered in *Figure 32.67* is the one corresponding to SCR = 4.5 and having MAP = 1.28 p.u. of rated power. With this much power margin, MAP is unlikely to get reduced for expected outage conditions below the rated power and such a system is termed a *high SCR a.c./d.c. system*.

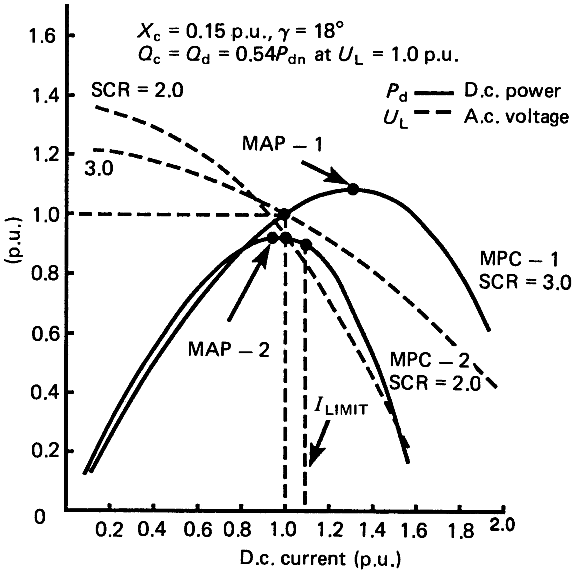


Figure 32.68 A.c./d.c. system: low SCR power and a.c. voltage curves; sudden change of SCR from 3 to 2

32.12.7.2 Low SCR a.c./d.c. system

For the system represented by SCR = 3 in *Figure 32.67*, MAP = 1.08 of the rated power, and it is likely that MAP will reduce below the rated power for some outage conditions. For the example shown in *Figure 32.68* MAP is reduced to 0.96 p.u. of rated power at point A, following the trip of one line. This intermediate condition is termed a *low SCR a.c./d.c. system*. As described in, Section 32.10.5, the current of a d.c. link is automatically adjusted so that the measured power is equal to the power order. If the power order has a higher value than MAP, then an increase of current beyond I_{MAP} would result in power reduction, because the line voltage drop increases at a higher rate than the current increase due to the increased inverter reactive power consumption. In general, there are three possible ways to deal with such outage conditions.

(1) For operation beyond MAP, change from power mode of control to constant-current control, and accept operation at a lower power level until a.c. system conditions are restored. The disadvantage of this approach is that the transmitted power would vary in sympathy with the a.c. voltage variations and would not be suitable for schemes where the d.c. power is used to respond to an a.c.-system-stabilising signal.
(2) Reduce the power level order and continue to operate in power control mode on the 'stable' part of the power curve, that is at the left side of MAP. This has the advantage that the control of d.c. power would be retained, albeit at a lower power level.
(3) If it is important to maintain full power without waiting for system conditions to improve, automatic switching of additional shunt capacitors should be used to maintain a.c. voltage and, therefore, the power level.

32.12.7.3 Very low SCR a.c./d.c. system

If the normal operating point is on the 'unstable' part of the power curve, i.e. at the right of MAP, where d*P*/d*I* is negative, as for the case of SCR = 1.5 (point A in *Figure 32.67*) then the system is described as a *very low SCR a.c./d.c. system*. (There may appear to be an operating point (B in *Figure 32.67*) on the stable part of the curve on the left of MAP on MPC for SCR = 1.5 for 1 p.u. power. However, examination of *Figure 32.67* for SCR = 1.5 will show that this point corresponds to an a.c. voltage which cannot be utilised as it is much higher than the rated voltage.)

Figure 32.67 shows that the a.c. voltage at total load rejection (TOV_f), $I_d = 0$ for SCR = 1.5, may be in the region of 1.7 p.u., ignoring transformer saturation.

The main criterion in the design of synchronous compensators, when used to reduce the a.c. system impedance as seen by the converter, was the limitation of TOV_f to acceptable values. This inevitably changed a *very low SCR* system to a *low SCR* system and shifted the normal operating point to the left of MAP, i.e. to the 'stable' part of the power curve. However, the availability of zinc oxide arresters has made it possible to control TOV without the need to use a synchronous compensator for that purpose.

There are two possible ways of operating with very low SCR systems, and thus avoiding the need to use the synchronous compensators for the sole purpose of reducing TOV.

(1) A.c. voltage can be controlled by a fast static var compensator (SVC). This can be a satisfactory solution, provided the SVC operation is continuous and faster than the required d.c. power-control loop. Automatically switchable shunt-capacitor banks are needed to keep the current of the thyristor-controlled reactor (TCR) or saturated reactor (SR) of the SVC in the controllable range.
(2) A more economic way is to use the inverter to control the voltage as shown in *Figure 32.69*. Controlling the d.c voltage by varying γ makes d*P*/d*I* positive in the normal operating range (along the line A–B), providing stability in power-control mode. For operation at constant γ, d*P*/d*I* is negative (point B and beyond) and an increase in current would cause a decrease in power; as discussed in Section 32.12.7.2 for transient operation, in that region constant current control mode must be employed.

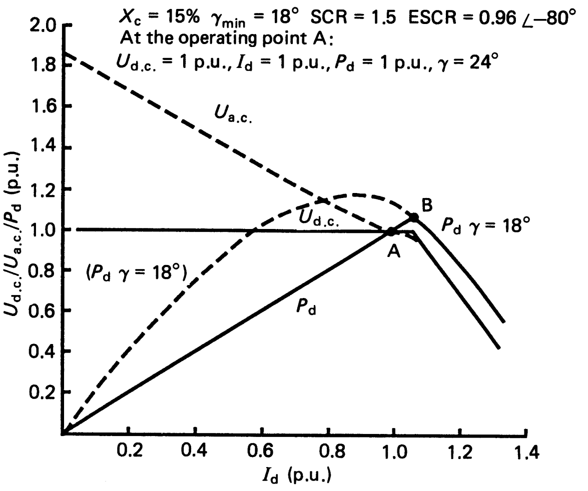


Figure 32.69 Operation with variable γ control to maintain the d.c. voltage constant

The inverter must be designed to operate at γ_0 larger than the minimum in the steady state. In the example in *Figure 32.69*, γ_0 is assumed to be 24° at the rated conditions (point A). By keeping d.c. voltage constant, the a.c. voltage changes near point A are reduced. Capacitor switching and transformer tap-changing are used to control the a.c. voltage in the steady state to keep γ_0 in the required range, between, say, 30° and the minimum value of 18°. The effects of a.c. voltage control and the use of the inverter in TCR mode to control a.c. overvoltages are not shown.

Inverter data for the examples given in *Figures 32.67* and *32.69* are the same, except that in *Figure 32.67* the 1 p.u. power corresponds to $\gamma_0 = 18°$ and in *Figure 32.69* to $\gamma_0 = 24°$. As Q_d is larger for $\gamma_0 = 24°$, the potential value of TOV_f is larger, as can be seen by comparing a.c. voltages on the two figures for SCR = 1.5 for zero current. It should be remembered that, in practice, these overvoltages would be reduced by transformer saturation; the value of the oscillatory component of TOV will depend on the a.c. system damping and the system resonant conditions.

By increasing the current from 1.0 p.u. and allowing γ_0 to reduce from 24° to the minimum value of 18°, power will be increased to the maximum value of 1.07 p.u. More power could be obtained only by restoring the a.c. voltage as discussed above.

The method of operation described has been employed in a number of recent d.c. schemes. To summarise, the most economic operation is at the minimum γ_0 (minimum cost of equipment, minimum losses, minimum generation of a.c. harmonics, and minimum consumption of vars), which can be done for high and low SCR systems. For operation with very low SCR systems the variable γ_0 mode of control must be adopted, at a normal value of γ_0 above the minimum, or fast SVCs should be used.

32.12.8 Critical short-circuit ratios

32.12.8.1 Definition and calculation

If the operating point coincides with MAP (SCR = 2 in *Figure 32.67*), then the corresponding SCR is termed critical (CSCR, CESCR, or CQESCR). The exact equation for CESCR can be found in references 25 and 27.

The angle ϕ, representing the system damping, has a small effect on CESCR in the region of 70–90°. Therefore, by assuming $\phi = 90°$, a simple formula is obtained for CESCR:[25]

$$\text{CESCR} = \frac{1}{U^2}\left[-Q_d + P_d \cot\tfrac{1}{2}(90° - \gamma_0 - u)\right] \qquad (32.13)$$

where U is the converter a.c. bus voltage per unit, P_d is the power supplied by the inverter per unit, u is the overlap angle of the inverter, γ_0 is the commutation margin (extinction angle) of the inverter, and Q_d is the reactive power consumed by the inverter. CSCR can be calculated from equation (32.13) by adding to it Q_c in per-unit of P_d.

For a given P_d and U, CESCR depends on γ_0 and u. As u is a function of γ_0 and the commutating reactance (x_c), CESCR is a function of x_c and γ. In *Figure 32.70* CESCR (calculated using the exact formula given in references 25 and 27), CSCR and CQESCR are plotted against x_c for $\gamma_0 = 15°$ and 20°, $\phi = 90°$ and 70°; and for $Q_c/Q_d = 1$ and 1.5.

CSCR varies by just over 50% for the assumed data. As discussed earlier, CESCR takes into account the value of Q_c and the variation is reduced to 27%. CQESCR takes account of Q_c and Q_d, and varies by 10%, from 0.9 to 1.0.

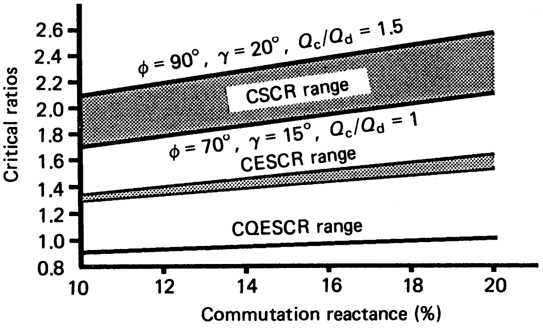


Figure 32.70 Sensitivity of critical short-circuit ratios

32.12.8.2 Significance of critical short-circuit ratios

Control-mode requirements CSCR represents a borderline between 'stable' and 'unstable' parts of the power curve when the rectifier is in power control and the inverter is in constant γ_0 control. Normal operation at the right side of MAP can be carried out only if a.c voltage is closely controlled. Moreover, if the operation is very close to MAP, even on the stable side, fully satisfactory operation can be achieved only if a.c. voltage is well controlled.

Expected levels of temporary overvoltages CSCR provides an indication of the expected temporary overvoltages. TOV_f for load rejection from MAP, i.e. at SCR will have a value near 1.4 p.u.; for higher values of SCR, TOV will be smaller and for lower values it will be higher, as can be seen from *Figure 32.67.*

Expected resonant frequency The resonant frequency which may occur between the converter station and the system is principally governed by the system impedance and the station shunt capacitors and can be expressed as

$$f_r = f_0\sqrt{S/Q_c}$$

where f_r is the resonant frequency and f_o is the system frequency.

From SCR = S/P_d for the case when $Q_c = 0.5\ P_d$

$$f_r = f_0\sqrt{2SCR} \qquad (32.14)$$

From equation (32.14), for SCR = 2 the resonant frequency would be near the second harmonic. From *Figure 32.67* it can be seen that SCR = 2 corresponds to CSCR for average inverter data.

Instantly available additional power The value of the short-circuit ratios gives an indication of additional power immediately available for high and low SCR systems, as indicated by the difference between the rated power and the MAP value (*Figure 32.67*). For the average inverter data used in *Figure 32.67*, the immediate power margin for SCR = 1.5 is equal to 1.28 p.u. at $I_d = 1.74$ p.u., and for SCR = 3 it is 1.08 p.u. at $I_d = 1.32$ p.u.

32.12.9 Short-circuit ratio as a guide to system planning[25]

For average converter data ($x_c = 15\%$, and minimum $\gamma_0 = 18°$) it can be concluded from *Figure 32.67* and the discussion in Section 32.12, that SCR will have the following approximate average values for different system strengths:

High SCR (strong),	a.c. system: SCR $\gg$3
low SCR (weak),	a.c. system: SCR $\gg$ 2
Very low SCR (very weak),	a.c. system: SCR $\ll$ 2

The above values of the SCR are only approximate. As described in the previous section, it is the value of the critical SCRs which determines whether a system behaves as having high, low or very low SCR values; as discussed and seen from *Figure 32.70* for different inverter data the CSCR values vary greatly.

The cost of a converter station will be higher for low and particularly for very low SCR systems. It is therefore important to obtain an approximate value of CSCR as early as possible in the planning stages of a scheme. In the early planning stages, the short-circuit MVA of the a.c. system, the proposed d.c. power and the amount of desirable reactive power compensation, may be the only known data, which would have a bearing on the CSCRs. An approximate value of CSCR can be obtained quickly from equation (32.12), by using the following equation, expressing all quantities in per unit of P_d,

$$\text{SCR} = \text{QESCR}[1 + Q_d] + Q_c \qquad (32.15)\Leftarrow$$

and, as CQESCR is approximately equal to 1, a good approximation for CSCR is

$$\text{CSCR} = \text{CQESCR} + Q_d + Q_c \qquad (32.16)\Leftarrow$$

In *Figure 32.71* CQESCR and Q_d (average value for γ= 18°⇐ to 20°) are plotted against the commutating reactance, which is normally equal to converter transformer reactance. Consider two examples.

(1) $x_c = 15\%$, γ= 18°⇐and $Q_c = Q_d$ (average data): using values of CQESCR and Q_d, from *Figure 32.71* equation (32.16) gives CESCR = 2.05, while the correct value is 2.0.
(2) $x_c = 20\%$, γ= 20°⇐and $Q_c = 1.5\ Q_d$: as for (1), equation (32.16) gives CESCR = 2.6, while the correct value is 2.58.

32.12.10 'Island' receiving system[25]

It is possible to supply most or all of the system power requirements to an island system by h.v.d.c. The example for such a scheme is the d.c. link supplying power from the Swedish mainland to the island of Gotland.[30] A more recent scheme bringing power to the South Korean island Cheju has a similar requirement.[43]

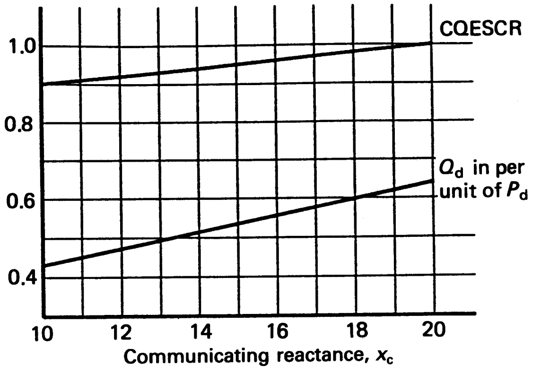


Figure 32.71 CQESCR and Q_d as a function of x_c

The a.c. line commutated converter, as used in h.v.d.c., depends for its operation on being supplied with an a.c. voltage of reasonable sinusoidal waveform. During a fault disturbance, say an a.c. system short circuit, the inertia of turbines and generators provides the energy to maintain the required system e.m.f. If all or most of the power is supplied to a network by h.v.d.c., the system inertia will be inadequate to provide transiently the required e.m.f. following a fault. In such cases a synchronous compensator is installed so that its inertia acts as a 'transient generator' to maintain the e.m.f. for sufficient time following a fault to enable the d.c. link to resume transmission.

Analogously to SCR, the required H constant of the synchronous compensator can be referred to P_d:

$$H_{dc} = H \cdot \text{MVA}_{sc}/P_d$$

where H is the inertia constant of the synchronous compensator and MVA_{sc} is its MVA rating.

H_{dc} can be calculated as

$$H_{dc} = P_{dc} \cdot dt \cdot f_o/2 \cdot df \qquad (32.17)\Leftarrow$$

If the frequency drop (df) is limited to 5%, for loss of power for $dt = 200$ ms (fault duration, breaker clearance, and time to get back to full power), then from equation (32.17) $H_{dc} = 2$ s.

A synchronous compensator dimensioned from the inertia point of view, will contribute sufficiently to the SCR value to convert a very low SCR system to a system having a high or, at least, low SCR value.

32.12.11 System interaction when the a.c. system impedance is high relative to d.c. power in-feed (low short-circuit ratio)[25–27,31]

The important and desired interaction between a.c. and d.c. systems is beneficial. D.c. can bring power into an a.c. system in a controlled way and can assist the a.c. system by improving its frequency control, stability and damping. The design of the converter system must allow for:

(1) steady-state stability;
(2) recovery after a.c. or d.c. faults; and
(3) a.c. and d.c. overvoltages.

32.12.11.1 Steady-state stability

Power and voltage instability This is a steady-state cumulative type of instability which can occur at low and very low SCR at the inverter end when power control modes are used. The simplest cure is to design the master controller to give a not-quite constant power-control when necessary. This and other methods[25,32,33] to prevent this form of instability have been discussed earlier in this section.

Core saturation instability This is an instability which involves saturation of converter transformers, and occurs when there is a resonance near to fundamental frequency on the d.c. side between d.c. reactors and d.c. line, with an antiresonance (or at least a high impedance) near to second harmonic on the a.c. side. The mechanism involves fundamental frequency on the d.c. side, d.c. components in valve windings which cause transformer saturation, and second harmonic on the a.c. side. The control system is also involved. Build up of this type of instability is very slow (several minutes) because of the transformer saturation effect. It can be cured by control system additions which effectively monitor the effect of saturation, and oppose it by feedback in the converter.[34]

Subsynchronous instability This can occur typically at frequencies of the order of 10–40 Hz, assuming the worst case of transmission via a long d.c. line or cable. It corresponds to ordinary control loop instability and can be studied by, for example, the Nyquist Criterion or by eigenvalues and cured by proper choice of control parameters such as loop gain settings at the rectifier and inverter. Synchronous machines with steam turbines exhibit mechanical shaft resonances in this frequency region, hence guaranteed stability is particularly important where these are present in the a.c. networks, to prevent their resonances being excited.

Suitably designed h.v.d.c. controls should not excite subsynchronous instability and in fact can be used to damp such instability.[35]

32.12.11.2 Recovery after a.c. and d.c. faults

Satisfactory recovery of the h.v.d.c. link after a major a.c. fault at the inverter can present problems, because the a.c. voltage distortion caused by magnetising in-rush current of transformers during recovery may be substantial. The most important criterion is the shock to machines caused by the loss of megawatt-seconds, which must be minimised to prevent pole-slipping in the a.c. system (transient instability). From the control point of view this requires a control system which re-starts as rapidly as possible on re-establishment of a.c. voltage, which has static and dynamic characteristics chosen to give fast restoration of power, and which gives freedom from commutation failures during recovery. As can be seen from oscillograms in the section on controls, modern controls can cope well even at low values of SCR. The situation can be helped further by suitable design of the converter transformers to give low magnetising in-rush current when economic, and of the a.c. filters to give substantial damping at low frequencies of the order of second to fourth harmonic.

Recovery following d.c. faults represents a less severe condition.

32.12.11.3 A.c. and d.c. overvoltages

Where a converter is connected to a low or very low SCR a.c. system, the effect of sudden blocking of the converter (load rejection) is to cause a substantial a.c. voltage rise. This may be caused for example by an a.c. fault at the remote station, or because of the necessity to cut off direct current to clear a d.c. line fault. Another case is where an a.c. fault near to the converter bus-bars is removed by a circuit breaker opening; when the a.c. voltage re-appears it is likely to be excessive, particularly if for some reason the converter does not commence commutation immediately. The control of these and other overvoltages is discussed in Section 32.9.

Overvoltages on the d.c. line are generally lower, at least with a d.c. cable scheme, because of the isolating effect of the d.c. reactors. However, with weak a.c. systems it is advisable to provide each station with voltage limiting loops in its controls; d.c. voltage surges can then be effectively reduced in most conditions, with benefit to cable insulation. A d.c. overhead line is subject to lightning overvoltages, which can be reduced by conventional methods.

32.12.11.4 Conditions at the rectifier end

The above are primarily concerned with the inverter (receiving system) end. The rectifier end source impedance has a much smaller effect on stability and recovery. However, the rectifier end is still liable to moderate a.c. voltage changes caused by disturbances. If the source is an isolated generating station without local loads, the a.c. voltage changes are generally acceptable. If there are local consumers then static compensators may again be a suitable solution to reduce a.c. voltage changes.

32.13 Multiterminal HVDC systems[36–38]

The applicability and flexibility of h.v.d.c. systems can be enhanced in some conditions if several converters are coupled to form a multiterminal h.v.d.c. system. The earliest application of this philosophy was the paralleling of bipoles I and II of the Nelson River Scheme onto one line in the event of an outage of one of the bipolar transmission lines. This procedure has now been carried out in response to a major storm damage to transmission lines. The first true multiterminal h.v.d.c. scheme has been achieved with the construction of a parallel tap (50 MW) on the Sardinia–Italian Mainland h.v.d.c. scheme (200 MW).[44]

Multiterminal h.v.d.c. systems may be divided into series and parallel types, illustrated in *Figures 32.72* and *32.73*, respectively. However, there are many permutations of each type, depending on system requirements.

32.13.1 Series connection

Figure 32.72 shows an example, in which one pole of a 500 kV, 1000 A rectifier suplies two inverter stations in series, respectively of 100 kV, 1000 A, and 400 kV, 1000 A.

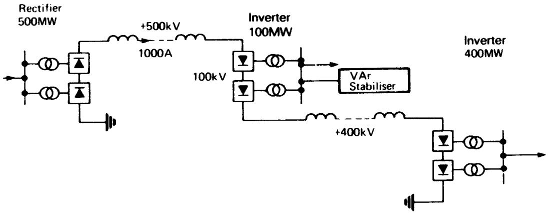


Figure 32.72 Three-terminal series scheme

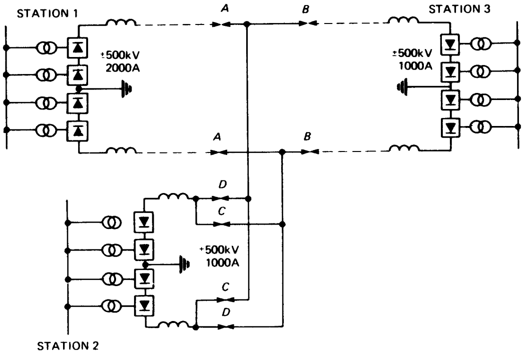


Figure 32.73 Three-terminal parallel scheme

In the case shown, one method of control is to operate the rectifier permanently at constant current equal to full rated current, and vary inverter powers individually as required by varying their voltages. This makes the series method expensive in terms of line losses, filter losses, damping losses, and reactive compensation, but does give virtually independent control (including smooth power reversal in any station), and very little interaction between stations during most disturbances. Some economy is obtained by reducing current when total power is less than full load, so that at least one inverter is at its rated voltage, but the sum of the ratings required of each item of plant (including filters, etc.) is always greater than for a parallel scheme.

32.13.2 Parallel connection

The example in *Figure 32.73* shows an arrangement for a three-terminal parallel scheme intended principally for operation as a 2000 MW rectifier supplying two inverters each of 1000 MW. This is an example of a system of medium control difficulty. Switches A and B are solely to isolate their respective lines in case of permanent d.c. line faults. Switches C and D form in addition a reversing system for station 2. The switches shown may be slow-acting isolators, plain fast switches, or true d.c. circuit-breakers.

Control of parallel rectifiers is easy since by using normal constant current control loops at each rectifier, their currents (and powers) are easily set to any desired values (including zero) and there are no current sharing problems.

The basic problem of control of parallel inverters is that an h.v.d.c. inverter operating in the most efficient mode, of constant extinction angle (γ), has an effective slope resistance which is negative, so that two such inverters in parallel are obviously unstable.

Many ways round this problem have been proposed, such as operating one inverter at constant extinction angle (to control direct voltage), and the other in a constant-current mode; this means that the second inverter requires higher plant rating. However, the system is in a rather delicate state for even minor transients.

Figure 32.74 shows an alternative control characteristic, originally developed to enable parallel operation of the two bipoles of the Nelson River Scheme, and which gives stable operation without requiring excessive plant rating. The current order at each station is communicated from a master control. Obviously the actual currents obey Kirchhoff's first law, so current orders should be restricted so that the sum of the rectifier current orders is equal to the sum of the inverter current orders.

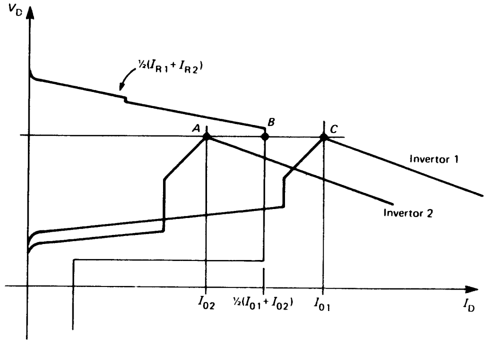


Figure 32.74 Normal steady-state operation

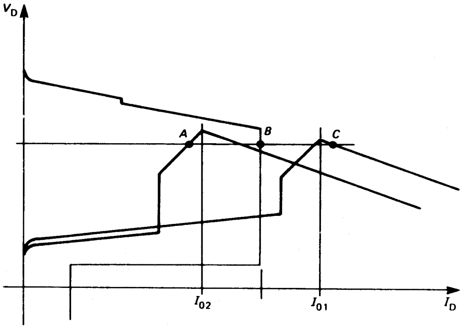


Figure 32.75 Inverter 1 tap-changer slightly low

Figure 32.74 has been drawn for the normal steady-state case where respective transformer tap-changers have been adjusted to the ideal positions, for the particular a.c. voltages, d.c. currents, etc. The working points of the inverters are right on the constant characteristic (in the ideal positions) at their correct currents, and the rectifiers have the usual small voltage margin in hand (i.e. their firing angles are in the range 2° to 15°).

The example of *Figure 32.75* shows the case in which inverter 1 a.c. voltage becomes relatively slightly low. The working points ABC must still obey Kirchhoff's and Ohm's laws, but are now not quite at their ideal positions, inverter currents being slightly high and low respectively, with inverter 2 operating at slightly high extinction angle. This is a stable operating condition. Powers are corrected in the short term by master control action, and normal (optimum) working conditions are re-established in a longer time by transformer tap-changers.

A commutation failure is a relatively frequent (and inevitable) minor disturbance, usually caused by a switching transient from the a.c. system at an inverter. Its effect is to cause temporary collapse of inverter operation for say 100–200 ms, momentarily giving zero d.c. voltage and a.c. current. In a multiterminal scheme there is a tendency for the whole of the d.c. current to be diverted temporarily into the 'failed' inverter. The rectifier controls will restrain this automatically after a cycle or two; the inverter will not be able to re-commutate (particularly with weak receiving system) until actual inverter current and firing angle fall below rated values.

32.13.3 D.c. circuit-breakers

One of the main advantages of an h.v.d.c. scheme is the controllability conferred by the ability of the converter valves to conduct or block the full load direct current as desired. This capability has meant that the need for true high voltage direct current circuit breakers has been small, since by appropriate converter action ordinary a.c. circuit breakers could be utilised for circuit switching on the d.c. side.

The first application for a d.c. breaker was the metallic return transfer breaker (MRTB). The MRTB is principally introduced to increase the independence between the poles of a bipolar scheme: A station fault requiring the blocking of a pole will divert the direct current from the healthy pole into the ground electrode. If the station fault is permanent, the faulty pole is isolated and bypassed, and transmission continues in the monopolar mode. However, although the metallic return conductor is then in parallel with the ground return, the major part of the current continues to flow into the electrode because of the lower resistance of this path. The MRTB can divert this current into the metallic conductor by developing a significant back e.m.f.

High voltage direct current breakers have now been developed and tested, but have not yet been commercially applied.[39] Such d.c. breakers may give some small reduction in the duration of outages caused by permanent faults in multiterminal d.c. networks. Since their duty is significantly harder than the transfer duty of the MRTB, they are still costly, and thus justifiable only when the rather small time improvements are of considerable benefit to the stability of the a.c system.

32.14 Future trends*

H.v.d.c. is now a maturing technology which has proved its technical and economic worth in the field. Future trends will be towards reduced complexity and lower cost, the latter achieved through both lower installed equipment cost and reduced running costs.

Very large transmission projects for which the economic transmission voltage lies above 600 kV will probably materialise, prompting the evolution of new insulation techniques for the d.c. side equipment. At lower voltages, developments will be generally confined to the specialist items of equipment such as valves and controls.

Two major developments which have contributed to present-day h.v.d.c. valves are open to further extension.

First, the increased rating per thyristor, which has led to thyristors based on 125 mm diameter silicon, may continue, albeit at a reduced rate as the technical and economic limits of processing large area, high voltage rated thyristors based on silicon is approached. It is unlikely that practical, silicon-based thyristors with blocking voltages in excess of 10 kv will be realised. Further extension from 125 mm diameter (the present limit) to a theoretically possible 150 mm seems unlikely because current ratings of all present applications can be covered by using 100 mm or 125 mm thyristors without paralleling. Development will be more towards better optimisation of dynamic characteristics and reduction of losses.

Semiconductor devices based on silicon carbide (SiC), rather than pure silicon, offer the exciting prospect of switching devices with much higher voltage blocking capacity than 10 kV. However, at present, the technology is only in its infancy and there are many technical challenges to overcome before practical SiC devices become available for use in H.v.d.c.

Second, the development of light triggering of individual thyristors, initially via auxiliary electronics mounted at valve potential, has already extended to the production of thyristors incorporating light sensitive gates suitable for direct optical triggering. In many applications, this reduces the need for 'local' electronics at each thyristor level, with the prospect of substantial economic advantage. For h.v.d.c., in the first instance, small direct optically triggered 'slave' thyristors have been employed to gate conventional thyristors, and to enable local protective circuits to act directly via the electrical gate. The value of these thyristors to h.v.d.c. is limited by the continuing need for local protective circuits to prevent damage to the thyristors during forward overvoltage and adverse forward recovery conditions. Work to develop direct triggered h.v.d.c. thyristors with fully integrated self-protection has met with some success with regards to overvoltage and $\mathrm{d}v/\mathrm{d}t$ protection (see *Figure 32.76*). However, integration of an effective forward recovery protection has proved illusive. Light triggered thyristors are practical and remain attractive but are generally uneconomic at present state of development.

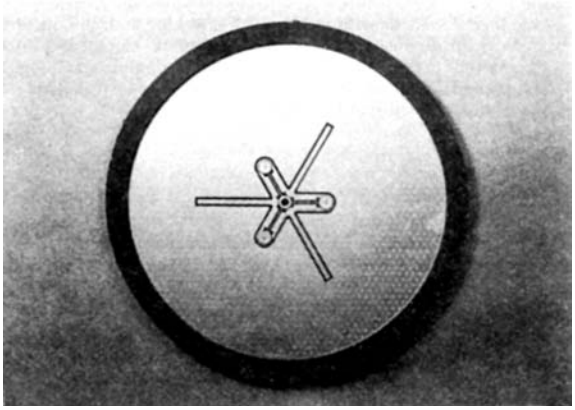


Figure 32.76 Silicon slice of a direct light-fired, self-protected thyristor. (Courtesy of Marconi Electronic Devices Ltd)

The universal adoption in recent schemes of indoor air-insulated valves, rather than outdoor oil or SF_6 insulated, stems largely from the need for regular, perhaps annual, access to thyristor levels which must be replaced or renovated to restore acceptable redundancy. Reducing the component count in a valve through the use of higher voltage self-protecting light-fired thyristors may make the outdoor oil or SF_6 insulated valve attractive once more, at least in regions where extreme climatic conditions are not encountered. However, this approach is unlikely to be adopted until satisfactory operating experience of light-fired self-protecting thyristors has confirmed the high availability and low replacement rates which are expected.

Today's h.v.d.c. inverter is line commutated, that is it relies on the e.m.f. of the receiving a.c. system to commutate current from one phase to the next against the restraining action of the commutation reactance. Forced commutation inverters, in which the energy for commutation is stored locally, usually in capacitors, and is released by auxiliary thyristors to force commutation at any desired time, is too expensive in first cost, in losses, and in reduced reliability through complexity, to be a serious competitor in large systems. Such circuits could, in principle, be applied to small schemes feeding isolated a.c. systems, where the advantage of being able to operate without the support of rotating machinery is greatest.

The advent of Gate Turn-off (GTO) thyristors and, in recent years, of Insulated Gate Bipolar Transistor (ITBT) technology at power levels of use in high power converters has opened up opportunities for small power (typically below 100 MW) h.v.d.c. converters based on voltage sourced converter technology.[45] At present, schemes employing the self-commutating capability of such converters are restricted to niche applications where special circumstances offset the higher capital cost/MW and reduced power conversion efficiency of voltage sourced converter technology compared with conventional, line-commutated technology. Improvements in the basic power semiconductor switching devices and the evolution of novel converter circuit topologies could well see voltage sourced converters become the technology of choice for h.v.d.c. transmission, initially at low power and later, as technology matures, at increasing levels of power.

---

*as considered in June 2000

The performance of an h.v.d.c. scheme is critically dependent on the central control system, particularly if the a.c. system at the receiving end is weak. The cost of the controls is a small fraction of the overall cost of a scheme, hence improvement of control performance and operational reliability will continue even where this incurs somewhat higher costs.

Some parts of the controls will be implemented on programmable digital systems with internal self-checking and, in critical areas, with duplicated circuits and automatic changeover in the event of faults. Because of their relative simplicity and low component count, analogue circuits are likely to be retained in those areas to which they are best suited.

References

1 KIMBARK, E. W., *Direct Current Transmission*, Vol. 1. Wiley Interscience, New York (1971)

2 THORP and MACGREGOR, *Design of the Sea Electrode System, Sardinia–Italian Mainland 200 kV Scheme* (*IEE Conference Publication No. 22*), London (1966)

3 BURGESS, R. P. and KOTHARI, R., *Design Features of the Back-to-Back HVDC Convertor Connecting the Western and Eastern Canadian Systems*, IEEE July (1989)

4 GAVRILOVIC, A., *HVDC Scheme Aspects Influencing Design of Converter Terminals*, International Symposium HVDC, Rio de Janeiro (March 1983)

5 IEC, *Insulation Co-ordination Application Guide for Insulation Co-ordination and Arrester Protection of HVDC Converter Stations* (*IEC Publication Nos 71-1 and 71-2*), London (1976) (CIGRE WG 33-05, Electra, **96** (1984))

6 GIBSON, H., BALLAD, J. P. and CHESTER, J. K., *Characterisation, Evaluation and Modelling of Thyristors for HVDC* (*IEE Conference Publication No. 255*), IEE London (1985)

7 WOODHOUSE, M. L., BALLAD, J. P., HADDOCK, J. L. and ROWE, B. A., 'The control and protection of thyristors in the English terminal cross channel valves, particularly during forward recovery', *IEE Conference Publication No. 205, Thyristor and Variable Static Equipment for A.C. and D.C. Transmission*, IEE, London (1981)

8 CHESTER, J. K., 'A new technique for deriving self-consistent electrical and thermal models of thyristors during surge loops', *IEE Conference Proceedings, Power Electronics, Power Semi-Conductors and their Applications*, IEE, London (1977)

9 EKSTROM, A. and JUHLIN, L. E., *Testing of Thyristor Valves*, (*CIGRE Publication No. 14-03*), CIGRE, Paris (1972)

10 BANKS, R., ROWE, B. A. and NOBLE, R. G., *Testing Thyristor Valves for HVDC Transmission* (*CIGRE Publication No. 14-07*), CIGRE, Paris (1978)

11 DEMAREST, O. M. and STOILS, C. M., *Solid State Valve Test Procedures and Field Correlation* (*CIGRE Publication No. 14-12*), CIGRE, Paris (1978)

12 LIPS, P., THIELE, G., HUYNH, H. and VOHL, P. E., 'Design and testing of thyristor valves for the HVDC back-to-back TIE Chateauguay', *International Conference on DC Power Transmission*, Montreal, Canada (June 1984)

13 INTERNATIONAL ELECTROTECHNICAL COMMISSION, *Publication 700*, IEC, Geneva (1981)

14 KRISHNAYYA, P. C. S., *Important Characteristics of Thyristors, of Valves for HVDC Transmission and Static Var Compensators*, (*CIGRE Publication No. 14–10*), CIGRE, Paris (1984)

15 CIGRE, *Guide for Planning DC Links Terminating at AC Locations having Low Short Circuit Capacities. Part I: AC/DC System Interaction Phenomena*, CIGRE Technical Brochure 68, Paris (1992)

16 *British Patent 1 300 226*. (Current-error characteristic)

17 AINSWORTH, J. D., 'The phase-locked oscillator—a new control system for controlled state converters', *IEEE Trans.*, **PAS-87**(3), 859–865 (March 1968)

18 AINSWORTH, J. D., 'Harmonic instability between controlled static converters and a.c. networks', *Proc. IEE*, **114**(7), 949–958 (July 1967)

19 AINSWORTH, J. D., 'Developments in the phase-locked oscillator control system for HVDC and other large converters', IEE Conference on a.c. and d.c. Power Transmission. *IEE Conference Publication No. 255* (September 1985)

20 *British Patents 1 170 249 and 1 258 974* (Master control and telecommunication)

21 AINSWORTH, J. D., 'Telecommunication for HVDC', IEE Conference on Thyristor and Variable Static Equipment for a.c. and d.c. Transmission, *IEE Conference Publication No. 205*, (November 1981)

22 MARTIN, C. J. B. and UHLMANN, E., *AC Network Stabilisation by DC Links* (*CIGRE Paper 32-01*), CIGRE, Paris (1970)

23 HAYWOOD, R. W. and RALLS, K. J., *Use of HVDC for Improving AC System Stability and Speed Control*, Manitoba Power Conference EHV-DC, Winnipeg, Manitoba (June 1971)

24 AINSWORTH, J. D. and MARTIN, G. J. B., 'The influence of h.v.d.c. links on a.c. power systems', *GEC Journal of Science and Technology*, **44**(1), (1977)

25 CIGRE, *Guide for Planning DC Links Terminating at AC Locations having Low Short Circuit Capacities. Part I: AC/DC System Interaction Phenomena*, CIGRE Technical Brochure 68, Paris (1992)

26 GAVRILOVIC, A., KRISHNAYYA, P. C. S., PEIXOTO, C. O. A., AINSWORTH, J. D., BOWLES, J. P., HAMMAD, A., LISS, G. and THIO, C. V., *Aspects of a.c./d.c. System Interactions: Peak Available Power, Second Harmonic Resonance, Low Inertia Systems, Controllability of h.v.d.c.*, MONTECH-IEEE Conference (September–October 1986)

27 GAVRILOVIC, A., KRISHNAYYA, P. C. S., AINSWORTH, J. D., BOWLES, J. P., BREUER, G. D., HAMMAD, A., LISS, G., PEIXOTO, C. O. A., POVH, D. and THIO, C. D., '*Interaction between a.c and d.c. systems', CIGRE Symposium: AC/DC Transmission Interactions and Comparisons*', Boston (September 1987)

28 GAVRILOVIC, A., *HVDC Scheme Aspects Influencing the Design of Converter Terminals*, International Symposium on HVDC Technology, Rio de Janeiro (March 1983)

29 THANAWALA, H. L., *Maximum Available Power Features of a.c. and d.c. Transmission Systems*, Power Technology International (1990)

30 LISS, G. and SMEDSFELT, S., *HVDC Links for Connection to Isolated a.c. Networks*, United Nations Seminar on High Voltage Direct Current (HVDC) Techniques (May 1985)

31 AINSWORTH, J. D. and GAVRILOVIC, A., *Interaction Between HVDC and a.c. Systems when the d.c. Link is Large Compared to the a.c. System*, United

Nations Seminar on High Voltage Direct Current (HVDC) Techniques, Stockholm (May 1985)

32 AINSWORTH, J. D., GAVRILOVIC, A. and THANAWALA, H. L., *Static and Synchronous Compensators for h.v.d.c. Transmission Converters Connected to Weak a.c. Systems* (*CIGRE Paper No. 31-01*), CIGRE, Paris (August 1980)

33 HAMMAD, A. and KAENFERLE, J., *A New Approach for the Analysis and Solutions of a.c. Voltage Stability—Problems at h.v.d.c. Terminals*, HVDC Symposium, Montreal, Canada (June 1984)

34 AINSWORTH, J. D., *Development in the Phase-locked Oscillator Control Systems for HVDC and Other Large Converters* (*IEE Conference Publication No. 255*), London (1985)

35 PIKE, P. J. and LARSEN, E. V., *HVDC System Control for Damping of Subsynchronous Oscillations* (*IEEE PZS-101*) (July 1982)

36 CARCANO, C., INESI, A., MAZZOLDI, F. and RICCI, F., *Rebuilding of the h.v.d.c. Sardinia–Corsica–Italy Mainland Link (SACOI): Installation of Two New Conversion Stations and a Tapping Station in Corsica for Multi-terminal Operation.* (*IEE Conference Publication No. 255*), IEE, London (1985)

37 AINSWORTH, J. D., *Multi-terminal h.v.d.c. Systems* (*CIGRE Meeting SC.14*), Winnipeg, Canada (June 1977)

38 LONG, W. F., REEVE, J. M'NICHOL, J. R., HARRISON, R. E. and BOWLES, J. P., *Considerations for Implementing Multi-terminal d.c. Systems*, W. F. Long, IEEE PES Winter Meeting (1985)

39 VITHAYATHIL, J. J., *HVDC Breaker and its Application*, International Symposium in HVDC Technology, Rio de Janeiro, Brazil (March 1983)

40 MELTA, H. and TEMPLE, V. A., *Advanced Light-triggered Thyristor, A.c. and d.c. Power Transmission Conference* (*IEE Conference Publication No. 255*), London (1985)

41 TAYLOR. P. D. and FRITH, P. J., *Recent Advances in High Voltage Thyristor Design, A.c. and d.c. Power Transmission Conference* (*IEE Conference Publication No. 255*), London (1985)

42 NISHIZAURA, J., *New Thyristor Applicable to d.c. Power Transmission, A.c. and d.c. Power Transmission Conference* (*IEE Conference Publication No. 255*), London (1985)

43 THANAWALA, H. L., WHITEHOUSE, R. S., GOO OURCK KWON and SUK JIN LEE, *Equipment and control features of Haenam–Cheju Link in South Korea,* (*CIGRE Paper 14-303*), CIGRE, Paris (1994)

44 MAZZOLDI, F., TAISNE, J.-P., MARTIN, C. J. B. and ROWE, B. A., *Adaptation of the Control Equipment to permit 3-terminal operation of the HVDC link between Sardinia, Corsica and Mainland Italy*, IEE Summer Power Meeting (1988)

45 APSLUND, G., ERIKSON, K., JIANG, H., LINDBERG, J., PALSSON, R. and SVENSSON, K., *DC Transmission based on Voltage Source Convertors*, (*CIGRE Paper 14-302*), CIGRE, Paris (1998)

Bibliography

An Annotated Bibliography of High Voltage Direct Current Transmission, 1969–1983, Bonneville Power Administration

An Annotated Bibliography of High Voltage Direct Current Transmission, 1984–1989, Western Area Power Administration, Bonneville Power Administration

An Annotated Bibliography of High Voltage Direct Current Transmission, 1989–1991, Western Area Power Administration, Bonneville Power Administration

An Annotated Bibliography of High Voltage Direct Current Transmission and Flexible AC Transmission (FACTS) Devices, 1991–1993, Bonneville Power Administration, Western Area Power Administration

An Annotated Bibliography of HVDC Transmission and FACTS Devices, 1994–1995, US Department of Energy, Western Area Power Administration.

An Annotated Bibliography of HVDC Transmission and FACTS Devices, 1996–1997, EPRI and Bonneville Power Administration

IEE Publications

IEEE 857 The Testing of Thyristor Valves for HVDC: (1996)

International Conference on DC Power Transmission, Montreal, Canada (1984)

Overvoltages and compensation on integrated a.c./d.c. systems, (*IEEE Conference Proceedings*), Winnipeg, Canada (June 1980)

Conferences other than CIGRE, IEE, IEEE

International Symposium on HVDC Technology–Sharing the Brazilian Experience, Rio de Janeiro (1983)

Manitoba Power Conference EHV-DC, Winnipeg, Canada (1971)

US DEPARTMENT OF ENERGY, *Incorporating HVDC Power Transmission into Power System Planning*, Phoenix, Arizona (1980)

Books by individual authors

ADAMSON, C., and HINGORANI, N. G., *High Voltage Direct Current Power Transmission*, Garraway, London (1960)

ARRILLAGA, J., *High Voltage Direct Current Transmission* (*IEE Power Engineering Series 6*) Peter Peregrinus, London (1983)

CORY, B. J. (Ed.), *High Voltage Direct Current Convertors of Systems*, Macdonalds, London (1965)

KIMBARK, E. W., *Direct Current Transmission,* Vol. 1, Wiley Interscience, New York (1971)

UHLMANN, E., *Power Transmission by Direct Current*, Springer-Verlag, Berlin (1975)

33

Power Transformers

D J Allan FREng, CEng, FIEE, FIMechE, FIEEE

Contents

33.1 Introduction 33/3

33.2 Magnetic circuit 33/3
  33.2.1 Core steel 33/3
  33.2.2 Magnetic circuit design 33/4
  33.2.3 Magnetic circuit characteristics 33/5

33.3 Windings and insulation 33/7
  33.3.1 Types of coil 33/7
  33.3.2 Insulation 33/8
  33.3.3 Winding design 33/9
  33.3.4 Leakage field 33/11
  33.3.5 Impedance voltage 33/12
  33.3.6 Losses 33/12
  33.3.7 Cooling 33/13
  33.3.8 Short-circuit conditions 33/13

33.4 Connections 33/13
  33.4.1 Phase conversion 33/14

33.5 Three-winding transformers 33/15
  33.5.1 Impedance characteristics 33/15
  33.5.2 Tertiary windings for harmonic suppression 33/15
  33.5.3 Tertiary windings for external loads 33/16

33.6 Quadrature booster transformers 33/16

33.7 On-load tap changing 33/16
  33.7.1 Tap changer control 33/17
  33.7.2 Line-drop compensation 33/18

33.8 Cooling 33/18
  33.8.1 Air insulated, air cooled 33/19
  33.8.2 Oil immersed, air cooled 33/19
  33.8.3 Oil immersed, water cooled 33/19
  33.8.4 Overload capability 33/20

33.9 Fittings 33/20

33.10 Parallel operation 33/21

33.11 Auto-transformers 33/21
  33.11.1 Auto-starters 33/22

33.12 Special types 33/22
  33.12.1 Static balancer 33/22
  33.12.2 Welding transformers 33/22
  33.12.3 Mining transformers 33/23
  33.12.4 Small transformers 33/23

33.13 Testing 33/23
  33.13.1 Routine tests 33/23
  33.13.2 Type tests 33/26
  33.13.3 Special tests 33/27
  33.13.4 Commissioning tests at site 33/27

33.14 Maintenance 33/27
  33.14.1 Insulating oil 33/28
  33.14.2 Insulation 33/28
  33.14.3 On-load tap changing equipment 33/28
  33.14.4 Reliability and condition monitoring in service 33/28

33.15 Surge protection 33/29

33.16 Purchasing specifications 33/30

33.1 Introduction

A transformer consists essentially of two or more electric circuits in the form of windings magnetically interlinked by a common magnetic circuit. An alternating voltage applied to one of the windings produces, by electromagnetic induction, a corresponding e.m.f. in the other windings, and energy can be transferred from the primary circuit to the other circuits by means of the common magnetic flux.

References to various specialist papers relating to design and operation of power transformers are given at the appropriate points in the text and listed at the end of this chapter. A recommended comprehensive textbook on all aspects of transformer design is *Large Power Transformers*.[1] The relevant British Standard Specification[2] is *BSEN 60076-1:1997. Power Transformers*. BSEN 60076 contains information relating to standard characteristics, guaranteed performance and tolerances, testing and operation, and conforms with the International Electrotechnical Commission (IEC) *Publication 60076: Power transformers: Parts 1 to 12*.

The British electricity supply industry issued a series of documents relating to transformers. All are listed in The Electricity Council's *Catalogue of Engineering Documents* and the more important are:

ESI Standard 35-1: Distribution transformers (from 16 kV A to 1000 kVA) (1985)
ESI Standard 35-2: Emergency rated system transformers (35/11.5 kV) (1982)
British Electricity Board Specification (BEBS) 72: Transformers and reactors (1966). In 34 sections, this specification relates to large transformers above 20 MVA
BEBS T3 Transformers for 33 kV and 22 kV systems (up to 20 MVA) (1962)

These documents are no longer maintained under revision by the Electricity Council as most Electricity Companies are now developing Functional Specifications, but they are available on demand.

33.2 Magnetic circuit

The magnetic circuit, or core, provides a closed ferromagnetic path for the flux. To prevent excessive eddy current loss within the metal of the core itself it must be laminated in a plane parallel to the flux path and the individual laminations must be insulated from each other.

33.2.1 Core steel

For many years power transformer core laminations were cut from sheets of special 4% silicon steel produced by a hot-rolling process. During the 1940s, an improved material was developed, known as cold-rolled grain-oriented strip (c.r.o.s.). This material has a silicon content of approximately 3% and is produced in strip form in rolls of up to 5 t.

Because of the effect of the cold-rolling process on the grain formation, the magnetic properties in the rolling direction are far superior to those in other directions.[3,4]

A heat-resistant insulation coating is applied by thermochemical treatment to both sides of the steel during the final stage of processing. The coating is approximately 1 μm thick and has only a marginal effect on the stacking factor.

Traditionally, a thin coat of varnish had been applied by the transformer manufacturer after completion of cutting and punching operations, but improvements in the quality and adherence of the steel manufacturer's coating and in the cutting tools available have eliminated the need for the second coating and its use has been discontinued.

Guaranteed values of loss (in watts per kilogram) and apparent power (in volt-amperes per kilogram) apply to magnetisation at 0° to the direction of rolling. Both real and apparent power loss increase significantly (by a factor of 3 or more) when c.r.o.s. is magnetised at an angle to the direction of rolling. Guarantees do not apply and the transformer manufacturer must ensure that a minimum amount of core material is subject to cross-magnetisation. This is to minimise the total core loss and (equally importantly) to ensure that the core temperature in the area is kept within safe limits.

Cold-rolled grain-oriented strip cores operate at nominal densities of 1.6–1.8 T. This compares with 1.35 T used for hot-rolled steel, and is the principal reason for the remarkable improvement achieved in the 1950s in transformer output per unit of active material. British c.r.o.s. steel is produced in two magnetic qualities (each having two subgrades) and four thicknesses (0.23, 0.27, 0.30 and 0.35 mm), giving a choice of seven different specific loss values. In addition, the designer can consider using Japanese-made steel of higher quality, available in three thicknesses (0.23, 0.27 and 0.3 mm).

The decision on which grade to use for a particular application depends on the characteristics required in respect of impedance and losses, and particularly, on the assigned capitalised value of the iron loss. The higher labour cost involved in using the thinner materials is another factor to be considered. The different materials are identified by code names. For example, the material previously known as 27 MOH is now called 103-27-PS, where the digits signify:

103: a guaranteed 50 Hz specific loss at B = 1.7 T of 1.03 W/kg;
27: a thickness of 0.27 mm; and
P5: the steel manufacturer's code for the higher quality steel previously identified by suffix 'H'.

The lower grade material is known as N5 and the first figure in the code is then an indication of the loss per kilogram at B = 1.5 T, e.g. the complete code might be 089-27-N5 (loss = 0.89 W/kg).

The Japanese grade ZDKH steel is subjected to laser irradiation to refine the magnetic domains near to the surface. This process considerably reduces the anomalous eddy current loss but the laminations must not be annealed after cutting.

33.2.1.1 Cutting and punching

Cold-rolled grain-oriented strip is produced in the form of strip up to about 850 mm wide. In the past it was common practice for the transformer manufacturer to buy full-width coils and slit these to the width required. It is now more usual to purchase the strip ready cut to width. This is more expensive unless the manufacturer has a very high turnover of core laminations, but the extra cost is offset by the elimination of the slitting operation and the wastage incurred by ever-increasing stocks of unused off-cuts. The only cutting process now undertaken by the transformer manufacturer is to crop to length by guillotine.

Where bolt-holes are employed, it is preferable to use a single hydraulic press to crop the strip material to length and punch bolt-holes simultaneously. Following cutting and punching, the individual laminations may need to be dressed to remove any edge burrs, but as deburring may harm the magnetic properties of the material, it is preferable that high quality cutting tools are maintained in good condition so that deburring is unnecessary.

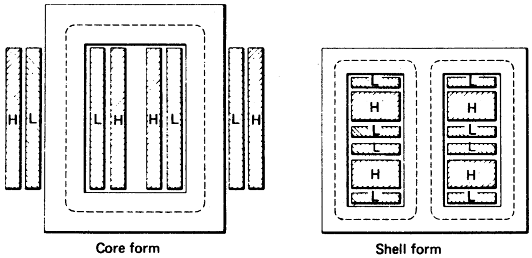
Cutting and punching adversely affect the magnetic properties of the material and, until recently, it has been considered desirable that the finished laminations (or for small units the complete cores) should be stress-relief annealed to remove cutting strains. Various types of annealing furnace have been used, including the batch furnace and the continuous belt furnace, in which small stacks of laminations (up to 10 or 12 plates deep) pass slowly through a heating zone at about 800 °C in an inert atmosphere of nitrogen. In the single-sheet roller hearth furnace, single laminations pass relatively rapidly through the heating zone; with this furnace the laminations are in the heated zone for a comparatively short time and it is unnecessary to provide an inert atmosphere to prevent oxidation.

The elimination of bolt-holes and improvements in cutting tools have led to a reduction in cutting strains and in the loss thereby incurred. The margin for reducing loss by annealing has been reduced and the process has generally been discontinued. (There is, for instance, obviously proportionately less area affected by cutting strains in a wide lamination than in a narrow one. As a rough guide, annealing is not considered to be necessary for strip over 200 mm wide.)

Although c.r.o.s. is now used for virtually all 'power' transformer cores from 1 kVA upwards, there are other special low-loss steels (e.g. Mumetal) used for the cores of instrument transformers. These materials have markedly superior magnetic properties (i.e. magnetising vars and power loss) at low densities, but they saturate at much lower levels than c.r.o.s. They are therefore not economic for power transformers, where the advantage of a high operating flux density overrides all other factors.

33.2.2 Magnetic circuit design

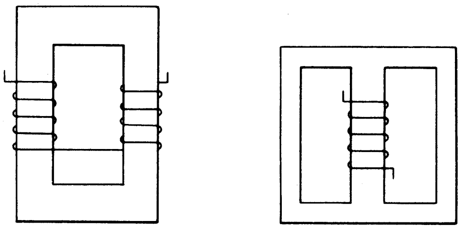
The two fundamental types of construction used, shown diagrammatically in *Figure 33.1*, are known as *core* and *shell*, respectively. The normal arrangement of a core-type transformer is for circular primary and secondary windings to be arranged concentrically around the core leg of substantially circular cross-section. In the shell type the magnetic circuit is of rectangular cross-section formed by a stack of laminations of constant width. The coils are straight-sided and the primary and secondary windings are interleaved in a sandwich fashion. The number of alternate high-voltage/low-voltage groups is dependent upon the required reactance characteristics of the transformer.



Core form
circular limb

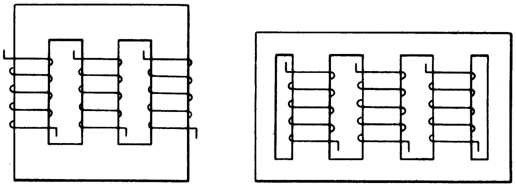
Shell form
rectangular limb

Figure 33.1 Basic cell and magnetic circuit arrangements



(a) Single-phase, two limbs wound

(b) Single-phase, centre limb wound



(c) Three-phase, three limb

(d) Three-phase, five limb

Figure 33.2 Core arrangements

In the UK the core-type transformer is used for all power-system applications. The shell form is used only for special applications, usually where very heavy current low-voltage outputs are required for such purposes as electric arc furnaces or short-circuit testing stations. The shell form is also used in the USA, France, Spain, Portugal and Japan for power transformers of the largest sizes and highest voltages. The fact that both core- and shell-type transformers have existed side by side for many years indicates that neither has any significant economic advantage. There is, however, a discernible trend away from the shell towards the core type of construction, due entirely to economic considerations.

Core-type transformers (circular coils on circular limbs) are built in various forms, as shown in *Figure 33.2*. The choice is dictated by the type of winding or the need to meet transport loading restrictions. Thus, a five-limb core will have a smaller overall height than a three-limb core for a three-phase transformer. A very large high-voltage single-phase transformer might also have to be designed with a multi-limb core.

The basic parameters of the core are: (a) the core circle diameter, (b) the yoke height, (c) the limb centres, and (d) the limb length. The normal practice in a design office is that the first three are related to a range of fixed standards, while a limited variation on core size is secured by change of limb length. The relation between the core circle diameter D and the rating S of a transformer is given approximately by

$$D = k(S)^{1/4}$$

where k is a constant depending on the type of transformer.

Thus, if a 500 kVA transformer has core circle diameter of 190 mm, then a 1000 kVA transformer of the same voltage ratio and service conditions would have $D = 220$ mm approximately. The actual diameter adopted would depend on the nearest standard diameter available.

The e.m.f. per turn E_t is a function of the frequency f, the peak flux density B_m and the net core area A_i:

$$E_t = 4.44 f B_m A_i$$

33.2.2.1 Limb space factor

To obtain the most economical use of core steel and copper it is essential to utilise as much as possible of the available

Table 33.1 Core utilisation factors

No of steps	1 (square)	2	3	4	5	6	7
Fraction of core circle areas	0.64	0.78	0.85	0.89	0.91	0.92	0.94

core circle. Consequently, the limb sections are built in steps; the greater the diameter, the larger the number of steps. Optimum utilisation factors for cores with 1–7 steps are given in *Table 33.1*.

In practice these areas are reduced by the lamination stacking factor, depending on the thickness of the interlaminar insulation and the tightness of the core clamping. With the latest core steel insulation and without varnish, stacking factors of up to 97% can be achieved. With large transformers allowance must be made for loss of area due to cooling ducts in the core, and the area will also be affected by the relative size of the individual core steps, which can depend on the width necessary to accommodate core clamping plates.

Cores having more than seven steps (up to as many as 15 or more) are common, but the improvement in utilisation decreases with an increasing number of steps and is offset from the production point of view by the increasing complexity involved in producing a greater number of widths.

33.2.2.2 Yoke dimensions

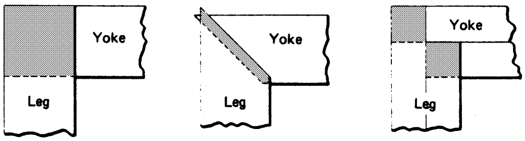
As the yoke height is not restricted by a winding, the area selected may be larger than that of the corresponding limbs. Dissimilar steps between the limb and yoke packets, however, can give rise to cross-flux, and the resultant increase in core loss may offset benefits accruing from the greater yoke area; this factor is of particular significance with c.r.o.s. Apart from very small cores, modern practice is to step the yoke and limb sections in a similar, or preferably identical, manner.

When the magnetic circuit is completed by a single-path yoke, e.g. a one-phase two-limb core (*Figure 33.2*(*a*)), or a three-phase three-limb core (*Figure 33.2*(*c*)), yoke areas at least equal to the limb section are necessary. Where the yoke path is split, the yoke area can be appropriately reduced, e.g. with a one-phase centre-limb-wound core (*Figure 33.2*(*b*)), the yoke area can be halved, and with a three-phase five-limb core (*Figure 33.2*(*d*)), the yoke area can be reduced to about 0.55–0.6 of that of the corresponding limb.

33.2.2.3 Core clamping and core joints

Bolt-holes cause local flux deviation and crowding, which results in increased noise and loss. These effects have led to the development of boltless cores, which are held together by circumferential bands. These bands are sometimes of steel with suitable insulated sections, but synthetic-resin-impregnated glass fibre tape is preferable, as it eliminates any risk of the band becoming involved in an electrical failure or becoming overheated owing to eddy currents. Banded cores are now used even in the largest transformers. Other methods of clamping include bolts passing through oil ducts in the core to avoid holes in the sheet, and reduction of the bolt area by the use of high-tensile steel.

The introduction of c.r.o.s. necessitated a modification of the traditional rectangular overlap (*Figure 33.3*(*a*)) between yoke and leg laminations commonly used with hot-rolled strip cores. To achieve minimum loss by taking maximum advantage of the directional properties of c.r.o.s., the area overlap has been minimised by means of mitred joints (*Figure 33.3*(*b*)). The affected area can also be reduced by using split laminations (*Figure 33.3*(*c*)), with the split joint butted in small cores, or separated to form a cooling duct in the case of large cores.



(a) rectangular overlap **(b)** Mitred overlap **(c)** Split lamination overlap

Figure 33.3 Types of joint in cores (shaded portions indicate overlap areas)

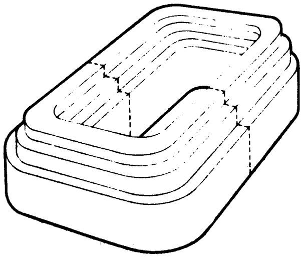


Figure 33.4 Strip-wound core (cuts in legs for butt-jointed core are shown by the dotted line)

By eliminating core joints, or by arranging that joints do not disturb the optimum flux path, maximum advantage can be taken of c.r.o.s. Such conditions are achieved with the wound core (*Figure 33.4*). This construction is widely used for transformers up to 25 kVA, the limitation in rating being imposed by special coil-winding requirements which involve a split former so that the coils can be wound on the completed core. A variation called a C core, of slightly reduced efficiency, is a wound core cut across the centre of the leg section, thus involving a butt joint in each limb.

33.2.2.4 Core building factor

The total core loss divided by the total mass of the completed magnetic circuit gives the specific loss of the built core. If this figure is divided by the specific loss of the material used (from tests on samples or from the steelmaker's guarantee), the result is the core building factor, a measure of the effectiveness of the magnetic circuit design. Building factors vary with the size of core, smaller cores generally giving factors near to unity. Building factors for large cores can be kept to less than 1.2 if full attention is paid in the design to make the best use of the directional properties of the steel.

33.2.3 Magnetic circuit characteristics

The magnetic circuit characteristics are the no-load core loss and the magnetising current. The former is commonly divided into hysteresis and eddy current components. The hysteresis loss depends on the peak flux density and the frequency, while the eddy current loss depends on the r.m.s.

flux density and on the degree of subdivision (i.e. lamination) of the core. The specific loss components are given by

$p_h = k_h f B_m^n$ and $p_e = k_e f^2 B^2 t^2$

where B_m is the peak and B the r.m.s. flux density, f is the frequency, t is the lamination thickness, and k_h and k_e are constants for a given material. The exponent n is empirical, with a value generally not very different from 2.

The measured no-load loss includes I^2R loss due to the no-load current; and dielectric loss (especially for large high-voltage transformers); these components can often be neglected.

The no-load current I_0 can be taken as comprising an active component I_{0a} in phase with the applied voltage V and accounting for the no-load power input P_0, together with a reactive or magnetising component I_{0r} (*Figure 33.5*). The two components are not physically separable, as both are concerned together in the magnetisation of the core. The no-load current is normally expressed in terms of the full-load current as a percentage or per-unit (p.u.) value. For distribution transformers in the range 500–1000 kVA and built of c.r.o.s., the no-load current is of the order of 1.5% (0.015 p.u.). Values for large high-voltage transformers may be less than 0.5%.

The saturation characteristic of the steel is such that the no-load current increases rapidly when the transformer is overexcited. If with c.r.o.s. the normal peak density is 1.6 T, the no-load current will be at least doubled by a 10% overvoltage. Further, the harmonics in the current waveform will be substantially increased. Provided that the circuit connections permit, the most pronounced harmonic is the third, although fifth and seventh harmonics can become significant at high densities. If the circuit connections do not permit third-harmonic currents to flow (in a three-phase transformer there must be either a neutral connection to a source of zero sequence current, or a delta connection within the transformer in which third harmonics can circulate), the flux waveform will distort because of the lack of the component. The distortion appears as a third-harmonic ripple in the secondary line-to-earth voltage.

This effect is most noticeable in single-phase transformers or in a five-limb, three-phase core. In both cases there is a free path in the core for a third harmonic component in the main flux. In a three-phase, three-limb core a third harmonic in the flux is in phase in each limb and the lack of a return ferromagnetic path results in the suppression of the third harmonic components. The effect on the magnetising current is that, although there is not any true third harmonic therein, there is a triple frequency component divided 1:2:1 in the core. To a large extent this satisfies the need for a third harmonic and the distortion of the flux wave (and hence of the output voltage) is negligible.

The magnetic path length associated with the central phase of a three-phase three-limbed core-type transformer (*Figure 33.2(c)*) is significantly shorter than that of either of the outer phases. The configuration shows that the outer path lengths include two half-yokes in addition to the limb. As a consequence, the magnetising current and core loss values are asymmetric, to an extent depending on the path length ratio. If the central path length is one-half that of either outer, then its magnetising current is likely to be about 30% less, and this is independent of the peak flux density level.

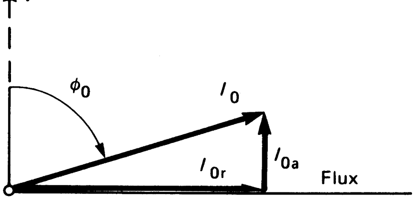


Figure 33.5 Phasor diagram of no-load current

33.2.3.1 *Magnetising inrush current*

When a transformer is 'switched-in' on no load, it may take an initial *inrush* current greatly exceeding normal no-load value, and sometimes greater than full-load current. The inrush transient decays to normal no-load level within a few periods. The first peak depends on the voltage at the instant of switching, and on the magnetic state of the core as left after the previous switching-out. If the instant of switch-on corresponds to a voltage zero, the flux must, in the first half-period, produce a complete change of $2\Phi_m$ from zero, as shown in *Figure 33.6*. The peak flux therefore rises to twice normal peak Φ_m. The maximum flux reached will be increased to $(2\Phi_m + \Phi_1)$ if there is a residual flux ϕ_1 already present in the core in the same direction as that to be taken by the first half-cycle of flux growth. These high-flux conditions demand very high peaks of exciting current with large harmonic content. As the core steel will saturate, much of the flux will follow an 'air' path, and the peak inrush current is consequently influenced by the area enclosed by the winding excited.

Under adverse conditions the magnitude and asymmetry of inrush currents may cause maloperation of overcurrent or balanced forms of protection, but in practice the worst conditions are statistically unlikely, and terminal voltage drop will reduce the peaks. In a three-phase transformer the inrush conditions must differ for each phase.

33.2.3.2 *Magnetostriction*

Magnetostriction is a property of magnetic material whereby a small change in linear dimensions (usually an elongation) follows the flux cycle in a complex pattern. In transformer steel the linear change is a few parts in a million. It causes vibration of the core at twice supply frequency and at multiples thereof, producing sound waves. Magnetostriction in the material of the core is the main source of transformer noise.

33.2.3.3 *Noise*

The unremitting hum of a transformer installed in a residential area can lead to complaints and in extreme cases to legal action. Careful design and manufacture is necessary to ensure that the noise emitted is within the level normally

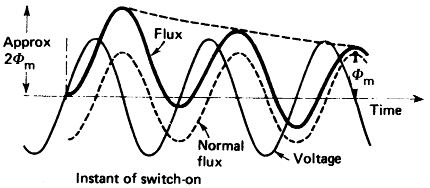


Figure 33.6 Flux transient for switching at voltage zero (no residual)

accepted as reasonable for a given size of transformer. The Electricity Supply Industry Standards (*ESI 35-1 Distribution transformers up to 1000 kVA* and *ESI 35-2 Emergency rated system transformers*) and British Electricity Board Specification (BEBS) T2 for the large units contain curves relating MVA rating and noise level which form part of the manufacturer's contractual obligations. BEAMA publication No. 227, *Guide to Transformer Noise Measurement,* is a useful reference work on the subject. IEC 60076-10 describes the way in which sound measurement tests are made.[5]

If it is expected that the noise emitted from a particular installation may create a nuisance, various mitigating measures are possible, including the following:

(1) Specifying a noise level less than the normal standard. Usually not economic (or even feasible) if reductions of more than about 10 dB below the standard level are likely to be necessary.
(2) Concealing the transformer behind a screen of trees or a wall. Sometimes the psychological effect is greater than the actual measured reduction in noise level at the point of complaint.
(3) Completely enclosing the transformer in a 'sound-proof' housing. This is expensive, as obviously special measures have to be taken to emit heat without emitting noise.
(4) A combination of (2) and (3) in which specially designed coolers form a screen wall completely surrounding the transformer. When introduced in the 1960s this appeared to be a promising development from the point of view of amenity in respect of both appearance and noise, but the idea never became popular.

In certain applications in Germany and Austria it has become necessary to install transformers with a very low noise level, up to 20 dB below the UK standard levels. In these cases it is necessary to design the transformers to operate at nominal flux densities of 1.2–1.4 T.

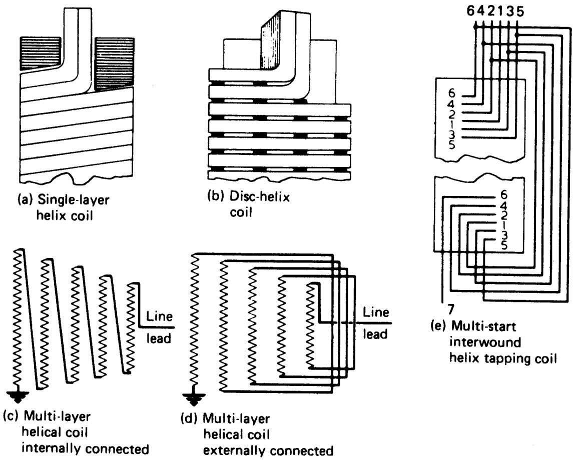
33.3 Windings and insulation

Because of their direct association with power systems, the windings and associated insulation are the most vulnerable parts of a transformer. They must be designed and constructed to withstand the voltage stresses and thermal conditions of normal service, the mechanical and thermal stresses resulting from system faults and short-circuits, and transient overvoltages such as those generated by lightning and switching. The core and windings together must meet specified impedance and loss requirements. In view of the almost total predominance in Britain of core-type transformers, the following section relates only to windings for this type.

33.3.1 Types of coil

The simplest *helix* coil consists of a single layer, formed by turns lying directly side by side, extending over the axial winding length. Each turn may comprise a single conductor or a number of conductors in parallel, the helix at each end of the coil being supported by a suitably shaped edge block to give adequate mechanical strength in an axial direction (*Figure 33.7(a)*). This type of coil, single or double layer, is used for the low-voltage windings of small and medium-sized transformers.

When the current rating necessitates a large number of conductors in parallel for each single turn, the individual strips can be laid one above the other in a radial direction, as either a single column or two columns in parallel, each turn or column separated by spacers to provide sufficient cooling surface (*Figure 33.7(b)*). This type of coil, sometimes termed a *disc helix,* is used for the low-voltage windings of large transformers.



(a) Single-layer helix coil
(b) Disc-helix coil
(c) Multi-layer helical coil internally connected
(d) Multi-layer helical coil externally connected
(e) Multi-start interwound helix tapping coil

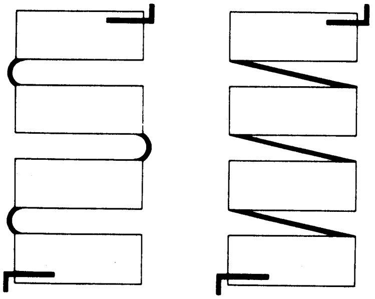
Figure 33.7 Forms of coil

Multilayer helix coils can be employed for the high-voltage windings of large transformers. Because of the high capacitance between the individual layers, this type of winding has good inherent strength against incoming surge voltages. It is fairly simple to calculate the surge response of the multilayer winding, but difficult to ensure that the long thin layers of conductors have adequate strength to resist axial forces. As ratings increase for a given primary service voltage, a *disc* coil (see below) becomes relatively less difficult to design from the point of view of predicting surge strength and more satisfactory because of its high inherent strength against axial forces. The use of the multilayer helical coil is thus generally confined to the smaller megavolt-ampere ratings of any given voltage class above about 200 kV.

This type of coil shows particular advantage when the transformer neutral is directly earthed. The inside layer of the winding is then near earth potential, thus reducing the major insulation between high- and low-voltage windings. The length of the individual layers decreases progressively towards the outer layer to provide increased insulation to-earth at the line end. Interlayer connections, usually top-to-bottom, can either be arranged internal to the coil between the layers, or formed by external joints (*Figure 33.7(c, d)*). Multilayer coils are often used for the high-voltage windings of small distribution transformers, in this case wound as continuous layers, with top-to-top and bottom-to-bottom interlayer connections.

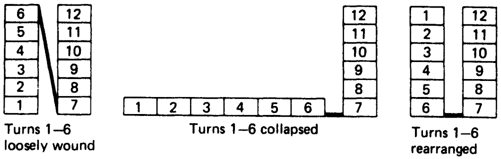
Another form of layer coil is a *multistart interwound* helix employed for the separate tapping windings of large transformers when a considerable range of voltage variation is required. The coil is so arranged that each single conductor forms one tapping section, with the requisite turns distributed over the winding length to provide axial ampere-turn balance for the various tapping positions. The individual conductors are physically located so that the voltage between them is reduced to a minimum (see *Figure 33.7(e)*).

A *crossover or bobbin* coil is wound on a former between side cheeks. It is, in effect, similar to a multilayer winding, except that the layers are short. The interlayer insulation needs to be extended at the ends to guard against failure by creep at the edges; alternatively, it can be folded round the last turn of each layer, or the overhang crimped to form a trough of depth equal to the thickness of the conductor. The assembly of separate coils is connected in series, with horizontal cooling ducts provided by spacers between the individual coils. Intercoil connections are usually made back-to-back and front-to-front, with adjacent coils reversed. Back-to-front connection requires insulation of the intercoil connections corresponding to the voltage developed across the coil (see *Figure 33.8*). Cross-over coils, generally of round wire, are employed for the high-voltage windings of distribution transformers up to about 1000 kVA.

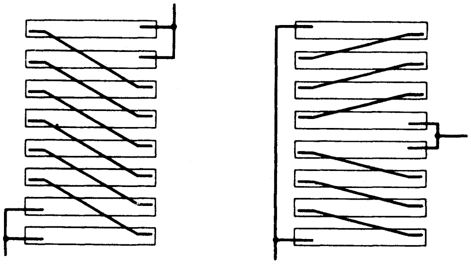

Front-to-front, back-to-back alternate coils reversed

Back-to-front

Figure 33.8 Cross-over coils: intercoil connections



(a) Steps in the formation of a disc coil pair


(b) Parallel disc pairs

(c) Parallel halves

Figure 33.9 Disc coil windings

Because of the increase in the cost of copper, more attention is being paid to the use of aluminium for transformer windings. Aluminium is a less effective conductor than copper, but despite this intrinsic disadvantage (which results in a larger transformer for a given efficiency) there is a net saving in overall cost. Aluminium strip can be used to wind any of the types of coil described above, but it is also uniquely used in foil form. This is because aluminium can be rolled to a thinner and more flexible foil than copper. The foil is used in bobbin-type coils of one turn per layer with intercoil connections as shown in *Figure 33.8*.

The *disc* coil differs from other windings in that adjacent turns, consisting of strip conductor, are wound one above the other in a radial direction from the centre outwards; thus, the coil might be compared with a cross-over coil comprising one turn per layer. To achieve the required disposition of turns, the coils are formed in pairs, the requisite turns for one disc being loosely wound so that the conductor finishes in a position to provide the start of the inside turn of the second disc, which can then be wound from the inside outwards, the turns of the first disc being rearranged by folding one inside another in such a manner that the start of the coil is located as an outside turn (*Figure 33.9*(*a*)). The coil is then tightened to remove slack from the reformed disc.

Disc coils can be wound either as individual pairs, which are then assembled and connected in series by external joints, or as a continuous winding where, following formation of one pair, the procedure is repeated without cutting the conductor or removing the coil from the lathe. The continuous coil avoids assembly and joining and saves joint space, an advantage with an inside winding. For large currents the coils can be wound with multiple conductors, or constructed as separate parallel connected pairs (*Figure 33.9*(*b*)). Alternatively, two half-windings, one half reversed with respect to the other, can be stacked as in *Figure 33.9*(*c*) with the line connection taken from the centre of the stack; this method is common for large high-voltage transformers having a directly earthed neutral. Horizontal cooling ducts are provided by spacers between individual discs and between disc pairs.

The disc coil is a general-purpose winding element, applied to the higher-voltage windings of transformers above 1 MVA up to the highest ratings and voltages, and also for the lower-voltage windings (normally 33 kV upwards) of medium and large power transformers. *Table 33.2* gives a survey of typical applications of the various forms of winding.

33.3.2 Insulation

Materials of classes A, E, B, F, H and C (in temperature sequence) are recognised in IEC 60076 as being suitable for transformers. A, B, H and C are most common, but F is being used to an increasing extent. All these are employed in dry-type transformers, the silicon treated materials being

Table 33.2 Transformer windings

Service	*Rating* (MVA)	*High-voltage winding*		*Low-voltage winding*	
		kV	Type	kV	Type
Distribution	<1	11, 33	Foil, cross-over or multilayer	0.43	Helix
System	1–30	33, 66	Disc	11	Disc, helix or disc-helix
Transmission	>30	132–500	Disc or multilayer	11, 33, 66	Disc or disc-helix
Generator	>30	132–500	Disc or multilayer	11–22	Disc-helix

advantageous (but at extra cost) by reason of their water-repellent properties. For windings immersed in hydrocarbon oil (BSEN 60296) or in synthetic liquids, the coil insulation is usually a class A material. Cotton is confined to small units, with paper and paper derivatives used for most purposes up to the largest ratings; but synthetic enamel is widely used as interturn insulation in small transformers and in the low-voltage windings (up to about 17.0 kV) of large transformers.

Most power transformers are immersed in hydrocarbon oil. To reduce fire risk, chlorinated biphenyls (e.g. Pyroclor) were used in the past, but as these liquids are toxic (and difficult to dispose of safely), they have virtually been banned on environmental grounds. Silicone based liquids and synthetic esters are available without these disadvantages, but they are relatively expensive, and the varnishes and binders usable with hydrocarbon oil may not be suitable for use with silicone fluids. Where fire risk is unacceptable, air-cooled 'dry'-type designs with glass and epoxy resin insulation are usually preferred. Gas filled transformers using electronegative sulphur hexafluoride gas (SF_6) at low pressure for distribution transformers and up to 6 atm for high-voltage power transformers eliminate fire risk, but the cost is high, the force-cooled heat exchangers are complicated, and the transformers have short thermal time constants, giving reduced overload capability. However, they are particularly suited to underground installations, and where land prices are high they can be justified when the cost evaluation recognises that no fire-fighting equipment is necessary.

33.3.2.1 Insulation design

For transformers of small and medium size, the inner (low-voltage) winding is insulated from the core by pressboard or a synthetic resin-bonded paper (s.r.b.p.) cylinder, with axial bars of pressboard or equivalent material arranged round it to form cooling ducts for the inside surface of the winding. With disc or disc-helix coils, the bars have a wedge section on which intercoil or interturn dovetail-slotted spacers can be threaded. Similar bars are placed over the outer surface of the inner winding and, if necessary, between layers of a helix winding. The main high-voltage/low-voltage insulation is provided by another pressboard or s.r.b.p. cylinder. The arrangement of bars and spacers is repeated for the outside (high-voltage) winding. Insulation to earth at the ends of the windings takes the form of blocks keyed to the axial bars in line with the spacers, to form a series of columns round the windings by which the windings can be effectively clamped. The cross-section of a typical arrangement is shown in *Figure 33.10*.

Insulation arrangements for dry-type transformers are similar to those described, except that the materials chosen are appropriate to the insulation class. For example, the cylinders and bars can be made from suitably impregnated glass fibre; ceramic materials may be used for the coil spacers and glass tape for interturn insulation.

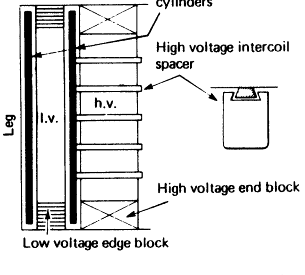
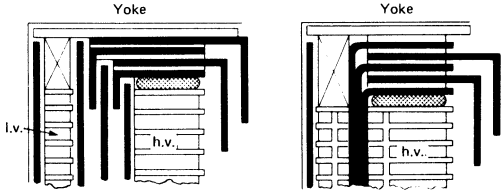
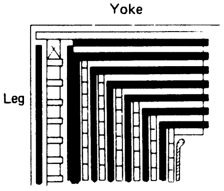


Figure 33.10 Typical insulation for small and medium-sized transformers



(a) s.r.b.p. oil barrier high to low insulation

(b) 'Solid' high to low insulation



(c) Arrangement of multi-layer helical High voltage winding

Figure 33.11 Insulation design

When operating voltages are such that thick major insulation is required between windings, it is customary to use a number of concentric thin-walled cylinders, spaced by axial bars, the insulation being carried round the ends of the outer (high-voltage) winding by means of flanged collars interfitting with the ends of the tubes (*Figure 33.11*(*a*)). Alternatively, with high-voltage windings at above 110 kV, so-called *solid* insulation can be used, comprising layers of pressboard or paper wound directly over the inner winding, the high-voltage winding being assembled directly over the outer layer, with cooling ducts transferred from the inside turns to some point in the winding. In this case the pressboard or paper wraps are extended beyond the axial winding length and flanged at the ends to assist with the insulation to earth (*Figure 33.11*(*b*)). A combination of s.r.b.p. cylinders and pressboard wrappings can also be employed, the wrappings being located over the outer cylinder again flanged at the ends. Further, the s.r.b.p. tube can be manufactured such that the bonding is limited to the central portion, leaving the ends plain so that these in turn can be flanged.

With multilayer helical high-voltage windings, the interlayer insulation is provided in the form of a combination of paper or pressboard wrappings, and bars for cooling ducts, with the paper between the individual layers flanged at the ends (*Figure 33.11*(*c*)). When internal interlayer connections are involved (*Figure 33.7*(*c*)), the interlayer paper wrappings may be formed in tapered halves so that the maximum thickness of insulation is provided at appropriate places between the actual connection and the layers. Paper is used almost exclusively for the interturn insulation of high-voltage windings.

33.3.3 Winding design

The main factors to be taken into account in winding design are:

(1) To provide adequate insulation strength to withstand (a) power frequency applied or induced overvoltage tests to prove the insulation to earth and between windings; (b) an induced voltage test for internal winding insulation (interturn, intercoil and between phases);

(c) an impulse voltage test to prove the ability of the insulation structure to withstand transient overvoltages such as may result from atmospheric surges; and (d), in some cases, a switching surge test to prove the strength to resist surges arising from switching operations. At system voltages of 500 kV and above switching surges rather than lightning control the design.

(2) To ensure that the load loss (i.e. the sum of the I^2R and stray losses) does not exceed the guaranteed performance figure.
(3) To provide adequate cooling to meet guaranteed temperature rise limits.
(4) To ensure adequate short-circuit strength.
(5) To achieve the required impedance characteristic.

Transformer windings may be fully insulated, where the insulation to earth at all points will withstand the separate source voltage test specified for the line terminals, or have graded insulation, where the insulation to earth is reduced from that required at the line terminals to a smaller amount at the neutral end. The insulation on a graded insulated winding will withstand only a separate-source voltage test corresponding to the insulation level at the neutral. In the latter case an induced voltage test is employed to test not only the internal insulation and that between phases, but also the insulation to earth at the line end.

The required level of insulation at the line end is normally designated as the *basic insulation level* (b.i.l.). This value is specified by the purchaser of the transformer and is determined by taking into account the maximum expected atmospheric or switching surges which may be imposed on the transformer, plus a small margin of safety.

The b.i.l. for any particular installation at a given system voltage is governed by the type and effectiveness of the protection against lightning and the combination of line and circuit-breaker characteristics that control the switching surge limit.

For systems where the neutral is isolated, or earthed through an impedance (e.g. an arc-suppression coil) such that, during a line-to-earth fault, the voltage to earth of the unfaulted lines can exceed 80% of the normal line-to-line voltage, fully insulated systems are essential. They are called *non-effectively earthed* systems. Graded insulation is permitted in effectively earthed systems, where the method of earthing at each transformer is such that the 80% value is not exceeded for any operating condition. The requirement is met when the zero phase sequence/positive phase sequence (z.p.s./p.p.s.) reactance ratio of the system is less than 3, and the z.p.s./p.p.s. resistance ratio is less than 1.

Uniform insulation must be provided for all delta-connected windings, and for star-connected windings where the neutral is not earthed. Grading may be applied in the latter case if the neutral is earthed. Where earthing conditions permit grading, they also, in general, allow forms of overvoltage protection for which test voltage levels are lower.

Over the past 50 years a major effort has been made by the International Electrotechnical Commission (IEC) to arrive at agreed international standard test voltages for lightning and switching surge impulse levels and the associated power frequency induced voltage test levels. The effort to accommodate all the different test levels already established by different supply authorities on the basis of their own assessment of the insulation level required on their systems to minimise total cost has led to a complicated set of tables, including a variety of alternative choices for the test voltages appropriate to the various system voltages. IEC 60076: Part 3 includes over six pages of tables and text relating thereto and reference should be made to this document if details are required.[6]

Two fundamental points arise:

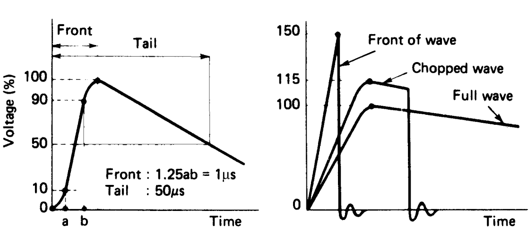
(1) The adoption of any one so-called 'standard' insulation level for a particular transformer installation is, of itself, of little or no value. There are at least ten major variables involved in the design of a large transformer (e.g. MVA rating, frequency, system voltages on both high and low voltage, and impedance) and the inevitable differences that arise in any one or more of these mean that each order will require a unique new design. It is of little importance whether or not the specified insulation level is one of the so-called 'standards levels'. The designer will provide what is appropriate.
(2) At the highest voltage levels (>700 kV) there is growing evidence that the trend (particularly in the USA and Canada) towards progressively lower insulation levels for a given system voltage has gone too far. In its simplest terms a reduction in impulse test level represents a reduction in the margin of safety against insulation failure.

References 7 and 8 both include details of recent changes in customers' specification to call for significantly higher impulse test voltages than those previously specified and which have now proved to be inadequate for reliable service.

33.3.3.1 Surge voltage distribution

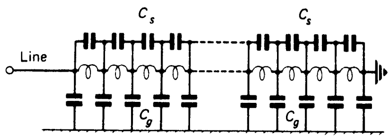
The voltage distribution in power frequency voltage tests is substantially uniform between turns and coils, and corresponds to the normal service condition in respect of voltage to earth. Under impulse test, however, the distribution can be far from uniform. An impulse voltage wave has a steep front and a long tail. The standard form, defined in IEC 60071, has a front rising from zero to peak value in 1.2 µs, and falling thereafter on the tail to 50% of peak value in 50 µs. This is termed a 1.2/50 µs wave (*Figure 33.12*(*a*)). Impulse voltage tests on transformer windings include the application of a *full wave* and a *chopped wave* impulse, the latter being a full wave shortened by sparkover of a rod gap or its equivalent. American practice adds a front-of-wave test, in which a voltage rising at approximately 1 MV/µs is chopped on the wavefront. The three forms of impulse test voltage are shown in *Figure 33.12*(*b*).

A full wave application tests the ability of the insulation structure to withstand voltage surges; a chopped wave test simulates stresses that occur on the collapse of a surge tail by operation of a rod gap or a flashover to earth. Transformers for systems operating at 300 kV or above may be also required to demonstrate ability to withstand switching surges, typically by a 100/1000 µs wave (see Section 33.14).



(a) Standard 1/50 µs impulse wave (b) Forms of impulse wave

Figure 33.12 Impulse testing



(a) Equivalent circuit of transformer winding

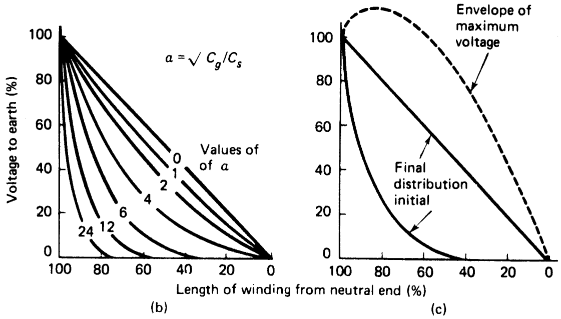


Figure 33.13 Voltage distribution

Ignoring resistance, a transformer winding and its surroundings may be represented by an inductor-capacitor network (*Figure 33.13*(*a*)). When a steep-fronted surge is applied to the line terminal, the initial voltage distribution is governed solely by the capacitance network, in particular by the ratio $\alpha = \sqrt{(C_g/C_s)}$ of the capacitance to earth and in series. The greater the value of α, the greater is the divergence from uniform voltage distribution from line to ground, as shown in *Figure 33.13*(*b*). For a uniform winding of identical sections having the same capacitance values C_s and C_g, the initial voltage to earth at any point x from the remote end (with x expressed as a fraction of the winding length) is:

isolated-neutral winding $e_x = E(\cosh \alpha x/\cosh \alpha)$

earthed-neutral winding $e_x = E(\sinh \alpha x/\sinh \alpha)$

for a surge voltage peak of E.

If the surge voltage is maintained, an approximately uniform voltage distribution will be reached, but between these two states (if they differ) a complex array of damped oscillations will occur, creating abnormal stresses in the insulation (*Figure 33.13*(*c*)).

A chop occurring between 3 and 10 µs after application of the impulse voltage will result in augmented interturn and intercoil stresses, although the voltages to earth are normally reduced. The stressing again depends on the initial distribution and on the actual chopping instant. The chop may be regarded as a unit-function voltage, of polarity opposite to that of the incoming surge, and superimposed on the conditions existing within the winding at the instant of chopping.

With the front-of-wave test a higher voltage is applied, but its duration is shorter and the increased stresses are generally confined to the entrance insulation (particularly the bushing). Internal voltage stresses are usually less than those arising from a chopped wave test.

The ratio α is high for a small transformer. For transformers of higher rating and especially for higher line voltage, C_g decreases because it is determined largely by clearances, while C_s increases because of the greater radial depth of the winding. Thus, $\alpha = \sqrt{(C_g/C_s)}$ is lower and the initial voltage distribution approximates more closely to the uniform.

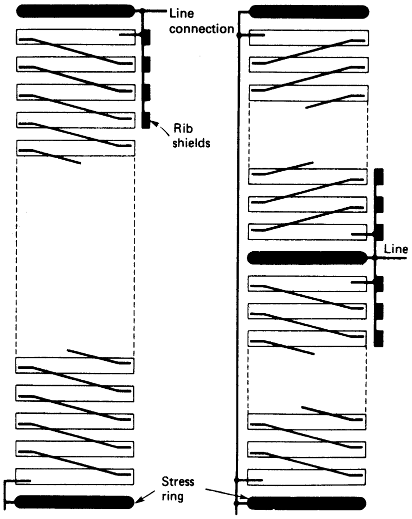


Figure 33.14 Disc windings fitted with stress rings and rib shields

To reduce stress concentrations at the ends, disc-coil windings are provided with stress rings that act as radial shields, although they do not materially improve the axial distribution. Axial improvement can be gained by the addition of rib shields to the line-end turns (*Figure 33.14*) or by several other means of controlling the electric field distribution.

Compared with a disc winding, a multilayer coil has a higher series capacitance and therefore a superior transient distribution, although electric field shielding at the ends of the layer may be necessary. The choice between disc-coil and multilayer windings depends, therefore, on the specified impulse level and on the rating.

33.3.4 Leakage field

Problems related to leakage field are significant factors in determining whether a transformer will be reliable in service. *Figure 33.15* shows a cross-section of the core and windings of a 240 MVA autotransformer and a flux plot of the leakage field set up by the load currents in the windings. The critical areas, where special precautions would be necessary to prevent local heating, are the tank wall, the coil support bracket and, to a lesser extent, the face of the core.

A precise knowledge of the shape and magnitude of the leakage field is essential for calculation of impedance, determination of stray loss and consequent heating and the magnitude and direction of short-circuit forces.

Even with modern computer programming, it is difficult to obtain an accurate three-dimensional plot of the leakage field. The computer gives a good two-dimensional plot but approximation and simplification are necessary to determine the third dimension. In spite of this limitation, however, numerical calculation gives the best results in determining leakage field strength with the best results obtained using Finite Element numerical solutions. The difficulties in calculation arise because of the irregular geometry of the windings and the presence of both non-magnetic and magnetic (with non-linear characteristics) components within the field.

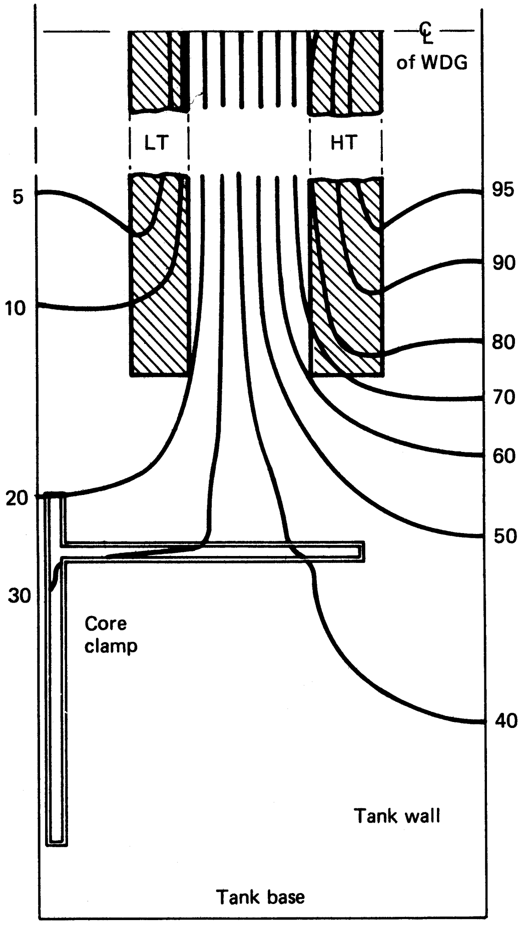


Figure 33.15 The leakage flux point of a 240 MVA transmission auto-transformer, showing leakage flux concentration in the tank wall and the bottom yoke core clamp

33.3.5 Impedance voltage

The impedance voltage of a transformer can be defined as that voltage required to circulate full-load current in one winding with the other winding(s) short-circuited. It comprises a component to supply the *IR* drop and another to overcome the e.m.f. induced in leakage inductance. On larger transformers the resistance component is usually negligible and the percentage value of the impedance is the ratio between total magnitude of the leakage flux and the main flux in the core. The leakage flux is a function of the winding ampere-turns and of the area and length of the paths of the leakage flux. By adjustment of these parameters the transformer can be designed for a range of reactances. The most economical arrangement of core and windings results in a 'natural' value of reactance. This value can be varied to some limited extent without any great influence on the cost and performance of the transformer. Interleaving the windings and so reducing the effective area of the leakage paths will reduce reactance. High reactance requirements usually result in greater stray load loss, because of the necessarily greater leakage flux.

It is usual to express the leakage impedance Z of a transformer as a percentage (or per-unit) value. The per-unit value is IZ/V, and the percentage value is 100 (IZ/V), where I and V refer to the full-load current and rated voltage of one of the windings, and IZ is the voltage measured at rated current during a short-circuit test on the transformer. In the case of a transformer with tappings, the impedance is conventionally expressed in terms of the rated voltage for the tapping concerned.

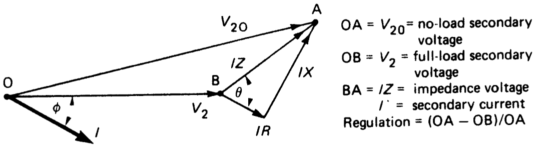

A
O
B
V_{20}
V_2
IZ
IX
IR
θ
ϕ
I
OA = V_{20} = no-load secondary voltage
OB = V_2 = full-load secondary voltage
BA = IZ = impedance voltage
I = secondary current
Regulation = (OA − OB)/OA

Figure 33.16 Regulation

33.3.5.1 Regulation

The regulation is the difference between the no-load and full-load secondary voltages expressed in terms of the former, with constant primary voltage. The difference is the result of voltage drops due to resistance and leakage reactance; it is proportional to load but is strongly influenced by the load power factor. *Figure 33.16* gives a diagram to show the voltage drop IZ in the leakage impedance Z of the transformer (as obtained from a short-circuit test). The regulation is a maximum when the load phase angle ϕ is equal to the angle θ in the impedance triangle of R, X, and Z. If the load is leading reactive, the regulation is reduced; it may even be reversed so that the secondary voltage rises on load. Formulae (IEC 60076) for the regulation ε of a two-winding transformer on full-rated load, in terms of the percentage IR drop voltage ε_r and the percentage reactance voltage ε_x are:

$$\varepsilon = \varepsilon_r \cos\phi + \varepsilon_x \sin\phi + (\varepsilon_x \cos\phi - \varepsilon_r \sin\phi)^2/200\% \qquad (33.1)$$

for transformers having impedance voltages up to 20% and

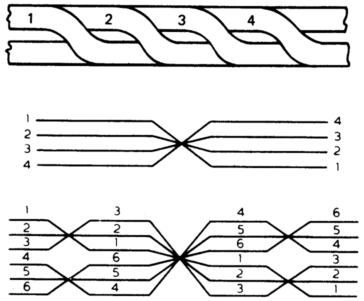
$$\varepsilon = \varepsilon_r \cos\phi + \varepsilon_x \sin\phi\% \qquad (33.2)$$

is a simplified expression for cases in which the impedance voltage does not exceed 4%. For unity p.f. load, (33.1) reduces to $\varepsilon = \varepsilon_r + \varepsilon_x^2/200$ and (33.2) reduces to $\varepsilon = \varepsilon_r$.

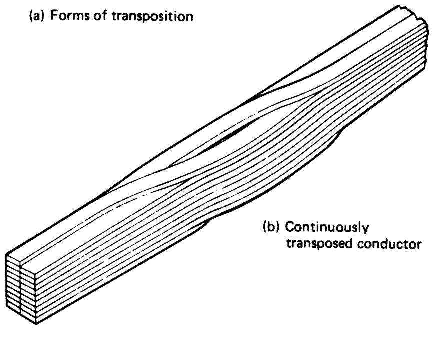
33.3.6 Losses

The load (or 'copper') loss comprises two components; a direct I^2R loss due to ohmic resistance of the windings, and a stray loss arising from eddy currents in the conductors due to their own flux, influenced by the tank and by steel clamping structures. The eddy loss is negligible when the section of the conductor is small. When the current is too great for a single conductor without excessive eddy loss, a number of strands must be used in parallel. Because the parallel components are joined at the ends of the coil, steps must be taken to circumvent the induction of different e.m.f.s in the strands, which would involve circulating currents and further loss. Forms of conductor transposition have been devised for this purpose.

Ideally, each conductor element should occupy every possible position in such a way that all elements have the same resistance and the same induced e.m.f. Transposition, however, involves some sacrifice of winding space. If the winding depth is small, one transposition halfway through the winding is sufficient; or in the case of a two-layer winding,



(a) Forms of transposition



(b) Continuously transposed conductor

Figure 33.17 Conductor transposition

at the junction of the layers. Windings of greater depth need more transpositions.

Typical forms of transposition are shown in *Figure 33.17*. The methods apply mainly to helical coils. In disc windings where there are two or more conductors in parallel, the connections between the discs can be arranged to give the necessary effect.

Stray loss may also be produced by radial components of leakage flux, but can be minimised by careful ampere-turn balance of the windings.

33.3.6.1 Efficiency

Efficiency is the ratio between power output and power input. Its actual value is less important than that of the magnitude of the losses, which determine heating, cooling, rating and the cost of supplying the losses under given loading conditions in a system.

For a given system voltage and transformer tapping, the flux density has constant peak value, and the core loss p_i is considered to be constant. The load loss p_c varies with the square of the loading, and because of the change of conductor resistivity with temperature is commonly stated at 75 °C. If the power input is P_1 and the power output is P_2, the efficiency is

$$P_2/P_1 = P_2/(P_1 + p_i + p_c) \text{ per unit.}$$

33.3.7 Cooling

The majority of transformer windings are cooled by thermal circulation of oil. Even where pumps are fitted to provide forced cooling, it is common for the transformer to operate without the pump up to some proportion of its forced cooled rating. It is, therefore, important to ensure that the natural thermal circulation of the oil is such that there is a flow over the main cooling surface of the coils. This is a simple matter where the main cooling surfaces are vertical, but when (as in disc windings and particularly on large transformers), the major cooling surfaces are horizontal, special means must be taken to ensure that the oil will flow across the horizontal surface under both natural and forced cooling conditions. Oil-flow barriers are introduced for this purpose.

33.3.8 Short-circuit conditions

Under conditions of system fault, mechanical and thermal stresses of considerable magnitude can be imposed on a transformer winding.

Mechanical forces of magnetic origin may be resolved into two components: (1) *radial*, due to the coil currents lying in the axial component of the leakage flux, tending to burst the outer and crush the inner winding; and (2) *axial*, due to the radial component of the leakage flux arising from ampere-turn unbalance, tending to displace the windings (or parts thereof) with respect to each other. Axial forces may increase the unbalance that produces them, and repeated short circuits may have a cumulative effect. Windings must be designed and built to withstand the mechanical forces, an important aim being to minimise ampere-turn unbalance between windings, especially that caused by tappings.

Thermal stresses arise from the temperature attained by the windings when carrying sustained fault current. The limits of temperature rise permitted by IEC 60076 for copper windings are: 250°C for class A insulation in oil and 350°C for class B in air. Aluminium windings are limited to 200°C for both classes.

Mechanical forces have to be considered relative to the asymmetrical peak current, whereas thermal stresses are governed by the symmetrical r.m.s. value and duration. The fault current is controlled by the leakage impedance of the transformer and the impedance of the supply system. IEC 60076-5 states that for transformers rated 3150 kVA or less, the system impedance shall be neglected in the total impedance if it is equal to or less than 5% of the leakage impedance of the transformer.

The duration for which a transformer can withstand short-circuit current depends on thermal considerations. IEC 60076-5 states that the duration of the short-circuit current to be used for calculation of the thermal ability to withstand short circuit is 2 s unless otherwise specified by the purchaser. The temperature attained by the winding can be calculated by the method described in IEC 60076-5 on the assumption that, during the period of short-circuit, all heat developed by the loss in the windings is stored in the conductor material.[9]

33.4 Connections

Transformer windings can be constructed for connection to a one-phase, two-phase or three-phase power supply. Combinations are also possible, namely three-to-two or three-to-one phase conversion. Six-, 12- and even 24-phase connections may be needed in rectifier transformers. Single-phase transformers can be independent; arranged to provide a two- or three-wire supply, centre-point earthed, or combined to form a three-phase bank. In the case of three-phase windings, three forms of connection are possible; star, delta and interconnected-star (zig-zag). When combined on the same core, a delta winding and an interconnected-star winding can be arranged to provide zero phase displacement, and when either of these is combined with a star winding, a 30° phase displacement results, either leading or lagging. By reversal of one winding with respect to the other when a combination of the same connection is involved, or where the combination is of connections giving the same phase displacement, a 180° phase displacement is produced. A summary of the combinations detailed in IEC 60076 and the corresponding e.m.f. diagrams is given in *Figure 33.18*.

In a star/star connection unbalanced load may result in neutral displacement and third-harmonic currents may circulate between lines and earth. These difficulties may be overcome by providing a delta connected stabilising (tertiary) winding with a rating sufficient to take short-circuit fault currents. The need for this winding depends, however, on the core construction. Where, as in banks of three one-phase units or in five-limb three-phase core-type transformers, there is an independent iron path for the zero-sequence flux in each phase, the z.p.s. impedance is consequently high, making a stabilising winding essential. With three-phase three-limbed cores the z.p.s. fluxes are forced out of the core and the z.p.s. impedance is lower; consequently, a tertiary winding may not be necessary.

33.4.1 Phase conversion

The Scott connection is the most common two-to-three phase conversion. Two one-phase transformers are generally used, with one pair of windings arranged to form a T connection for the three-phase supply. The 'main' unit is wound for the phase-to-phase voltage with a mid-point tapping, and forms the head of the T. The other unit, the 'teaser', is wound for 0.866 times the line voltage and both are designed to carry full line current. A neutral point can be provided by a tapping on the teaser winding at a point 0.577 of the turns from the line end (see *Figure 33.19*). The two units, for operational convenience, can be made interchangeable. A winding arrangement which results in a minimum leakage reactance between the winding halves of the main unit (for example, an interleaved winding) is essential to ensure correct current distribution and to avoid excessive voltage regulation. In the place of two single-phase units, a three-limb core, outer legs wound, may be used if the unwound centre limb is proportioned to suit the flux conditions, i.e. given 41.5% greater section.

As an alternative to the Scott connection, which is basically a one-phase arrangement, a Leblanc connection can be used. It employs a standard, three-limb, three-phase core and a standard three-phase delta connected winding (*Figure 33.20*). Compared with the Scott arrangement, the Leblanc connection requires a smaller core but involves a greater winding section.

Three-phase to single-phase transformation can be achieved by an open delta connection which involves a standard, three-limb, three-phase core, with outer limbs wound. Alternatively, an arrangement similar to the Scott connection can be used. With the open delta connection the output voltages from each limb are 120° apart, so that the voltage applied to the load is 3 times the voltage across one winding, and in the case of the T connection the output voltages are 90° apart, so that the voltage applied to the load is 2 times the voltage across one winding. With either form of connection, if a three-wire one-phase supply is required and the loads on each side of the mid-point are liable to unbalance, the windings should be subdivided and interconnected to distribute more evenly the out-of-balance current and to avoid excessive voltage regulation.

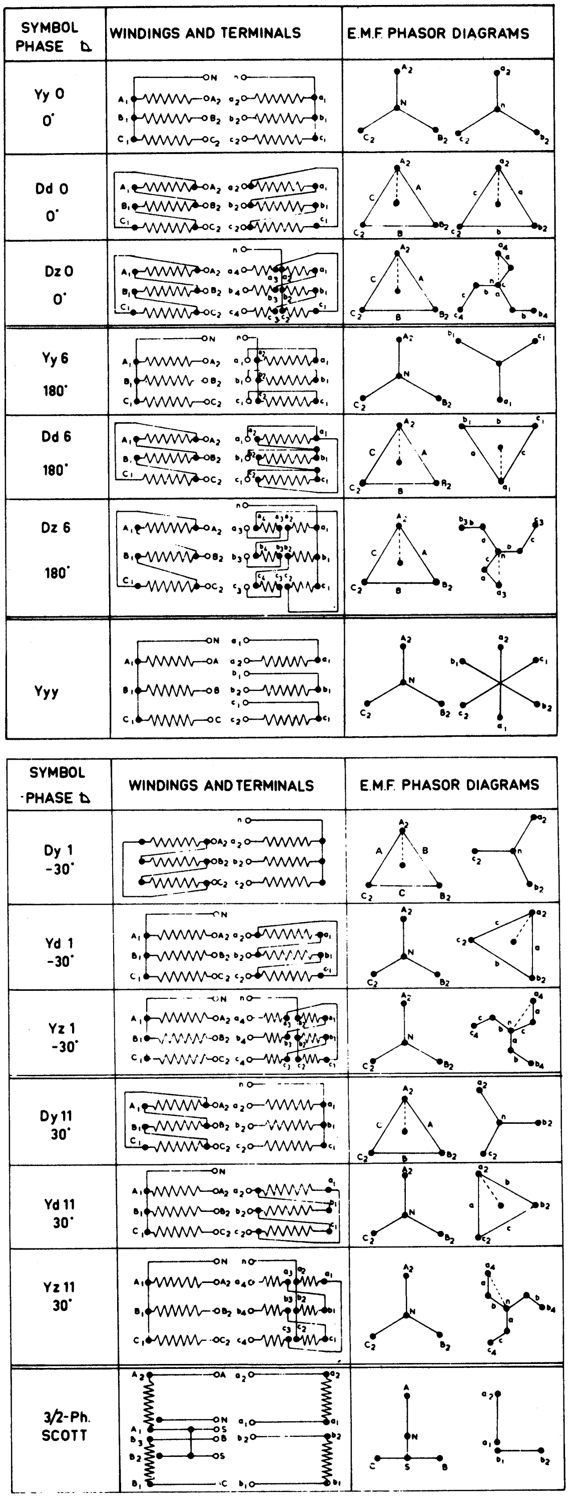


Figure 33.18 Standard connections for three-phase transformers

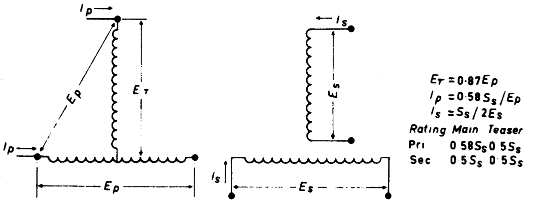


Figure 33.19 Scott three-to-two phase transformation

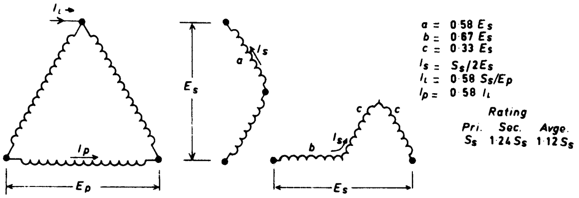


Figure 33.20 Leblanc three-to-two phase transformation. S_s is the kilovolt-ampere rating of the two-phase load

It is important to realise that, although a three-to-one phase connection can be used to transform a line–neutral load on the secondary side to a line–line current on the primary side, it does not result in a balanced primary load. The same degree of unbalance must appear on the primary; or in terms of symmetrical components, all zero- and negative-sequence currents on the secondary will also flow on the primary side.

33.5 Three-winding transformers

Typical applications of three-winding transformers are:

(1) feeding two secondary networks, of different voltage or phase relationship, from a common primary supply;
(2) connecting two generators to the same high-voltage system while maintaining a relatively high impedance between them to limit cross-feed of fault energy;
(3) feeding two parts of a sectionalised low-voltage network so as to limit the fault level of each part without too high an impedance between the high-voltage and the low-voltage sides; and
(4) providing a rectifier with a multiphase (e.g. 12- or 24-phase) supply.

Provided that there is at least one delta connected winding, to permit the flow of third-harmonic currents, there is freedom to adopt star or delta connection to meet phasing or earthing requirements.

33.5.1 Impedance characteristics

Two-winding technology does not apply. The essence of the procedure for a three-winding unit is that the leakage impedance can be represented by assuming each of the three windings to have an individual resistance and leakage reactance, and mutual impedance effects (other than those that result from these individual values) to be absent. The equivalent circuit can be represented by the star network in *Figure 33.21*. The leakage impedance values (in per-unit form to a common kilovolt-ampere base) are given in terms of the conventional two-winding impedances: for resistances

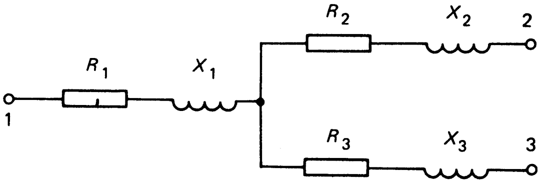


Figure 33.21 Equivalent circuit of a three-winding transformer

$$R_1 = \frac{1}{2}(R_{12} + R_{31} - R_{23})$$

$$R_2 = \frac{1}{2}(R_{23} + R_{12} - R_{31})$$

$$R_3 = \frac{1}{2}(R_{31} + R_{23} - R_{12})$$

X being substituted for R to give the leakage reactances. These for the individual arms are then combined to give the effective values between any pair of terminals: e.g. $R_{12} = R_1 + R_2$, $X_{23} = X_2 + X_3$, and so on. (As the equations for the individual arms include a negative term, some particular evaluations may be found to be negative.)

33.5.2 Tertiary windings for harmonic suppression

The most common three-winding transformer is star/star connected with a delta tertiary to provide a path for z.p.s. currents. If third-harmonic distortion of the main flux (and, hence, of the secondary voltage wave form) is to be avoided, a tertiary must be used on any configuration of core that provides a low-reluctance path for third-harmonic components of the flux. This requirement applies to shell types, and to three-phase core types having more than three limbs; a three-phase three-limb core-type transformer has no ferromagnetic return path for third-harmonic flux components, so suppressing z.p.s. distortion of the flux, and it will operate satisfactorily provided that the load is not significantly unbalanced. For an unbalance exceeding 10%—or if a low zero sequence impedance is required for protection purposes—it would be prudent to include a tertiary. However, a two-winding transformer called upon to supply zero-sequence loads could alternatively be provided with an external source of zero-sequence power (a) by a direct connection between the transformer in question and the neutral of a stand-by unit having a delta winding, or (b) by the installation of a zig-zag connected 'earthing transformer'. If several star/star units are acquired and undue unbalanced loading is not expected, it would be preferable on grounds of cost to specify them without tertiary windings, subsequent remedial action being taken if and when it becomes desirable. Whether or not third-harmonic problems arise depends on the complete installation and how it is operated. The transformer characteristics do not alone determine the issue; other factors are (a) the level of the transformer core flux density (and, hence, the magnitude of the third-harmonic component of magnetising current); (b) the expected degree of load unbalance; (c) the impedance of external sources of zero-sequence current; and (d) the proximity of telecommunication circuits to the external zero-sequence current path. Tertiary windings fitted only for harmonic suppression and not connected to external terminals must be rated to withstand the effects of primary and secondary earth fault currents: tertiary current then depends on the p.p.s., negative phase sequence (n.p.s.) and z.p.s. impedances of both the supply and the transformer. As such fault currents persist for only a few seconds, the temperature rise of the tertiary winding is determined by its thermal capacity.

33.5.3 Tertiary windings for external loads

Tertiary windings providing auxiliary power circuits, or supplying reactive power compensating capacitors or inductors, must have an appropriate continuous rating. The tertiary must withstand the effects of a short-circuit fault across its external terminals, as well as those due to earth faults on the main windings. The conductor sections and the bushings are designed for the most onerous operating condition.

33.6 Quadrature booster transformers

Consider two lines in parallel transmitting power between A and B (*Figure 33.22*): T_1 and T_2 are conventional transformers, either or both having on-load tap changing equipment. The division of load currents I_1 and I_2 between the lines is in inverse proportion to the impedances Z_1 and Z_2. For better load sharing it may be desirable to adjust the division of current. Tap changing the transformers to give a voltage difference V between T_1 and T_2 will cause a circulating current $I_t = V/(Z_1 + Z_2)$ to flow around the loop; and I_t will lag V by the natural phase angle of $Z_1 + Z_2$ increasing the current and I^2R loss in the branches without achieving the desired change in load sharing.

If a booster transformer T_b is introduced into the loop to inject a small voltage in quadrature with the supply voltage, the resulting circulating current will be approximately co-phasal with the active-power component of the load current. It will therefore add arithmetically to the active current in one branch and subtract from that in the other, thus controlling the relation between I_1 and I_2. The adjustment permits an increase in the total load that can be transmitted over the system. The quadrature booster is an important element for power control in parallel circuits. A simple method of deriving the necessary injected voltage is shown for one phase in *Figure 33.23*. A single transformer unit has a delta connected main winding and a tapped series winding to provide an adjustable voltage in quadrature with the phase voltage. For large high-voltage quadrature boosting, two transformers are commonly used in order to provide the necessary dielectric and short-circuit strengths. The first 'excitation' transformer supplies the second 'injection' transformer. By suitable choice of connections either quadrature or in-phase regulation can be obtained. By use of two tap changers and more complex tapping windings, separate control of both in-phase and quadrature components can be achieved.

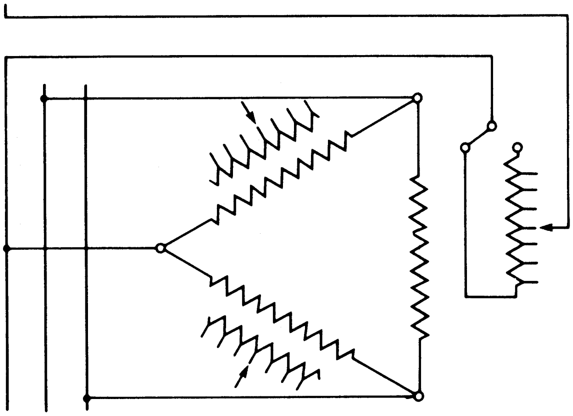


Figure 33.23 Connections for quadrature regulation

33.7 On-load tap changing

The essential feature of all methods of tap changing under load is that circuit continuity must be maintained throughout the tap stepping operation. The general principle of operation used in all forms of on-load tap changer is that, momentarily at least, a connection is made simultaneously to two adjacent taps on the transformer during the transition period from one tap to the next. Impedance in the form of either resistance of inductive reactance is introduced to limit the circulating current between the two tappings. The circulating current would represent a short-circuit between taps if not so limited. *Figure 33.24* shows in diagrammatic form the use of a centre-tap inductor or auto-transformer as the transition impedance. In *Figure 33.24(a)* the load current is shown passing from the maximum tap through the halves of the inductor in opposition, and hence, non-inductively. In *Figure 33.24(b)* one of the two tap-selector switch contacts has opened and the load current is carried inductively through one half of the inductor. In *Figure 33.24(c)* the inductor is shown bridging the two adjacent tappings. The load current is shared equally between the two tappings

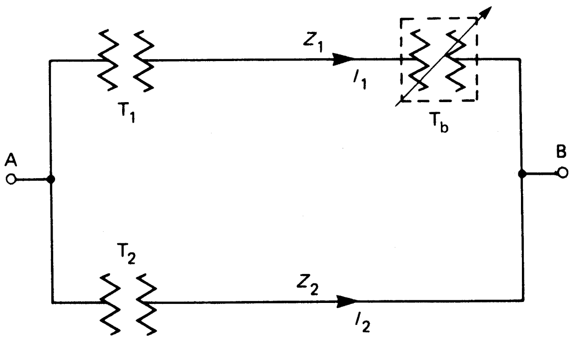


Figure 33.22 Power flow along two parallel lines of different impedance

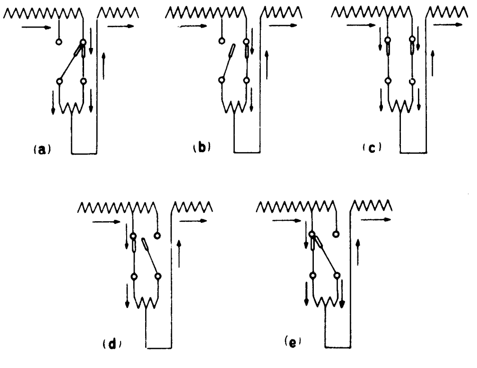


Figure 33.24 Inductor transition

and passes non-inductively in opposition through the halves of the inductor. The tap step voltage is applied to the whole of the winding of the inductor and the circulating current is limited by the total impedance. In this position, in which the tap changer can remain indefinitely, the effective voltage is equivalent to the mean of the two individual tap voltages. *Figure 33.24(d)* shows the momentary condition where one half is connected inductively to the incoming tap position, and *Figure 33.24(e)* shows the final stage of the transition where both selector switch contacts are connected to the incoming tap and the inductor is non-inductively connected.

Circulating-current limitation by centre-tapped inductor was common in the late 1940s, but has since been almost entirely superseded by high-speed resistor transition. The switching sequence is shown in *Figure 33.25* and is similar to that with inductors, except that two resistors are used. Backup main contacts are provided which short-circuit the resistors for normal running conditions.

Advantages of inductor transition were: the inductor could be continuously rated and a failure of auxiliary supply during a tap change did not necessitate the main transformer being taken out of service; also, that the intermediate or bridging position could be used as a running position, giving a voltage equivalent of one-half tap step. The main disadvantage was that the circulating current between taps during the bridging condition was at low power factor, adversely affecting diverter-switch contact life. The inductor itself was costly and occupied a significant amount of space in the transformer tank.

Resistor transition is now used almost exclusively by British and European tap changer manufacturers, although inductor transition is still used in the USA, possibly because it is common practice there to specify a large number of small tap steps, a requirement met by using the bridging position as a running tap.

Resistor transition requires one winding tap for each operating position. The basic disadvantage is that the resistors cannot be continuously rated, if their physical size is to be kept small. It is essential to minimise the period during which they are in circuit. Some form of energy storage has, therefore, to be incorporated in the driving mechanism to ensure that a tap change, once initiated, is completed irrespective of failure of auxiliary supply. Early resistor tap changers operated at low speed and the stored-energy mechanism was a flywheel or a falling mass. All modern tap changers use springs for energy storage, and the total time that a resistor is in circuit during a tap change is limited to a few periods. The advantages of the high-speed resistor tap changer are its compactness and lack of wear of diverter-switch contacts because of the high speed of break, and because the circulating current is at unity power factor. Contact life of 250 000 operations is common, compared with the 10 000–20 000 for reactor tap changers.

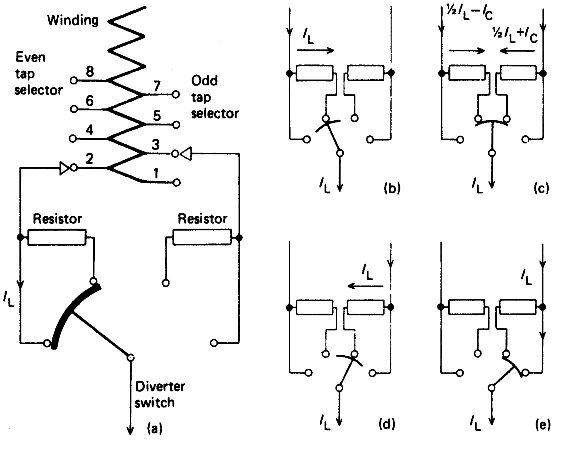


Figure 33.25 Resistor transition. (a) Outgoing tap operating; (b) load current in resistor; (c) resistors in parallel to load current and in series to circulating current; (d) load current in resistor; (e) incoming tap operating

Irrespective of the form of transition, all on-load tap changers fall into one of three categories in respect of the switching arrangement. These are as follows:

(1) The oldest (and now least common) arrangement is for a separate contactor to be connected to each winding tap. Contactors are operated by a camshaft to ensure the correct sequencing. A later development was to use mercury switches instead of open-type contactors, giving the advantage of freedom from carbonisation of the coil.
(2) The winding tappings are connected to a series of fixed contacts of a selector switch of either linear or circular form, and an associated pair of moving contacts operates to provide the required switching sequence. Current making and breaking occurs at the selector switch contacts and some degree of oil carbonisation and contact wear is inevitable. This type of tap changer, usually called 'single-compartment', is now common for transformers up to about 20 MVA rating.
(3) For the largest and most important transformers the tap selector switches do not move when carrying current. Current making and breaking is carried out by two separate diverter switches, usually in a separate compartment of the tap changer to minimise the amount of oil contaminated by carbon. The diverter switches operate to make and break the current and are mechanically interlocked with the selector switches, which move only when not carrying current to provide the correct sequence of connection to the winding taps.

At one stage of development of medium-sized tap changers, large mercury switches were used as diverter switches; these could be mounted in a selector switch compartment as there was no risk of oil contamination. A recent development along the same lines is the application of vacuum-switch diverters, capable of many thousand operations without attention and with freedom from contamination. Trials are being made with thyristors to provide 'contactless' switching: while this might yield marginal advantage in minimising maintenance and perhaps improving reliability, the cost and complexity of the control arrangements are likely to be inhibiting factors.

33.7.1 Tap changer control

Control gear can vary from simple local push-button control to a complex scheme for the automatic control of as many as four transformers operating in parallel. The objective of automatic tap change control is to maintain output voltage either constant or with a compound characteristic rising with load. The main component is the automatic voltage-regulating device, which consists of a voltage governor, a time-delay relay and compounding elements. The time-delay element prevents tap changes occurring due to minor short-time fluctuations of voltage. It can be set for delay periods of up to a minute. Tap change control circuits necessarily involve auxiliary switches mounted within the driving mechanism of the tap changers themselves, and this has led to a proliferation of types of control schemes each designed to operate with a particular type of tap changer. These variations have made it extremely difficult for the British Electricity Supply Authorities to develop a national standard control scheme. Some of the differences arise in the method adopted to arrange simultaneous or near-simultaneous tap changes of transformers operating in parallel. A further complication arises because of differences between one transformer and another in the tapping range and the number and size of tapping steps. Some of the various types of parallel control schemes are as follows:

(1) Simultaneous operation of two or more tap changers initiated from one voltage regulating relay. Closing of the 'raise' or 'lower' contact in the voltage-regulating relay closes the appropriate motor contactor in each tap changer, and auxiliary switches lock out further movement until all the tap changers have completed a single change of position. This scheme is, in general, only applicable to tap changers with identical main and auxiliary characteristics.

(2) Master–follower operation initiated from one voltage-regulating relay. This is also suitable for two to four transformers in parallel, although the length of time to complete an initial tap change may be considered to be excessive where more than two transformers are involved. The voltage relay initiates the movement of the 'master' tap changer: when this has completed one tap step, the auxiliary switches operate to cause the second transformer of the chain to come into line with the first. This is followed in sequence by the movement of the remaining units. The time for completing one tap step on all units in the bank is, therefore, the sum of the individual operating times. As in (1), this scheme is suitable only for substantially identical tap changers.

(3) Master–follower scheme with multiple voltage-regulating relays; a more complicated version of (2) in that each transformer is equipped with its own voltage-regulating relay and any one of the group may be selected to act as master, the remaining units following in any desired sequence.

(4) Circulating-current control schemes. The foregoing schemes all require multipoint switches in each tap changer, interconnected by multi-core cable. The circulating-current schemes depend on the fact that if two transformers operating in parallel are out of step, a circulating current will flow between them in a direction depending on the relative ratio of transformation. Each transformer is controlled by its own voltage-regulating relay, and as individual characteristics are not absolutely identical, it follows that, when a change of voltage occurs, one relay of the group will initiate a tap change on its associated transformer earlier than the others. As soon as the first transformer of a group has completed a tap change, there will be an imbalance of ratio and a circulating current will flow in the main circuit connections between the transformers. This circulating current is used to control auxiliary relays in each tap changer of the group, so that no further movement can occur to increase the imbalance, i.e. the second and later transformers in the group can change step to come into line with the first unit which has already moved; alternatively, the first unit can move in the reverse direction to bring itself back into its original position and therefore directly in step with the others. This type of scheme allows transformers to operate indefinitely in a 'one step out' condition. The wiring is relatively simple, as the only interconnections between units are the secondary leads of the circulating-current transformers. It permits automatic parallel control of transformers of different rating, impedance, tapping range and number and size of tapping steps, as the tap changers take up positions to minimise the circulating current between units. Provided that the c.t.s. are selected to correspond to the rating of each main transformer, optimum loading of the group is achieved.

(5) Parallel control by reverse compounding. All of the schemes mentioned above necessitate secondary connections between the transformers that are to operate in parallel, and all except (4) suit only transformers with near-identical characteristics. Where parallel operation is required between transformers with differing characteristics, or where the transformers are situated some distance apart, it is possible to achieve stable operation with each being controlled by its own voltage-regulating relay by introducing negative compounding of the reactance element of a line-drop compensator. This tends to give a negative compounding characteristic (i.e. output voltage drops as load increases), but compensation can be provided by increasing the positive compounding of the in-phase element. Unless the negative reactance characteristic is introduced, any two tap changers controlled by independent voltage-regulating relays which are not positively locked together will inevitably move quickly to opposite extremes of their range.

33.7.2 Line-drop compensation

This device permits the output voltage of the transformer to be compounded so that it rises with load to compensate for voltage drop in the cables connected to the secondary side, and to maintain approximately constant voltage at a remote point. The line-drop compensator comprises an 'artificial line' circuit consisting of adjustable resistances and reactances connected into the voltage relay operating-coil circuit (*Figure 33.26*). A current transformer in the main secondary connection injects current into the adjustable resistance and reactance coils and thus biases the voltage applied to the voltage sensitive element of the regulating relay.

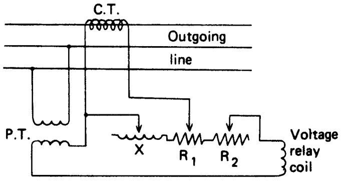


Figure 33.26 Line-drop compensation. R_1 and X are adjustable to suit line characteristics; R_2 adjusts the output-voltage level

33.8 Cooling

Small transformers are air cooled and insulated. For units of larger rating and higher voltage, oil cooling becomes economical, because oil provides greater insulation strength than air for a given clearance, and augments the rate of removal of heat from the windings. With the exception of certain special installations, such as in coal mines or within occupied buildings where mineral oil is undesirable because of the fire risk involved, almost all power transformers are oil immersed. The combination of oil and paper insulation has been used in transformers for many years and there appears little likelihood of its being superseded by any modern synthetic material. A principal reason for this situation is that both materials can be operated safely at the same maximum temperature, approximately 105°C. Any alternative materials would have to show significant advantages in either or both insulating and heat transfer properties compared with the combination of oil and paper.

The most common types of cooling arrangements, detailed below, are identified in IEC 60076 by a system of symbols which indicate the cooling medium in contact with the windings; the cooling medium in contact with the external cooling system; and the kind of circulation for each. The symbols for the cooling media are:

O for mineral oil or an insulating liquid with fire point <ς300°C
K for an insulating liquid with fire point >ς300°C
L for an insulating liquid with no measurable fire point
W for water
A for air
G for gas.

The symbols for circulation in a liquid immersed transformer are:

N for natural thermosyphon flow
F for forced fluid circulation, but thermosyphon cooling in the windings
D for forced fluid circulation, with fluid directed into the windings.

IEC specifications stipulate temperature limits for windings (measured by resistance) and insulation and define normal standard values for the temperature of the cooling medium.

33.8.1 Air insulated, air cooled

33.8.1.1 Natural air cooling (AN)

The temperature rise measured by resistance is limited by the class of insulation used. Typical figures are 60 K for class A, 90 K for class B and 150 K for class C materials; all rises being above a maximum ambient temperature of 40°C and a daily average of not more than 30°C. Type AN cooling is generally limited to relatively small units, although the development of high-temperature insulation, such as glass and silicon resins, has resulted in its use on transformers up to 1500 kVA and for special application as in mines.

33.8.1.2 Forced air cooling (AF)

The temperature conditions are the same as for AN, but the improved heat transfer properties resulting from the forced air stream enables the current density in the windings and the flux density in the core to be increased and greater output to be obtained from a given size of unit.

33.8.2 Oil immersed, air cooled

IEC 60076 recognises two maximum oil temperatures: 60°C when the transformer is sealed or equipped with a conservator, and 55°C when the transformer is not so equipped. Winding temperature rise by resistance for oil immersed transformers with ambient air temperatures of 40°C and a daily average of not more than 30°C is limited to 65 K, irrespective of the type of cooling or the cooling medium. The various types of cooling and the new symbols are as follows.

33.8.2.1 Natural oil circulation, natural air flow (ONAN)

The great majority of power transformers up to ratings of 5 MVA are of ONAN type. A plain sheet-steel tank radiates about 13 W/m^2 per degree Celsius rise. Above about 25 kVA, three-phase, an increased cooling surface becomes necessary. This extra surface may be obtained by using fins or corrugations, but the most common method is to employ a tubed tank. The tubes are usually 40–50 mm in diameter, of welded steel construction, having a wall thickness of about 1.5 mm. For medium sizes (2–5 MVA) tubes of elliptical section are preferred, as a greater number can be accommodated on a given tank. For transformers larger than 5 MVA it becomes necessary to employ radiator banks of elliptical tubes, or banks of corrugated radiators.

Transformer tanks have been constructed with finned tubes in order to augment the surface. They are difficult to paint, and are liable to collect water if employed out of doors.

The power dissipated by a tubular tank is a function of the ratio between tank envelope and total surface, for radiation is a function of the envelope, while convection depends upon the whole surface.

33.8.2.2 Natural oil circulation, air blast (ONAF)

By directing an air blast on to an ordinary tubular tank or on to separate radiators the rate of heat dissipation is increased; thus, while the transformer itself is not reduced in size, less external cooling surface is required.

33.8.2.3 Forced oil circulation, natural air flow (OFAN)

OFAN is an uncommon system, but is useful where for reasons of space the coolers have to be well removed from the transformer. The oil is pumped round the cooling system, from which heat dissipation is by natural air convection. The forced oil circulation permits high current densities to be employed in the windings, so that there is a reduction in transformer size.

33.8.2.4 Forced oil circulation, air blast (OFAF)

The OFAF system is employed for most large transformers. The forced oil enables the windings and core to be economically rated, while the forced air blast reduces the size of the radiating surfaces, an important point for transformers of 30 MVA upward. Depending upon the type and disposition of the radiators and on the purchaser's requirements, *mixed* cooling is employed in which the transformer operates as an ONAN unit up to 50% of its forced cooled rating (66% in the USA). As the load increases further, temperature sensitive elements start the pumps and fans of the forced cooling equipment.

33.8.3 Oil immersed, water cooled

Cooling-water inlet temperature limits as defined in IEC 60076 must not exceed 25°C. Oil temperature and winding temperature limits are as for air cooled units.

33.8.3.1 Natural oil, water (internal cooler) (ONWF)

A copper cooling coil is mounted above the transformer core in the upper portion of the transformer tank.

33.8.3.2 Forced oil, water (external cooler) (OFWF)

The OFWF system uses oil/water heat exchangers external to the transformers. The arrangement has a number of advantages over ONWF cooling:

1. The transformer is smaller in size, as the windings can be more highly rated owing to forced oil flow and the tank is not required to accommodate a large cooling coil.
2. Condensation troubles are non-existent.
3. Water leakage into oil is improbable with the oil pressure maintained higher than that of the water. In cases where the cooling water has a high head at the transformer plinth level, it is necessary to employ two heat exchangers in series, i.e. oil/water and water/water; the water in the intermediate circuit between the two coolers is separate from the main water supply and at a low head.
4. The cooling tubes may be easily cleaned. Water cooling of transformers is common at generating stations (particularly hydroelectric ones), where ample cooling water supplies are available.

33.8.4 Overload capability

The temperature limits for oil, windings and insulation laid down in IEC 60076 are chosen to ensure that a transformer operating within these limits will have a satisfactory life of 20 or more years. The relationship between operating temperature and life is complex. Experimental work indicates that each 8°C increase in operating temperature halves the life of the insulation, but there is little information available to enable an operator to determine precisely the actual life expectation under any given operating conditions. This obviously depends on other factors, including the incidence and severity of short-circuit forces to which the ageing insulation may be subjected. To some extent also the loss of insulation life due to overload operation during an emergency can be offset, provided that for some further and more protracted period operating temperatures are kept well below the specified limits.

Because of the considerable thermal inertia of the mass of oil and metal in a transformer, appreciable overloads can be carried for short periods without endangering insulation life. IEC 60076-7 gives details of permitted overloads.[10] These allowable overloads are, however, normally regarded as conservative and have been exceeded in service without adverse results. The overload recommendations in IEC 60076-7 are based on a maximum hot-spot temperature of 90°C at rated load and loss of life doubling for each 6°C increase in temperature above 98°C. It is important that the windings are not operated at a temperature above 140°C as air/water vapour bubbles are formed from cellulose insulation at higher temperatures. Bubble formation can lead to dielectric failure of the windings.

33.9 Fittings

The number of fittings and their siting on the transformer tank constitute major problems of transformer standardisation. It is recommended, therefore, that the fittings enumerated in IEC 60076 be used, careful consideration being given to the essential points before any refinements and extras required for satisfactory operation are specified.

Terminals and bushings Transformer terminals must of necessity cover a wide range of voltages, currents and operating conditions. Outdoor type bushings are standardised in Britain and are detailed in ESI Standards 35-1 and 35-2 and BEB Specifications T_1, T_2 and T_3. Some outdoor bushings are fitted with arcing horns which provide a safety gap to discharge an incoming surge which might otherwise damage the transformer windings. It is important that there is the correct correlation between the insulation strength of the transformer winding and the flashover characteristics of the gap. The setting of the gap must be small enough to protect the windings without causing too frequent interruption of the supply due to power arcs following gap flashover due to surges on the system.

Cable boxes These are essential when paper or plastics insulated cables require connection to the transformer. Cable box arrangements depend on the number of cores in the cable and the number of cables connected in parallel. Several factors must be considered in designing a cable box: adequate electrical and mechanical clearances are necessary; the compound must not ooze out of the box nor any transformer oil leak in; and voids must not easily occur in the compound, either during pouring operations or during normal thermal expansion and contraction. In order to make transformers interchangeable the box flanges are standardised (ESI 35-1 and 35-2).

Oil conservator Transformer oil has a coefficient of thermal expansion of 0.000 725 per degree Celsius at 0°C, equivalent to a 7.25% volume change over an oil temperature range of 0–100°C. The expansion may be accommodated in a free-breathing or sealed space at the top of the tank, or in a conservator tank mounted on the tank cover. It is desirable to avoid oxidation of the oil, which causes sludging and acidity to develop. With either the sealed tank or the conservator a nitrogen cushion can be maintained above the surface of the oil, thus preventing oxidation. A free-breathing conservator is preferable to an unsealed tank, as the temperature of the oil in contact with the air is lower and this in itself reduces the rate of oxidation. Some form of oil pressurisation is essential for large transformers and for those working at 33 kV and above; it is desirable also for small transformers, especially those subjected to heavy periodic peak load.

Breather This device allows ingress and egress of air to compensate for changes in oil volume. Except on small transformers, the breather should incorporate either a chemical or a refrigerating system for removing moisture from the air entering the transformer.

Oil gauge This can be a direct-reading glass window, or a dial instrument operated by magnetic coupling from a float on the oil surface. The dial gauge can be fitted with contacts to give a low-oil-level alarm.

Oil temperature indicator Normally of the dial type, oil thermometers can be arranged for remote electrical indication of oil temperature.

Winding temperature indicator The so-called winding temperature indicator is basically an oil temperature thermometer in which the bulb is associated with a heater coil which carries a current proportional to the load on the transformer.

The heater coil introduces an increment of temperature above that of the oil, to correspond to the gradient between winding and oil temperatures. The instrument thus indicates a figure which, although a reasonable analogue of the temperature of the windings, is not a direct measurement thereof.

Important new developments in direct measurement of temperature in high-tension windings are being introduced.[11] This measurement is achieved by the use of fibre-optic strands attached to one of the winding conductors. Two different applications are possible. The first uses a gallium-arsenide based sensor tip at the end of the fibre-optic strand to provide a single-point temperature detector.

The second application is to use the whole length of the fibre-optic strand itself as a distributed sensor. A technique known as optical time domain reflectometry interrogation from a single end enables temperature measurements to be made at points 1 m apart along the full length of a fibre-optic strand up to 4 km long, i.e. at up to 4000 data collection points. These applications may lead to considerable advances in determining safe overloading rules.

Buchholz relay Any electrical fault occurring inside a transformer is accompanied by an evolution of gas. Appreciable quantities of gas may be produced before the fault develops to such an extent that it can be detected by the normal protection equipment. The Buchholz relay, connected between the transformer tank and the conservator, contains two elements: (1) a float which operates a gas alarm device to give warning of gas discharge from within the transformer, and (2) a surge element connected to trip the transformer out of circuit in the event of a massive surge of oil and gas resulting from a major internal fault. In the event of a gas alarm being given, it is sometimes possible to deduce the nature of the defect within the transformer by observing whether the gas emission is at a constant rate (voltage dependent) or varies with load (current dependent).

The Buchholz relay provides sensitive protection against certain types of fault (e.g. flashover between tapping leads) to which the normal protection is relatively insensitive, and may not operate until extensive damage has been caused.

Relief or explosion vents Many users specify relief or explosion vents, which are intended to act as safety valves to reduce internal pressure in the event of a major fault within the transformer and thus to protect the tank from damage. The vents can be spring loaded or fitted with thin non-metallic material to fracture under pressure. Provided that the build-up of pressure is relatively slow, the relief vent can operate satisfactorily and prevent the tank from bursting, but in the event of a violent fault leading to a shock wave of pressure, the tank may be burst before the relief vent has time to operate.

Tapping switches (off-load) Almost every distribution and medium-sized power transformer is fitted with voltage adjusting tappings, usually for five positions corresponding to ±2.5% and ±5% of the supply voltage. To obviate opening up the transformer to change tapping links, a tapping switch is necessary. Such a switch is often fitted on top of the transformer core and is gang operated on all phases by means of a hand wheel on the tank end.

33.10 Parallel operation

The following information is required when the parallel operation of transformers is considered:

1. output and temperature rise of the transformers;
2. polarity for one-phase units; angular displacement for three-phase units-IEC group reference or phasor diagram;
3. turns ratio on all tappings;
4. percentage impedance at 75°C; and
5. percentage resistance drop at 75°C, or the load loss.

The polarity is basically determined by the direction of the primary and secondary windings and the position of the transformer line leads with respect to the start and finish of the windings. British transformers usually have a subtractive polarity. In the case of three-phase units it is the angular phase displacement, i.e. the angle between the primary and secondary phase to neutral voltages, which has to be considered. This angle may be 0°, 180° or ±30° depending on the direction of the windings and the interphase connections.

All groups of transformers having the same angular displacement may be connected in parallel. Those having +30° displacement may be paralleled with −30° units, provided that the line leads are suitably transposed. Parallel operation is not possible between a 0° or 180° group and a ±30° group.

Some typical examples of connections are:

Connection	*IEC ref. no.*	*Displacement* (°)
Delta/star	Dy.11	+30
Star/delta	Yd.1	−30
Star/star	Yy.0	0
Star/inter-star	Yz.11	+30
Delta/inter-star	Dz0	0

From this it will be seen that parallel operation is not possible between star/star and delta/star units. If the turns ratios of the transformers are not identical, a circulating current traverses the transformer windings, increasing the no-load losses. The magnitude of this circulating current for the case of two transformers A and B is

$$I_s = (V_A - V_B)/(Z_A + Z_B)$$

where V is the no-load secondary voltage for a common primary voltage and Z is the leakage impedance, all quantities being complexors.

The relative values of the percentage (or per-unit) impedance determine the proportion of the total load shared by each transformer. Thus, when all the percentage impedances are identical, each transformer will take its fair share of the load. Although it is advisable to have this condition, dissimilar impedance units may be connected in parallel in an emergency, provided that the current carried by any transformer does not exceed its normal rating or an acceptable overload value.

The percentage resistance drops of the transformers need not be the same. A difference in resistance drop, when the percentage impedances are numerically equal, results in an angular displacement of the individual transformer currents and reduces slightly the maximum permissible output. This is not normally of serious consequence.

33.11 Auto-transformers

The great advantage of the autoconnection, as distinct from the usual double-winding arrangement, is that the transformer physical size and losses are much smaller, provided that the primary to secondary turns ratio is not large. The amount of apparent reduction is termed the autofraction

$$n = (V_1 - V_2)/V_1$$

where V_1 is the higher and V_2 is the lower voltage. Thus, the equivalent frame rating of an auto is equal to n times the load or throughput rating. For a 2/1 ratio this means that the auto-transformer is half the size of a double-wound transformer for the same duty. A requirement for tappings can have a marked effect on the apparent economy of using an auto-transformer.

Unfortunately this economy is not obtained without certain liabilities, so that care is required in specifying auto-transformers unless the conditions are known and appreciated.

The calculated reactance of an auto-transformer on a frame kilovolt-ampere base has to be multiplied by n to obtain the reactance on a throughput base, e.g. a 2/1 ratio auto-transformer with a frame reactance of 4% would present an impedance of $4/2 = 2\%$ to through faults. Given a high fault MVA in-feed on the system, this could lead to short-circuit currents of more than the maximum permitted value of 25 times the normal one. System operating conditions must be clearly specified and, if necessary, additional impedance introduced to limit fault currents. It is the joint responsibility of purchaser and manufacturer to ensure that the transformer will not be subjected to excessive stresses.

The common electrical connection between the primary and secondary sides is a potential source of danger. The position of the earth connection with a three-phase star-connected auto-transformer is important and it is normally preferable to connect the supply neutral (assumed earthed) to the auto-neutral, and not to have the auto-neutral floating.

The use of the auto-arrangement on transformers interconnecting different voltage levels (e.g. 400/275 kV) of a transmission system enables significant cost savings to be achieved. Small units are most useful as voltage regulating devices: they lend themselves readily to the provision of tappings, and as the loads generally have constant impedance characteristics, a small unit can control a large load. Consider a 10 kW, 400 V heating load, taking 25 A in an equivalent resistance of 16 Ω. It is required to control the heat in five steps by adjustment of the secondary load voltage E_s, using an auto-transformer. The secondary current for each tap is $I_s = E_s/16$, and the corresponding primary current is $I_p = I_s(E_s/400)$. The winding currents are I_p, in the part corresponding to $(400 - E_s$, and $(I_s - I_p)$ in the remainder. The winding could be graded to carry the maximum current in each portion, and it should be noticed that the 100 V tapping currents are less than those already determined for any portion of the winding. It is an axiom that, for constant impedance auto-transformers, any tappings below half the supply voltage do not influence the transformer size. The equivalent kVA is the sum of the part winding kVA values, divided by 2, and for the example this is 1.65 kVA.

33.11.1 Auto-starters

The principles above are applied to auto-starter transformers for three-phase induction motors. The value of the equivalent motor impedance (assumed constant) is:

$$Z = 1.25V/\sqrt{3}kI = 0.72V/kI$$

where V is the supply voltage, I is the full-load current of the motor and k is its ratio between short-circuit current and full-load current. The winding currents are determined according to the number of tappings. These currents are of short duration.

33.12 Special types

33.12.1 Static balancer

The static balancer is a simple apparatus which in its three-phase form comprises an ordinary three-phase transformer core carrying two windings per limb in zig-zag connection. Normally, when connected to a three-phase line the balancer draws only a small magnetising current. When a load is connected between one line and neutral, however, so that the current balance of the feeders is upset, the load current flows through the balancer windings. This condition is illustrated in *Figure 33.27*, the current distribution being based upon a 100 A line-to-neutral load. Each balancer winding carries one-third of the neutral or out-of-balance current, and has one-third of the line voltage impressed across it. The rating of the balancer as a three-phase transformer is therefore $2(\sqrt{3}V/3)(I/3)(3/2) = 0.58\ VI$, where V is the line-to-neutral voltage, I is the neutral current and VI is the one-phase load being balanced.

(1) To supply a one-phase load. A balancer will be found to be cheaper than a one-phase transformer connected across two lines, and much cheaper than a three-to-one phase transformer. Overload protection is provided by a fuse in the load circuit as shown in *Figure 33.27(a)*, and the balancer neutral current corresponds to the load current.
(2) To improve the voltage regulation of four-wire networks. Balancers have in the past been used chiefly for this application. The improved regulation has been useful on rural distribution systems with isolated one-phase loads. The balancer rating depends upon the out-of-balance current of the system, which usually is taken as 20% of the three-phase line current.
(3) To transform a three-wire system into a four-wire one. Fuses, or even a circuit breaker, must not be employed in the balancer line. If one fuse were to blow, the neutral current would have to flow through the high-impedance paths offered by the sound balancer limbs and there would be a rise in the line-to-neutral voltage on the sound phases.

33.12.2 Welding transformers

Owing to their simplicity, economy and efficiency, transformer welding sets predominate over their rotary machine d.c. counterpart. Stick electrodes have been specially developed

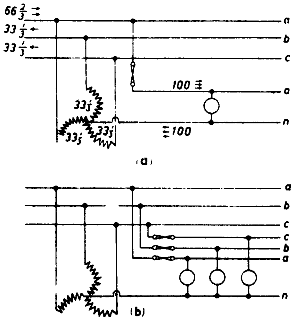


Figure 33.27 A.c. static balancer applications

for a wide range of applications for use on a.c. sources, but even where a d.c. source is imperative, preference is shown for transformer/rectifier equipment over rotary machines.

The basic requirement for a.c. welding is a low-voltage power source (70–100 V), with an adjustable series inductor to ensure stability of welding current and provide phase shift between the source voltage and the welding current, enabling the arc to be re-struck in each half-period after the current has passed through zero.

Power sources are supplied for use either by individual operators when part of the series inductive reactance may be incorporated in the transformer, or by groups of operators when a single multiphase transformer of relatively low impedance provides low-voltage (90 V) supply through a number of separate adjustable inductors. In the latter case advantage can be taken of the diversity factor in minimising the power rating of the transformer. For general-purpose applications standard a.c. single-operator welding sets are also available with inbuilt rectifiers and a smoothing inductor enabling the operator to use a wider range of electrodes.

The inherently high reactance of the source produces a very low p.f. load on the supply which, owing to its variability and intermittent nature, cannot be continuously corrected by capacitors.

However, some correction is both possible and desirable, and most single-operator sets are designed to house a capacitor of a size recommended by the supplier. The performance of welding power sources, specified in BSEN 60638, covers the basic requirements for the majority of applications, but for other uses such as consumable electrode shielded gas systems, the characteristics of the welding set play a fundamental part in determining the quality of the weld. Typically, in transformer/rectifier power sources the internal impedance of the sets is very low and the open-circuit voltage is little more than the arc voltage; this produces a high rate of change of welding current when the arc length varies, and automatically adjusts the burn-off rate. In the past most welding sets supplied for the British market have been oil cooled, and these are most suitable for the onerous conditions found, for instance, in shipyards; but the development of new insulating materials has enabled air cooled dry-type sets increasingly to take their place in less onerous conditions and where light weight and portability are valuable.

33.12.3 Mining transformers

The operating conditions in coal mines impose special requirements on transformers for use underground. There must be no possibility of a defect in the equipment itself causing an explosion of the gaseous atmosphere in the mine, and head-room is normally extremely limited.

Until the early 1950s, special low-height oil filled non-flameproof mining transformers were used underground up to within 300 m of the coal face. The switchgear directly mounted on these transformers was of certified flameproof construction. As the size of the load increased, the problems of voltage drop in the low-voltage cables between the transformer and coal face machinery became greater, leading to the need for a completely flameproof transformer (and associated switchgear) which could be taken close to the coal face. The modern flameproof underground transformer is an air insulated unit constructed with class C insulation and contained in a flameproof case.

33.12.4 Small transformers

Small transformers are made in large numbers for electronic apparatus and similar equipment. The open construction has been superseded by a hermetically sealed arrangement in air filled or oil filled metal containers. The connections are brought out through metal/glass or metal/ceramic seals soldered to the container. Transformers for mobile equipment, which may be operated at 400–1600 Hz for the sake of the saving in weight and size, are made of relatively costly materials to obtain larger magnetic and electric loadings. Thus, very thin cold-rolled silicon steel or thin nickel-iron cores may be used with coils insulated by high-temperature dielectrics such as glass fibre, silicones, etc. In this way a 30 VA, 1.6 kHz, fully sealed transformer weighs about 0.1 kg compared with its open-type equivalent manufactured in 1940 weighing 1 kg or more.

33.13 Testing

The normal practice for testing power and distribution transformers is to carry out a comprehensive set of tests at the maker's works—the number and nature of which depends on whether the transformer is the first of a new design or otherwise—and a few relatively simple tests after installation at site to prove that the transformer is ready for service. The three classes of works tests are referred to as 'type' and 'routine' and 'special'. The first transformer of a particular design or contract is subjected to both type and routine tests, while routine tests only are applied to later units. Special tests are only required at the specific request of the purchaser.

33.13.1 Routine tests

Routine tests consist of:

(1) voltage ratio, polarity and phase displacement checked;
(2) winding resistance measured;
(3) insulation resistance measured;
(4) load loss and short-circuit impedance measured;
(5) no-load loss and magnetising current measured;
(6) dielectric routine tests; and
(7) tests on on-load tap changers, where appropriate.

33.13.1.1 Ratio, polarity and phase displacement

Figure 33.28 shows the type of circuit used for measuring ratio. This involves the use of a 'ratiometer', which basically consists of a multi-ratio transformer from which tappings are taken to coarse and fine adjusting switches. The ratiometer and the transformer under test are connected in opposition. When the ratiometer is adjusted to give a ratio exactly equal to that of the transformer under test, no current will flow in the secondary circuit. The ratio can then be read directly from the dial readings on the ratiometer. Polarity and interphase connections are checked by measuring voltages between various terminals when the transformer is energised at a low voltage.

33.13.1.2 Winding resistance

The d.c. resistance of each phase of each winding is measured separately by the voltammeter method and is recorded together with the temperature of the winding at the time. This information is required for use in connection with later measurements of the load loss and the temperature rise of the transformer under rated load.

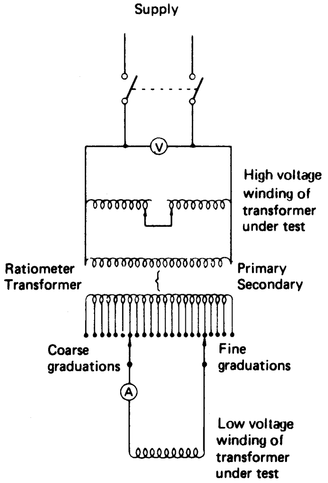

Supply
V
High voltage winding of transformer under test
Ratiometer Transformer
Primary
Secondary
Coarse graduations
Fine graduations
A
Low voltage winding of transformer under test

Figure 33.28 Ratio test

Because of the inductive effect of the core, care must be taken to ensure that a steady d.c. value is reached before voltage and current readings are recorded.

33.13.1.3 Insulation resistance

The insulation resistance between windings and from each winding to earth is measured by a special instrument such as the Megger or Metrohm.

The insulation resistance is commonly used as one of the criteria for determining that the transformer has been properly dried out. It varies widely and inversely with temperature, and care is necessary to ensure that the readings are correctly interpreted.

33.13.1.4 Load loss and impedance

Load loss and impedance are measured by short-circuiting the terminals of one winding of the transformer and applying a low voltage to the other winding sufficient to cause rated full-load current to flow. Because the applied voltage, and, hence, the magnetisation of the core, is extremely low, the core loss can reasonably be neglected and the measured input power represents the total load loss at rated load on the complete transformer. The short-circuit is applied to the low-voltage winding and the supply is connected to the high-voltage winding, at which side all readings are taken. In principle, the same result would be obtained if the high-voltage winding were short circuited and the supply connected to the low-voltage winding, but this involves measuring the heavier low-voltage rated currents, which may be too large for convenience.

In the past, a 2-wattmeter method was allowed to measure load loss of smaller transformers, but this method is no longer accepted and three wattmeters are necessary to make the measurement. Alternatively, an Electronic Power Analyser may be used to measure all quantities on a single instrument with higher accuracy.

It is important that the test is performed at normal frequency to ensure the correct proportioning of the PR and stray losses in the windings and structure of the transformer, which are dependent on frequency.

In transformers with exceptionally high reactance the voltage that has to be applied to circulate rated full-load current through the windings, even under short-circuit, may be sufficient to magnetise the core to a level at which the loss therein may be significant. In such cases the core loss during the short-circuit test may be determined by removing the short-circuit and measuring magnetising loss when the transformer is excited on open circuit, at the previously measured voltage required to circulate full-load current. The true load loss is then the difference between the two successive measurements.

It is important that the short-circuiting connections are substantial and applied in such a manner that the loss therein does not represent a significant fraction of the loss within the transformer. It is also important that the temperature is measured at the time that the load loss measurements are made, so that the necessary correction can be made to deduce the copper loss at the temperature (75°C) at which guarantees normally apply. IEC 60076 details the manner in which the load loss measurements at any given temperature can be corrected to the equivalent figure at 75°C.

33.13.1.5 No-load loss and magnetising current

The fundamental principle of this test is that normal rated voltage is applied to one winding while the other is left open circuit. The current flowing in the winding to which the supply is connected is the magnetising current and this is recorded as part of the test records. This magnetising current is normally a small percentage of the full-load current and the I^2R loss is negligible compared with the core loss. To avoid unnecessarily high voltages in the test circuit during the core loss test, it is normal practice to connect the supply to the lower voltage winding of the transformer (see Section 33.2.3).

33.13.1.6 Dielectric routine tests

Dielectric routine tests consist of an applied-high-potential (separate source test), a short duration a.c. test or a long duration a.c. test in combination with a swithing impulse test, dependent on the voltage rating, and a lightning impulse test, dependent on the voltage rating.

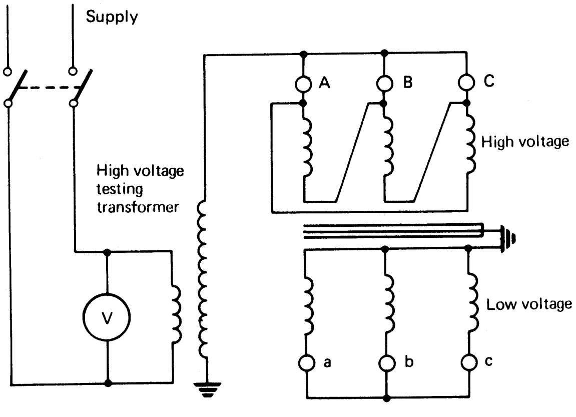


Figure 33.29 Connections for applied-high-potential test

Applied-high-potential tests These tests are normally made, in turn, between each winding, and the core and all other windings connected to earth.

Figure 33.29 shows the arrangement with the high-voltage winding under test and the low-voltage winding and core connected together and to earth. The magnitude of the applied potential test depends on the rated voltage of the winding in question and on whether the major insulation between it and earth is uniform or graded. For a uniformly insulated winding the applied voltage test provides the principal dielectric test of the main insulation. It is usually of the order of $(2E+1)$ kV, where E is the 'highest system voltage' for the winding in question. Full details are given in IEC 60076. On a three-phase transformer an applied voltage test of $2E$ raises the line terminals to 3.46 times their normal operating voltage to earth.

For graded insulation windings the applied voltage test is at a value appropriate to the insulation level at the neutral point and therefore does not adequately prove the strength of the line-end insulation.

Short-duration a.c. test This induced voltage test is a short duration a.c. test that involves exciting the transformer on open circuit at a voltage higher than normal for a short period. This test is a routine test on transformers of up to 170 kV rating and a special test for transformers of higher voltage. For transformers with uniform insulation on which the applied high-potential test provides the principal check on the strength of the major insulation, the purpose of the induced test is to prove the strength of the insulation between turns and between other parts of the transformer operating at different potentials. The magnitude of the test is usually twice rated voltage, and to prevent overexcitation of the core the frequency of supply also has to be increased to at least twice normal.

On transformers up to 170 kV rating, with graded insulation the induced overvoltage test constitutes the main test of the major insulation. The magnitude of the test is fixed so that the potential to earth of each of the high-voltage terminals in turn is raised to the appropriate test voltage for the system on which the transformer is to operate. The magnitude of the test may be as high as 3.46 times normal rated voltage, and the interturn and other insulation is obviously tested to this degree at the same time.

The duration of the test is 60 s for any test frequency up to and including twice rated frequency. When the test frequency exceeds twice rated frequency, the duration of the test is for 6000 periods (i.e. 1 min at 100 Hz) or a minimum of 15 s, whichever is the greater. The magnitude of the test voltages for different system operating voltages and conditions are given in IEC 60076.

Long-duration a.c. test For transformers rated above 170 kV, the routine test is a combination of a long duration a.c. test, lasting 60 m and a switching impulse test.

The standard wave shape for a switching impulse test on air insulated equipment is of the order of 250/2500 μs, i.e. a relatively slow rise of front followed by a tail of long duration. The practical difficulty in producing such a wave from an impulse generator connected to a transformer winding (related to the generator capacity and the transformer-core saturation) has led to a relaxation of the requirements for the switching surge wave specified for application to transformer windings. Limiting features are a wavefront rise time of at least 90 μs, a time above 90% of the specified amplitude of at least 200 μs, and a total duration from virtual origin to first zero passage of at least 500 μS.

Partial discharge measurements are made during the long duration a.c. test to give a reliable check on the capability of the insulation in normal service.

Details of the tests and of the specified test levels are given in IEC 60076-3.

Lightning impulse test Although this is a type test for transformers rated up to 72.5 kV, the lightning impulse test is a routine test for all transformers of higher voltage.

The lightning impulse test simulates the conditions that exist in service when a transformer is subjected to an incoming high-voltage surge due to lightning or other disturbances on the associated transmission line.

Lightning impulse tests were introduced originally solely as type tests and, because of fears that they might cause undetected damage to the insulation of a transformer, the tests were largely confined to specially built prototype assemblies and were not applied to production units prior to going into service. Gradually it was realised that with increasing sensitivity of the equipment provided to detect insulation breakdown during the test, there was little risk of a service transformer suffering undetected damage. Still later it was appreciated that the sensitivity of the failure detection equipment was such that it would disclose hitherto unsuspected damage that might have occurred during a preceding power-frequency over-potential test. In Britain, therefore, common practice is for the lightning impulse test to follow the short-duration a.c. test, or the combined long-duration a.c. test and switching surge test.

Details of the lightning impulse test levels are specified in IEC 60076 for transformers for various system voltages. The precise form of a complete impulse test varies with customers' preferences and on whether the test is being applied as a basic type test on the first unit of a new design, or as a routine check on insulation strength following power-frequency test as described above.

The normal sequence for impulse tests in Britain is:

(1) one reduced-level full-wave (voltage between 50% and 75% of the full-wave voltage test level);
(2) one full-wave at specified test level;
(3) two chopped waves with a crest value not less than the specified full-wave test level; and
(4) one full-wave at the test level.

Evidence of insulation failure during an impulse test primarily depends on oscillograph records of the impulse voltage wave and either of the current passing through the winding under test or of the voltage induced by inductive transfer in another winding of the phase under test.

The normal sweep time for the oscillograph recording the voltage wave is 100 μs for a full-wave and a shorter time for a chopped-wave test, e.g. 10 μs. Current oscillograms are taken with the same sweep time, or preferably simultaneous records are taken of current with three different sweep times, e.g. 10, 100 and 500 μs.

It is becoming difficult or impossible to purchase new analogue impulse recording instruments and many laboratories have now installed digital recording equipment. It is important that digital recording equipment used when testing power transformers has at least a 10-bit analogue-to-digital converter in the circuit.

The current and voltage wave shapes as shown on the oscillographic records are carefully compared with each other and particularly with the traces taken during the preliminary reduced voltage shots. Any discrepancy, however slight, in the wave shape of any one of the records compared with the others indicates a change in the conditions either in the test equipment or in the transformer under test. Such discrepancies must be investigated and satisfactorily explained, and any failure to do so involves the risk of a transformer with damaged insulation being put into service. IEC 60722 states that in cases of any doubt as to the interpretation of discrepancies three subsequent 100% full-wave shots shall be applied, and that if the discrepancies are not enlarged by these tests, the impulse test is deemed to have been withstood.[12]

Before the development of the sensitive failure detection techniques now available, the principal criteria for determining whether an impulse test had been successfully withstood were that there should be no noise from within the transformer during the test, and no sign of smoke or bubbles in the oil. These remain as accepted criteria, but are now of secondary significance compared with the oscillographic procedures. The chopped wave test is not usually included when impulse tests form part of the routine tests on a transformer, because of the time involved in setting up the chopping gap and some doubt as to whether the chopped wave test is useful for the detection of defects of workmanship or material. The main purpose of the chopped wave test as part of the impulse-type test series is to increase the stress in the insulation within and between the coils adjacent to the transformer line terminal. This is normally regarded as a feature of the design rather than of the individual unit. The chopped wave test is, nevertheless, representative of the conditions that arise when an incoming surge is suddenly chopped because of a flashover of an insulator in or near the substation, and it is prudent to ensure that the transformer is capable of safely withstanding events of this nature.

33.13.1.8 Tests on on-load tap changers

With the tap changer fully assembled on the transformer the following test sequences are performed:

1. With the transformer un-energised, eight complete cycles of operation are performed.
2. With the transformer un-energised, and with the auxiliary voltage reduced to 85% of its rated value, one complete cycle of operation is performed.
3. With the transformer energised at rated voltage and frequency one complete cycle of operation is performed.
4. With one winding short-circuited, and rated current in the tapping winding, 10 tap change operations over two tap steps on either side of the middle tapping, or where any reversing switch operates, is performed.

33.13.2 Type tests

Type tests consist of:

1. dielectric type tests;
2. a temperature-rise test.

Until the 1950s the distinction between type and routine tests was clearly defined, but operational experience, particularly on very large e.h.v. transformers, seems to be leading towards a compromise arrangement being adopted on transformers other than on the first of a new design. In particular, a simplified impulse test is frequently specified for application to all transformers purchased by certain customers. The logic is that many weaknesses disclosed by impulse test have been found to be due to poor materials or workmanship and not to fundamental errors of design. In the light of this experience it is obviously prudent to adopt the highly sensitive impulse test, even in modified form, as a routine check on materials and workmanship.

33.13.2.1 Dielectric type tests

The lightning impulse test is a type test for transformers rated below 72.5 kV.

The long-duration a.c. test is a special test for transformers rated above 72.5 kV but below 170 kV; it is however, often requested as a type test on new designs.

33.13.2.2 Temperature test

Each new design of transformer should be subjected to a test to determine that the temperature rise at rated load will not exceed the guaranteed values. It is uneconomic, if not totally impossible, to test a large transformer at the maker's works with both full voltage applied and full-load current in the windings, as the total output would have to be supplied and dissipated in some way. On small transformers where the rating of the available test plant is equal to or greater than twice the rating of the transformer under test, it is possible to arrange two units connected back-to-back in the manner shown in *Figure 33.30*. If the ratio of both units is identical, no current will flow in the windings, but if the ratio is deliberately unbalanced (by connecting the two units on different tappings), a circulating current will flow through the two units, of magnitude governed by the out-of-balance voltage and the total impedance of the two units in series. The current is largely reactive (since the reactance of a transformer is normally considerably greater than the resistance), and the net power taken from a supply is equal to the total power loss in the two transformers under test. If tappings are not available, or are unsuitable to circulate approximately rated full-load current, it is possible to inject the requisite circulating current through a small

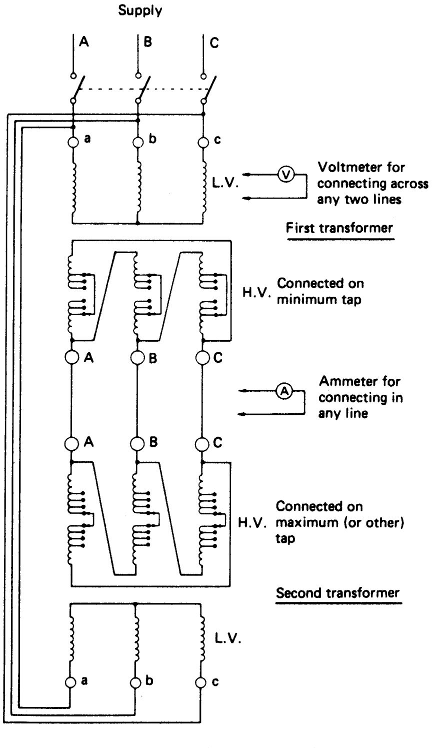

Supply
A
B
C
a
b
c
L.V.
V
Voltmeter for connecting across any two lines
First transformer
H.V.
Connected on minimum tap
A
B
C
A
Ammeter for connecting in any line
A
B
C
H.V.
Connected on maximum (or other) tap
Second transformer
L.V.
a
b
c

Figure 33.30 Method of connection for a back-to-back heat run

booster transformer connected in the leads running between the two main transformers.

The back-to-back connection, or direct loading, must be employed when making temperature tests on dry-type transformers, because the heat transfers from the core and windings to the cooling medium (air) are largely independent. The test conditions must therefore represent actual conditions in service when the core is heated by the magnetisation loss and the windings are heated by the load current. In oil filled transformers the heat generated in both core and windings is transferred to the oil, and the total heat then has to be dissipated from the oil to the cooling medium. Because the loss in the core is usually appreciably less than the loss in the windings, it is possible to employ the so-called short-circuit method of test.

Under short circuit the loss in the core is almost negligible and the current in the windings is adjusted to a value slightly above the rated figure, so that the total loss in the windings is equal to the sum total of the separately measured load and core losses. The procedure during a short-circuit temperature test is to load the transformer with this increased current until such time as the observed oil temperature rise becomes sensibly constant, or is not rising at more than 1°C/h. This part of the test proves that the cooling equipment of the transformer is adequate to dissipate the total losses under normal full-load conditions and the oil temperature rise is recorded accordingly.

The increased current obviously results in the temperature gradient between windings and oil being higher than when the current is at the rated value. The thermal inertia of the windings is relatively low, however, and any change in current is quickly followed by a corresponding change in the temperature gradient between windings and oil. After the oil temperature rise has been recorded, current is reduced from the increased to the normal rated value and maintained at this level for 1 h. The supply is then disconnected and the d.c. resistance of the windings measured in a manner similar to that employed prior to the loss measurement tests when the transformer was cold. By taking temperature measurements over a period of 10 m plotting a graph against time from shut-down and extrapolating back to the time of shut-down, the resistance of the windings at the instant of shut-down can be determined. The winding temperature rise is then calculated by comparing the cold and hot resistances, with an allowance being made for any fall in oil temperature during the last half-hour at rated current. Full details of the methods are given in IEC 60076.[12]

33.13.3 Special tests

Special tests consist of:

(1) Dielectric special tests.
(2) Determination of capacitances between windings and earth and between windings.
(3) Determination of transient voltage transfer characteristics.
(4) Measurement of zero-sequence impedances on three phase transformers.
(5) Short circuit withstand test.
(6) Determination of sound levels.
(7) Measurement of harmonics in the no-load current.
(8) Measurement of the power taken by the fan and oil pump motors.
(9) Measurement of insulation resistances to earth of the windings, and measurement of the loss angle (tan δ) of the insulation system capacitances.

Special tests are only carried out by agreement between purchaser and manufacturer. The most common special test carried out in Britain is the sound level test.

33.13.3.1 Sound level measurement

Because of the importance of noise as an environmental factor, the specification of a sound level limit for power transformers is increasingly common. The British Electricity Board's Specifications for various sizes of transformer stipulate maximum acceptable sound levels for each size of transformer purchased. The sound level tests, normally carried out at the manufacturer's works, are usually made at times when factory noise (including that of running test plant) can be kept to a level well below that of the noise emitted by a transformer on test. When it is necessary to make sound level tests in a more noisy environment than the noise emitted by the transformer it is necessary to use the sound intensity measurement method. It is also important that the transformer on test is well clear of walls or other large areas which would reflect sound and cause a build-up of noise which would give a false reading.

Sound level measurements made using sound pressure, sound power or sound intensity methods have bee standardised and are described in IEC 60076-10.

33.13.4 Commissioning tests at site

Commissioning tests vary considerably with the size and importance of the installation.

For a small or medium-sized distribution transformer the minimum requirements would be a visual examination for transport damage and an insulation test with a portable instrument. Preferably there should be a check of the ratio (by applying a medium voltage to the high-voltage terminals and measuring the induced voltage at the low-voltage terminals), and on the oil level and condition to confirm that ingress of moisture has not occurred. Measurements of ratio and of polarity are essential if a transformer is to be connected into a circuit where it will operate in parallel with other transformers.

On large units which are normally despatched either without oil or are only partially filled, checks must be made on the filling procedure and of the condition of the oil prior to filling, in addition to ensuring that the insulation has not become wet during transport.

After filling, a sample of oil may be taken for dissolved gas analysis (see Section 33.14.4) so as to provide a basis for comparison with similar samples taken as part of a routine maintenance procedure when the transformer is in service.

Auxiliary equipment such as on-load tap changing gear and any protective relays and current transformers associated with the main transformer must also be checked for correct operation.

In general, a repetition of high-voltage tests carried out at the works is not considered to be necessary. Where the transformer is subjected to retesting on site at high voltage, the test voltage level is normally restricted to 75% of that applied during tests at the works.

33.14 Maintenance

Maintenance can be described as the measures adopted to ensure that equipment is kept in a fully serviceable and reliable condition. Of necessity it is therefore mainly a routine involving attention at regular intervals to particular features based on service experience and manufacturers' recommendations. In the former, the measures involved tend to be general, whereas the latter tend to cover, in addition to general points, particular measures depending on the constructional characteristics of the individual manufacturer's designs. Consideration here is limited to two particular issues, namely insulating oil and solid insulation.

33.14.1 Insulating oil

Oil forms part of the main insulation of most transformers, but it tends to deteriorate in service owing to (a) operating temperature, (b) atmospheric conditions (applicable to unsealed, non-conservator-type transformers) and (c) presence of moisture or fibres.

In (a) deterioration is accelerated by prolonged high operating temperatures leading to the development of acidity and sludging, which in turn have a deleterious effect on the solid insulation. Poor ventilation in a transformer chamber (b) results in condensation inside the transformer, which similarly is liable to promote acidity and sludging. The electrical strength of the oil is considerably reduced by included moisture or fibres and particularly by a combination of both (c).

Deterioration can be greatly reduced or even arrested by attention to operating conditions and by routine precautions. Samples of the oil should, therefore, be taken from the transformer at regular intervals and the characteristics checked. With large transformers, generally speaking, little trouble is experienced with acidity, mainly owing to the lower operating temperatures. Standard British practice is to specify conservator-type transformers, and apart from one experimental installation associated with a group of generator transformers, it has not been necessary to consider the use of 'inhibited' oil. Similarly, any form of 'sealing' has in the past been considered unnecessary, for the same reasons. More sophisticated oil preservation equipment is now being specified for 400 kV transformers.

In small transformers there is a greater tendency for the development of acidity, but with modern oil (and provided that reasonable precautions are taken) no serious inconvenience should be experienced in this respect. Discharge under oil may result in the flash point of the oil being reduced, although after a relatively short period of time it may recover. Similar reduction in flash point can occur as the result of abnormal local heating such as may be experienced during the development of an incipient fault, e.g. a core fault: involving circulating currents within the core itself due to a breakdown of interlamination resistance or failure of core bolt insulation.

33.14.2 Insulation

The standard method for checking the state of insulation is by measurement of the insulation resistance. It should be noted, however, that a transformer with relatively 'wet' insulation may have a high insulation resistance when the measurement is made with the transformer cold, but the value may drop rapidly as maximum operating temperatures are approached. A hot insulation resistance reading below 1 MΩ per 1000 V rating of the tested windings is generally indicative that drying out is necessary.

When suitable equipment is available, measurement of dielectric loss angle gives a reliable check on the state of the internal solid insulation. As the value of the loss angle will depend on the transformer design, a reference value taken on the actual transformer during works test is necessary for comparison before any useful assessment can be made. A useful method mainly used in the factory for checking the state of dryness in solid insulation is the 'dispersion' test or 'recovery voltage' test, based on the fact that in dry insulation the distribution of elements is such that individual shunt paths of time constant greater than 3 ms are effectively absent and that the presence of moisture introduces time constant within the range 3–300 ms. The method of test involves the application of a 300 ms pulse followed by a 3 ms short-circuiting pulse. Any shunt paths having time constants greater than 3 ms retain their charge, which is measured as a voltage by a suitable measuring device, the measurement indicating the moisture content.

The windings of a transformer should be inspected at long-term intervals. Any slackness due to insulation shrinkage or to the falling out of packing can then be remedied either by the adjustable coil clamping screws or by packing out the winding. This operation is particularly advisable in the case of transformers subject to heavy load surges, such as furnace transformers. Large transformers are usually built with preshrunk windings, so that slackening in service is almost entirely eliminated.

33.14.3 On-load tap changing equipment

From an operational aspect, an important factor is the period of time that can be allowed to elapse before attention to the switch contacts becomes necessary. If practicable, this period should be such that it can be co-ordinated with other outages of plant. This is of particular importance with generator transformers, where outage of the unit means the non-availability of generating plant. Until comparatively recently it was normal practice to carry out maintenance after every 10 000 operations, reasonably corresponding to a normal 12-month period between generator overhauls. The modern high-speed resistor tap changer, however, requires diverter switch maintenance only after 100 000 or more operations, and tap changer maintenance is no longer a limiting factor from the operational point of view.

33.14.4 Reliability and condition monitoring in service

Ignoring short-term outages due to defects in components, the reliability of a transformer depends on the electric strength of the insulation being designed and maintained at an adequate level to withstand the stresses imposed on it by either steady-state or transient voltages.

Although rare, unexpected conditions can arise to cause insulation failure despite an apparently adequate margin of safety. Wagenaar *et al.*, refer to such incidents where discontinuities in the winding arrangement of a large generator step-up transformer responded to critical frequencies arising on the system.[7] Resonant voltages within the windings—particularly in the tapping zone—were significantly higher than those arising during impulse testing and insulation breakdown occurred. The supply authority has increased test voltage levels and introduced further tests to detect critical frequency response within the windings and to prove capability to withstand the consequent stresses.

In the long term, insulation failure can be caused by ageing. This may be due to overheating because of local cooling deficiency or overloading, or simply because of long periods of operation at or near full load. The economics of supply system operation dictate that maximum utilisation is made of installed transformer capacity, by normally loading well up to rated value and accepting some loss of life due to deliberate overloading in an emergency. In these circumstances it is essential to be able to monitor the condition of important units at regular intervals to allow action to be taken to prevent a failure in service.

The simplest and cheapest diagnostic technique available for on-line tests is by laboratory analysis of dissolved gases (DGA) in samples of oil taken from the transformer while in service. DGA can be used to diagnose the type of fault causing gas to be produced, e.g. arcing, intense local heating of conductor joints or other metal parts, with or without involvement of cellulose insulation.

IEC 60599 gives guidance for fault diagnosis which involves determining the concentration of the various hydrocarbon gases and calculating the ratio of the concentration of different pairs of gases.[14]

Table 33.3 provides a useful summary of the key gases produced by different types of fault.

A very important development has been to determine the amount of furfuraldehyde present in the oil samples.[15] Furfuraldehyde is the liquid residue of the breakdown of paper by molecular chain scission and is the only non-destructive indication of the thermal ageing of paper. It enables an assessment to be made of the 'amount of life' left in the insulation before total ageing has occurred and can be used to determine whether a transformer should be replaced.

33.15 Surge protection

Any transformer connected to an overhead transmission line must be protected against surges resulting from lightning striking the line conductors. National and international standards exist for the insulation level of lines, switchgear and transformers at all usual transmission voltages, in the context of *insulation coordination*.

Transformers are protected against lightning surges by discharge gaps which may be in the form of simple arcing horns attached to the transformer bushings, or more sophisticated surge arresters. The minimum spacing of the gap electrodes, is chosen to ensure that a flashover will not occur under normal steady state or transient power-frequency operating conditions. The voltage at which the gap will flashover following a lightning surge is known, and the insulation level of the transformer is chosen to withstand this gap voltage together with a margin of safety. (The nominal margin is 20%, but in practice it is somewhat higher, owing to the fact that the actual strength of the transformer insulation is necessarily greater than its specified level.)

Modern surge arresters are based on metal oxide technology. Under normal service conditions the arrester draws only a very small current, but under surge or overvoltage conditions the non-linear characteristics of the material allow it to draw a much larger current to supply a discharge path to the surge or overvoltage.

Whether surge arresters are used instead of the simple discharge gap depends on such factors as the severity of lightning in the area and the importance of ensuring that the individual supply circuit remains intact. A lightning surge that causes a gap discharge with either form of protection is inevitably followed by a power arc sustained by normal operating voltage. With surge arresters the non-linear characteristic causes the resistance to rise immediately the lightning discharge current (which may amount to several kiloamperes) ceases to flow. With simple arcing horns, however, a flashover of the gap due to lightning is followed by a power arc which is not self-extinguishing and which forms a direct earth fault on the phase in question. This leads to disconnection by the normal earth fault protection, although if the associated switchgear is arranged for automatic reclosing, there is only a momentary interruption.

In Britain it has been found that, in general, surge arresters are unnecessary largely because of the relative freedom from severe lightning storms and the small statistical probability of a lightning strike close to a transformer. Open-type gaps (usually referred to as 'co-ordinating gaps') give perfectly adequate protection against lightning surges with relatively slow 'fronts' (of the order of 5 μs or greater), although necessarily with the disadvantage described above in respect of continuity of supply following a gap flashover. In other parts of the world where lightning storms are intense, it is normal practice to provide surge arresters because of the better protection that they provide.

Lightning arresters are used in Britain on certain lower-voltage lines (as an alternative to the use of automatic reclosing circuit-breakers) and on some 132 kV wood pole lines without earth wires. Such lines have an inherently high insulation level to ground (compared with a steel tower line, in which a lightning surge is usually immediately discharged by a flashover of a line insulator on the nearest tower) which results in high-amplitude surges travelling along the line to the transformer.

Surge arresters are also used in certain special cases (e.g. in association with shunt inductors) where circuit characteristics are such that it is desirable to limit the magnitude of switching surges by installing an arrester to provide a discharge path. Conversely, some circuit configurations (e.g. the complex bus-bar system of a main switching station) are such as to reduce to an insignificant level any likelihood of a co-ordinating gap flashover being caused by lightning. In such cases surge arresters may be found unnecessary, even on such important installations as main generator step-up transformers.

Table 33.3 Sources of 'key gases' from decomposition of cellulose and oil

Material	*Condition and temperature*	*Key gases*	*Symbol*
Cellulose	Overheated >150°C	Carbon monoxide Carbon dioxide (Water)	CO CO_2
Cellulose	Excessive heat >1000°C	Carbon monoxide Carbon dioxide (Carbon/tar)	CO CO_2
Oil	Overheated >150°C	Methane Ethane Ethylene (Organic acids)	CH_4 C_2H_6 C_2H_4
Oil	Electrical stress (partial discharges and arcing to 1000 °C)	Hydrogen Acetylene (Waxes and water)	H_2 C_2H_2

33.16 Purchasing specifications

The essential basic information to be given is the following:

(1) standard (national or international);
(2) rated winding voltages;
(3) frequency;
(4) cooling medium (external);
(5) cooling medium (internal);
(6) ambient temperature of cooling media;
(7) transport limitations (weight, dimensions);
(8) nominal rating;
(9) number of phases;
(10) phase connection;
(11) phase relation;
(12) terminal arrangement;
(13) altitude (if >1000 m);
(14) impulse levels for lightning and switching surges (if applicable);
(15) impedance (with acceptable minimum or maximum values for limitation of fault level or voltage regulation);
(16) characteristics of transformers with which parallel operation is required;
(17) tapping range (number and size of steps, whether on-load or off-load, restriction of impedance variation over tap range);
(18) performance requirements for cyclic or emergency overloading; and
(19) special requirements (fittings, paint finish, etc.).

In addition to the technical information, listed above, the manufacturer should be advised on the basis of purchase (minimum first cost, maximum efficiency or capitalised loss). If a loss-capitalising formula is to be used, details should be provided to minimise the work of preparing and evaluating bids.

A working group within CIGRE Study Committee 12 has prepared a guide for customers' specifications for transformers 100 MVA and 123 kV and above.[16]

Specifications should set down a clear definition of technical requirements and operating conditions, and exclude detailed clauses on constructional points better left to the maker. The highly competitive transformer market, together with the normal practice of buying on the lowest tender that complies with the specification, means that the manufacturer must put forward the minimum-cost design with no extra capability not specified. It is thus important to specify every abnormal requirement.

References

1 KARSAI, K., KERENYI, D. and KISS, L., *Large Power Transformers*, Elsevier, Amsterdam (1987)
2 British Standards Institution, BSEN 60076 Power transformers
3 IEC 60064-1: Magnetic materials classification
4 IEC 60064-8-7: Specifications for individual materials — Cold-rolled grain-oriented electrical sheet steel and strip delivered in the fully processed state
5 IEC 60076-10: Power transformers—Determination of sound levels
6 IEC 6 60076-3: Power transformers—Insulation levels, dielectric tests and external clearances in air
7 WAGENAAR, L. B., SCHNEIDER, I. M., PROVANZANA, J. H., YANNUCHI, D. A. and KENNEDY, W. N., 'Rationale and implementation of a new 765 kV generator step-up transformer specification', *CIGRE*, Paper 12-202, Paris (1990)
8 MALEWSKI, R., DOUVILLE, I., LAVALLEE, L. and TSCHUDI, D., 'Dielectric stress in 735 kV generator transformers under operating and test conditions', *CIGRE*, Paper 12-203, Paris (1990)
9 IEC 60076-5: Power transformers—Ability to withstand short circuit
10 IEC 60354: Loading guide for oil-immersed transformers
11 WHITE, A., DANIELS, M. R., BIBBY, G. and FISHER, S., 'Thermal assessment of transformers', *CIGRE*, Paper 12-105, Paris (1990)
12 IEC 60722: Guide to the lightning impulse and switching impulse testing of transformers and reactors
13 IEC 60076-2: Power transformers—Temperature rise
14 IEC 60599: Mineral oil-impregnated electrical equipment in service—Guide to the interpretation of dissolved and free gases analysis
15 CARBBALLEIRA, M., 'HPLC contribution to transformer survey during service or heat run tests', *Electra* pp. 45–51, No. 133
16 'Guide for customers specifications for transformers 100 MVA and 123 kV and above', *CIGRE*, Brochure No. 156, Paris (2000)

34

Switchgear

B M Pryor DipEE, CEng, FIEE
Power Systems Consultant Substations & Main Plant

Contents

34.1 Circuit-switching devices 34/3
    34.1.1 Disconnectors (isolators) 34/3
    34.1.2 Switches 34/3
    34.1.3 Switch disconnectors 34/4
    34.1.4 Earth switches 34/4
    34.1.5 Fuses 34/5
    34.1.6 Fuse switches 34/10
    34.1.7 Contactors 34/12
    34.1.8 Circuit-breakers 34/13

34.2 Materials 34/19
    34.2.1 Insulating materials 34/19
    34.2.2 Contact materials 34/21

34.3 Primary-circuit-protection devices 34/21
    34.3.1 Current transformers 34/22
    34.3.2 Voltage transformers 34/22
    34.3.3 Combined-instrument transformers 34/24
    34.3.4 Surge arresters 34/24

34.4 LV switchgear 34/25
    34.4.1 Fuse cut-outs 34/25
    34.4.2 LV fuse cabinets 34/26
    34.4.3 LV switchboards 34/27

34.5 HV secondary distribution switchgear 34/27
    34.5.1 Urban networks 34/27
    34.5.2 Rural networks 34/29

34.6 HV primary distribution switchgear 34/29

34.7 HV transmission switchgear 34/29

34.8 Generator switchgear 34/30

34.9 Switching conditions 34/31
    34.9.1 Normal switching conditions 34/31
    34.9.2 Abnormal switching conditions 34/33

34.10 Switchgear testing 34/34

34.11 Diagnostic monitoring 34/35

34.12 Electromagnetic compatibility 34/35

34.13 Future developments 34/35

'Switchgear' is a term used to refer to combinations of switching devices and their interconnection with associated control, protective and measurement systems. It facilitates the interconnection of different parts of an electricity supply network by means of cable or overhead-line connections in order to allow the control of the flow of electricity within that network. Switchgear is also designed to be capable of safely clearing any fault which may occur on any part of the electricity network in order to protect the network itself, connected equipment and operational personnel. It also provides the facility, by means of disconnectors, for segregating parts of the network and, with the application of earth switches, allows safe access for maintenance or repair to component parts of the electricity supply network. Switchgear connected into an electricity supply network via transformers, overhead lines and/or cables, together with its mounting structures, housings, protective fencing and ancillary equipment and connections is referred to as a 'substation'.

These terms generally apply to switchgear of all voltage ratings ranging from, for example, supplies to domestic consumers at say 240 V with nominal short-circuit currents of up to 16 kA, to equipment used on major transmission networks at say 420 kV or higher with short-circuit currents of typically 55 kA. Switchgear may also be used for direct connection of main generators having rated short-circuit currents of typically 200 kA at 24 kV with normal rated load currents of 24 kA.

Switchgear is used to provide supplies to major items of plant such as motors, arc furnaces, reactors, capacitors or other machinery as well as supplies to commercial, industrial and domestic consumers. In doing so it must also provide for the safe operation of connected items of plant and limit the effects on connected equipment of transient overvoltage phenomena caused, for example, by lightning or other surges generated by operation of the switching devices themselves. In order to perform these varied functions, switching devices used within switchgear assemblies have evolved in different ways to suit particular applications. They may be mounted separately in an outdoor fenced area and be interconnected by means of air-insulated bus-bars, this arrangement generally being referred to as an 'open-type substation'. Alternatively, they may be coupled together by the switchgear manufacturer to form an assembly within a metal enclosure. The insulating media may comprise combinations of air, solid insulating material, insulating oil, bitumen compound or, with more recent equipment, an inert dielectric gas known as sulphur hexafluoride (SF_6).

The major components of the switchgear are described in the following sections.

34.1 Circuit-switching devices

34.1.1 Disconnectors (isolators)

A disconnector is a device comprising movable contacts capable of being mechanically closed to form an electrical circuit to other equipment or of being mechanically opened to physically disconnect one part of the electrical network from an adjacent part and at the same time providing an isolating distance. This provides circuit segregation for operational purposes or, with appropriate earthing equipment and application of safety procedures, provides facilities to allow work to be safely undertaken on disconnected and earthed primary electrical equipment.

Disconnectors are generally hand operated but may have power operating mechanisms if remote operational facilities are required. Disconnector moving contact systems are generally slow in operation as a disconnector is not intended to interrupt circuit currents. It will, however, be required to interrupt the small capacitive current flowing in the associated disconnected equipment. When air-insulated disconnectors are operated a short duration 'buzz' is generally heard when this capacitive current is interrupted. High-frequency currents are generated during such an operation. These can result in primary circuit overvoltages, which in turn can be coupled into earthed circuits or secondary connected equipment resulting in electromagnetic interference. In its closed position a disconnector must be capable of carrying the full rated load current of the associated circuit and of safely withstanding the rated short-circuit current of the system for the maximum circuit-breaker clearance time which may be between 1 and 3 s.

Disconnection may be achieved in metal-enclosed switchgear of voltage ratings typically up to 36 kV by physical withdrawal of the circuit-breaker either horizontally or vertically. No separate disconnector as such is provided, as the separation of the plug in contacts of the circuit-breaker provides an isolating distance. Earthed metallic shutters are commonly provided to segregate both sides of the circuit, but this is not an essential requirement for a disconnector. For higher voltage metal-enclosed equipment or for open-type substations separate disconnectors are used. Depending on the substation electrical layout, disconnectors are generally provided to isolate either side of each circuit-breaker and all circuits connecting into a substation. In addition, they are generally used on overhead rural distribution networks to segregate parts of the network in the event of a faulted section.

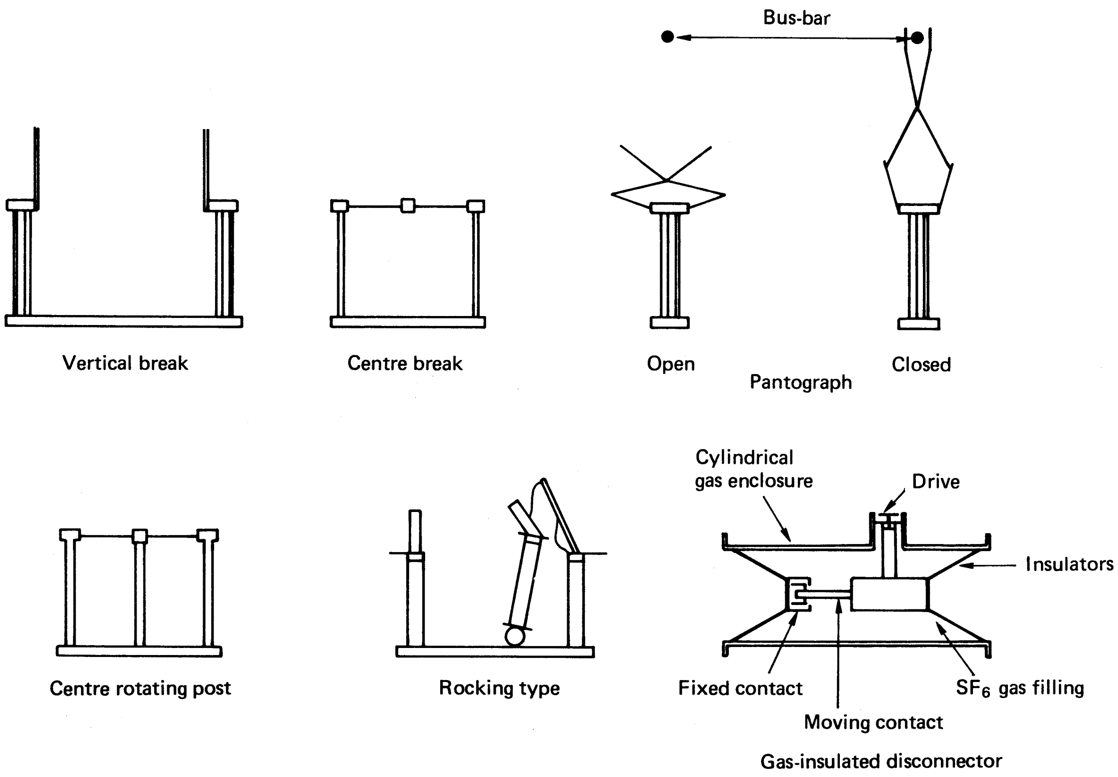
Disconnectors can take a number of different physical forms depending on particular applications and substation layout. Typical arrangements are shown in *Figure 34.1*.

The isolating distance of a disconnector is generally visible, but this is not an essential requirement. With open-type equipment, visible indication is readily achieved, but with metal-enclosed equipments windows may be provided. These may induce additional problems. Very many disconnectors are in use which do not provide visual indication of the isolating gap and with this type of equipment the operator must be able to rely fully on the external operation indicators accurately representing the positions of the main contacts.

34.1.2 Switches

A switch is a mechanical switching device that, unlike a disconnector, is capable of closing against, and interrupting, circuit load currents. In its closed position it must be capable of carrying the rated short-circuit current of the system for the rated time, i.e. 1–3 s. An additional optional design feature is the ability to close satisfactorily against the rated short-circuit current of the system. A switch is not capable of breaking short-circuit currents. A further optional design feature is the ability to close and open satisfactorily against specified overload conditions as may occur, for example, when switching motor circuits.

In view of the need to make and break specified circuit conditions, consistency of operation is an essential requirement. As direct manual operation cannot achieve this, power operation is necessary. Power operation of the switch is invariably achieved by the use of a spring operating mechanism whereby the spring is compressed during the



Vertical break

Centre break

Open

Pantograph

Closed

Centre rotating post

Rocking type

Gas-insulated disconnector

Figure 34.1 Typical disconnector arrangements

early part of the operation by manual means and then travels over centre to release its stored energy to operate the moving contact system independently of the operator. Switches normally are either air or oil insulated, with more recent designs being SF_6 insulated and they usually have some rudimentary method of controlling the arc during a breaking operation. Although the major application of switches is in low-voltage networks they are also widely used on distribution networks at voltages of up to about 20 kV. Use at higher voltages is minimal and at transmission voltages very few current designs exist and these have very limited specific applications.

34.1.3 Switch disconnectors

A switch disconnector is a switch which in its open position provides the isolating facilities required of a disconnector. Interlocking and padlocking facilities are generally provided. In LV applications switch disconnectors are not always employed, whereas at higher voltages switch disconnectors are almost invariably employed. The oil switch and SF_6 switch shown in *Figure 34.2* normally provide an isolating distance and are thus defined as switch disconnectors.

For open-type equipment at distribution voltages a switch may be built into a disconnector. This usually comprises an air-insulated interrupter as shown in *Figure 34.3.* Such a switch disconnector provides the facility for switching a downstream length of overhead line without the need to de-energise the whole of the line as would be required for a disconnector.

Switch disconnectors were also commonly used at transmission voltages some years ago but they now show little economic advantage over the use of a circuit-breaker and their application has declined. A typical 420 kV switch disconnector is shown in *Figure 34.4.*

34.1.4 Earth switches

An earth switch is not a switch as described in Section 34.1.2 because it does not have to make or break load current. An earth switch is a mechanical switching device used to connect the disconnected and de-energised primary conductors of a circuit to earth and to allow work to be undertaken on the earthed part of the circuit. An earth switch does not have to be capable of making or breaking current on its contacts and, consequently, earth switches are usually dependent manual operated. In its closed position it must be capable of carrying the rated short-circuit current of the system for the rated time. This requirement is necessary to safeguard against the circuit being inadvertently re-energised at its remote end. Earth switches are usually interlocked by mechanical or electrical means with the associated disconnector such that they can only be operated to apply an earth to the system when that part of the system has been de-energised and isolated. However, when applying an earth to an outgoing circuit comprising a cable or overhead line, interlocking with the remote end of the circuit is not feasible and safety instructions and permits must be relied upon. These may not always be foolproof and some authorities insist that line-end and cable-end earth switches must, in addition, be capable of safely closing onto an energised circuit. Such earth switches are generally referred to as ‘fault-making earth switches’ and power-operated mechanisms are an essential requirement.

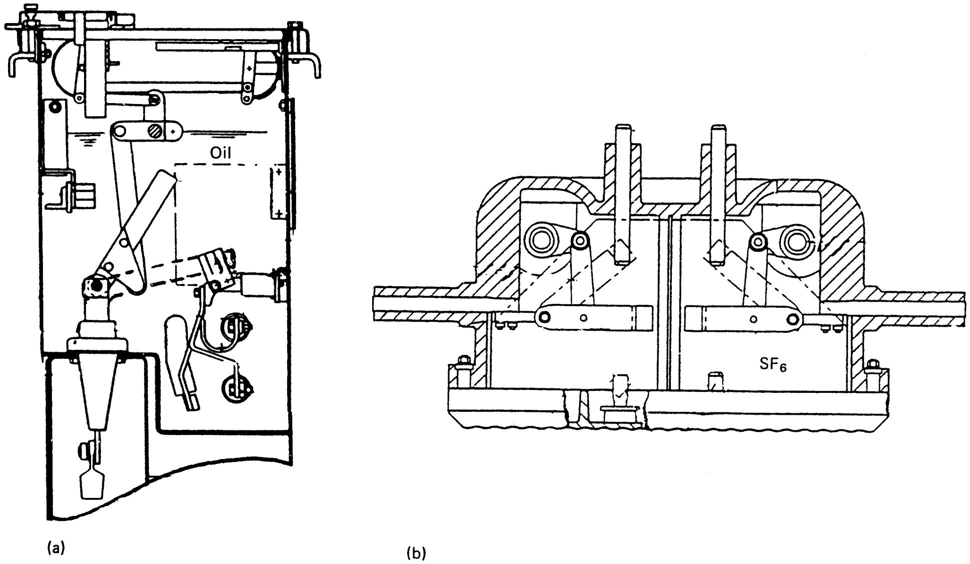


Figure 34.2 Arrangement of (a) an oil switch and (b) an SF_6 switch disconnectors

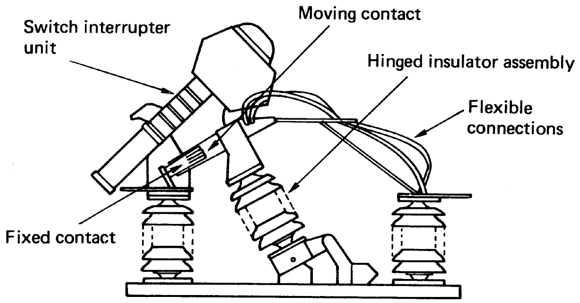


Figure 34.3 Pole-mounted switch disconnector

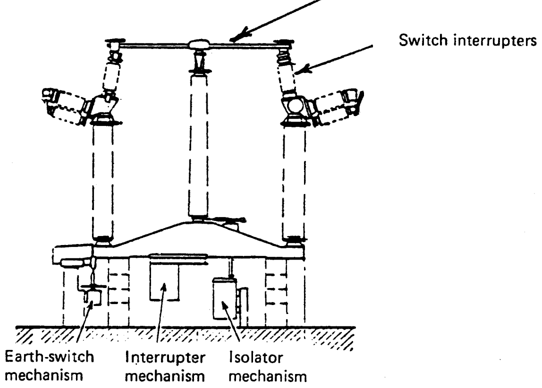


Figure 34.4 A 420 kV switch disconnector

At low voltages earth switches are seldom applied, as circuits can readily be tested to check whether they are live. Once proven to be de-energised, temporary earth connections can then be fitted. At distribution voltages many accidents have occurred over the years with operators inadvertently applying an earth to a live circuit and current practice is for fault-making earth switches to be provided for all applications. At transmission voltages, where instruction and permit systems can more readily be relied upon, and particularly for open-type substations where visible isolation is readily available, non-fault-making earth switches are nearly always used. For gas-insulated earth switches at transmission voltages where isolating distances may not be readily visible, the practice of using the safer fault-making earth switch is now becoming more common.

34.1.5 Fuses

A fuse is a 'one-shot' device capable of carrying the rated load current of the circuit in which it is situated, defined circuit overload conditions for predetermined times and of clearing overcurrents or short-circuit currents associated with faults which may occur on the system. Once operated on short-circuit, the fuse link, a component part of a fuse, must be replaced. (A fuse comprises all the parts which form the complete device and incorporates terminations for connecting into a circuit.)

A fuse is a thermally operated device. It comprises some form of conductor which, when subjected to predetermined currents for predetermined times, melts or blows to clear the circuit. The simplest form of fuse link comprises a wire held between terminations in air.

34.1.5.1 Rewirable fuses

For LV applications a wire fuse is generally described as a 're-wirable' or 'semi-enclosed' fuse. These are commonly used in domestic applications for protecting radial and ring circuits, their main advantage being that they are readily repairable by the layman, providing appropriate fuse wire is available. However, they can readily be abused. Their rated short-circuit breaking capacity is low, typically 2–3 kA at 240 V, although some designs may achieve a rating of 10 kA. As they break in air they require the passage of a current zero in order to clear, hence the total let-through energy of such a fuse is high and the 'downstream' circuit must be designed to cater for this passage of short-circuit current and associated energy. In addition, the fuse holder, which contains the wire fuse element tends to carbonise with repeated fault operations and the fuse may eventually fail to clear. A further shortcoming is that for the fuse to operate, the fuse element must reach a temperature which will cause it to melt; depending on the element material, this may be of the order of 1000°C or more. Also, for long-duration overloads over-heating may well occur.

Notwithstanding these potential limitations re-wirable LV fuses are very economical and are still widely used in domestic applications, although latterly the tendency has been to use the inherently better performance of the cartridge fuse or miniature circuit-breaker.

Re-wirable fuses are also used at distribution voltages up to 33 kV, but in some countries they may be used at even higher voltages. Such a re-wirable fuse is generally referred to as an expulsion fuse. This comprises a fuse element, i.e. that part of the fuse designed to melt when the fuse operates, which is generally enclosed within a small insulated tube. Each end of the element has a flexible tail for connecting to the terminals of the fuse carrier. The fuse carrier also generally comprises an insulating tube through which the fuse element is inserted. The fuse element is then held under tension by spring-loaded contacts into which the fuse carrier is fitted. These contacts are secured to the fuse base which comprises one or more insulators and a metallic base for installation on to an overhead-line pole.

The assembly, as described, is placed up a pole and connected onto the overhead line. It operates in atmospheric air and may be subjected to all types of weather conditions. When the fuse operates the fuse-carrier assembly usually hinges about its bottom contact and drops down, thus providing a visible indication of operation. Such a fuse assembly may also be designed to provide isolating facilities and thus can be used as a disconnection point.

When an expulsion fuse operates an arc is formed at the appropriate point of the fuse element, gases are produced from the small insulated tube surrounding the element; for small overloads these gases may be sufficient to build up pressure to extinguish the arc. For higher overloads or short-circuit currents the rapid pressure build up will cause the inner tube to burst and gases will then be generated by the arc impinging on the inner wall of the fuse-carrier tube. When sufficient pressure is built up the arc is extinguished and this pressure expels the ends of the fuse element. During such an operation a loud detonation or bang will be heard and hot ionised gases will be expelled from both ends of the fuse-carrier tube. A typical expulsion fuse is shown diagrammatically in *Figure 34.5*.

Like the semi-enclosed LV fuse, the expulsion fuse also requires the passage of a zero current in order to clear and thus the let-through energy is high. This results in the breaking capacity being relatively low, typically 8 kA at 12 kV. Expulsion fuses are very economical devices and are particularly effective at protecting overhead-line spur connections. However, replacement of an operated expulsion fuse can be costly in terms of the manpower required to be sent to the appropriate location. The problem can be alleviated to a large extent by the use of an electronically operated sectionalising link which fits into an expulsion fuse base. The link differentiates between transient and permanent faults and operates to isolate a permanently faulted circuit during the open period of the feeding auto-reclosing circuit-breaker.

34.1.5.2 *Current-limiting fuses*

The most widely used type of fuse for industrial applications is the current-limiting cartridge fuse. The main advantage of a cartridge fuse over the re-wireable fuses described is that they are perhaps the most effective circuit-protecting device available in that they will operate very rapidly during

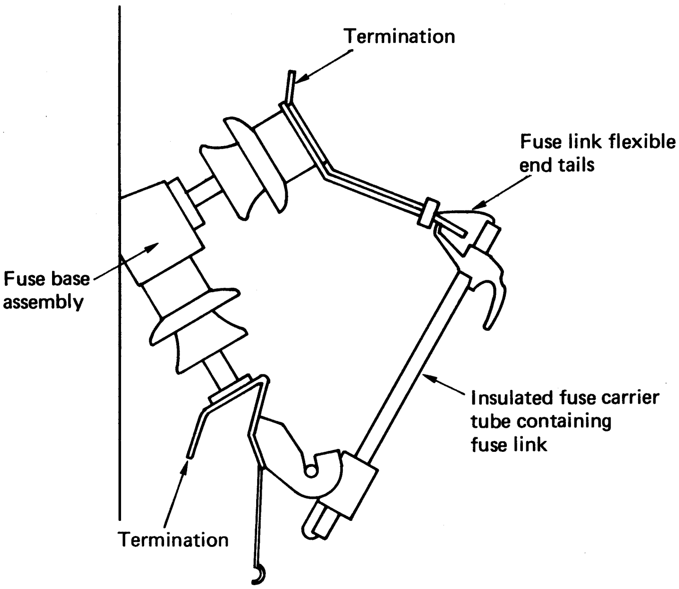

Termination
Fuse link flexible end tails
Fuse base assembly
Insulated fuse carrier tube containing fuse link
Termination

Figure 34.5 Typical expulsion fuse

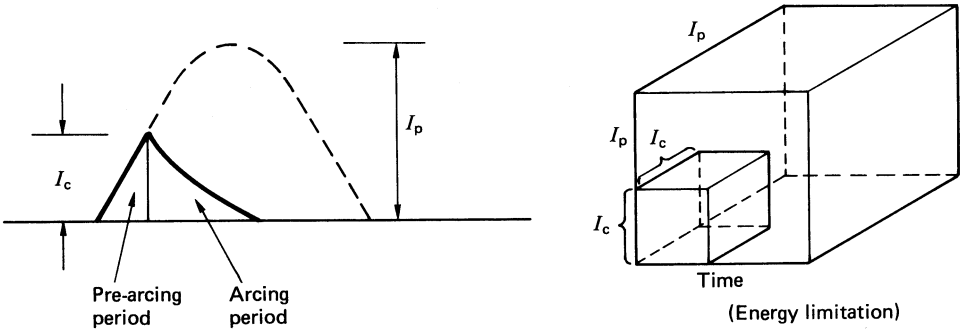


Figure 34.6 Fuse cut-off current

the initial part of the power frequency short-circuit current to clear the current at a low level before it reaches its peak value. Such a phenomenon is referred to as 'cut-off' and is shown in *Figure 34.6*

In consequence, the downstream circuit will never see the full short-circuit current of the system and economies in circuit design can be achieved. A further advantage of this rapid cut-off of current is that energy fed through to the faulted point is also considerably limited; this is shown diagramatically in *Figure 34.6*.

A cartridge fuse exhibiting these properties is known as a 'current-limiting fuse'. Current-limiting fuses are generally available from LV up to about 72.5 kV. Their major application in HV distribution being for voltages in the range 10–20 kV. Typical short-circuit breaking currents at LV are normally up to 80 kA and in the HV range up to 20 kA. However, short-circuit ratings as high as 200 kA at 415 V can be achieved and recent development of a HV fuse at 200 kA, 20 kV has been successfully achieved. In general interruption of high short-circuit currents does not constitute a difficult problem to a current-limiting fuse since the current and resultant energy is kept to a very low level by the phenomenon of cut-off.

The construction of LV fuses is somewhat different from HV fuses, but both operate in a similar manner. The current-limiting fuse is always in the form of a cartridge comprising a specially designed and proportioned element or number of parallel elements enclosed within an insulating barrel, usually of a ceramic material. The barrel is fitted with end caps to which the fuse elements are attached, additional end caps may then be fitted to carry the terminations for connecting to the fuse holder. Before the outer caps are fitted the inner assembly of fuse link is filled with fine-grained silica sand. The general construction of a typical LV fuse link is shown in *Figure 34.7*.

The fuse element may take a number of different forms, but the basic design criterion is to achieve melting of the element at defined points to create a number of arcs in series along the length of the fuse element. Restrictions are incorporated in the element to ensure a greater current density at these points and, for high short-circuit currents, there is little time for heat to dissipate along the length of the element and the element thus melts at these restrictions. An arc is formed at each point and energy is removed from the arc by the fine granules of sand which melt and fuse together to form a glass-like substance known as 'fulgurite'. The rate of extraction of energy must be greater than the rate of build up of energy in order for the arc to be extinguished. In practice this process is very rapid and will only last for some 2–3 ms or less and it is this phenomenon which gives the cartridge fuse its rapid current-limiting capabilities. Typical fuse-element forms are shown in *Figure 34.8*.

The fuse-element material is generally silver, but copper is now widely employed for LV applications.

In order to avoid overheating problems as might occur with small, long-duration overcurrents, the fuse element

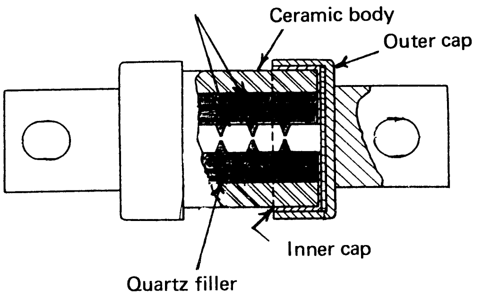


Figure 34.7 Construction of a typical LV fuse link

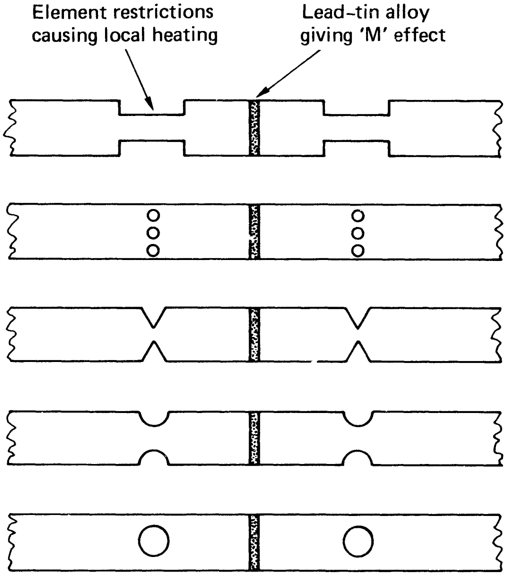

Element restrictions causing local heating
Lead-tin alloy giving 'M' effect

Figure 34.8 Typical fuse-element forms

often employs a small deposit of low-melting-point alloy at some point across its width. The alloy solder deposit causes an eutectic effect whereby the melting point of the base material is considerably reduced, e.g. from 960°C for silver to about 230°C for the alloy. This is often referred to as the 'M' or 'Metcalf' effect.

For HV current-limiting fuses it is necessary to employ a significantly longer fuse element comprising a larger number of restrictions along its length. A greater number of arcs in series must be generated in order to produce a back-e.m.f. capable of overcoming the applied voltage to ensure arc extinction. A typical 11 kV fuse link may require an element length of approximately 1 m. Clearly, a fuse link of this length would be impracticable and costly. The problem is overcome by winding the fuse element or elements in a spiral form around an inner ceramic star core support. This enables the element of an 11 kV fuse to be accommodated in a fuse barrel length of some 250 mm. Since the spacing between adjacent sections of the spirally wound elements must be sufficient to ensure that flashover does not occur across this gap, only a limited number of parallel elements can be employed. This limits the normal current rating to typically 80 A for a fuse link of some 250 mm length and 50 mm diameter. Higher current ratings at the same voltage can be achieved by increasing the barrel length, a typical length being 350 mm for which ratings of some 125 A may be achieved at 11 kV. Even higher current ratings can be achieved by connecting a number of fuse links in parallel. This requires more space, becomes more costly and is seldom applied at HV

A typical HV fuse construction is shown in *Figure 34.9.*

A further feature of a HV current-limiting fuse is that it can generally be fitted with a fuse-striker device which ejects when the fuse operates. The striker can be used as an indication device or, more generally, to cause an associated switch to trip so that for single-phase faults all three phases can be disconnected by operation of the striker activated switch. This approach prevents two phasing of HV circuits and alleviates consequential problems that might occur for example on three-phase motors.

Striker assemblies are of two basic types, being either activated by an explosive charge or compressed spring. The former is generally used within the UK where three-phase tripping is a requirement. This tends to necessitate the larger output of an explosively operated striker. Where indication or three-phase tripping is not required or where very light loaded switch trip mechanisms are used the lower energy spring-operated striker can be applied. Spring-activated strikers are commonly used in continental European fuse designs. Both types of striker are operated by a very thin fuse wire running the length of the fuse, generally located within the centre section of the star core assembly.

The striker wire may only require a few amperes to operate. When it does so it activates and ignites the gunpowder charge in the explosive design of striker or allows the release of the precharged spring in the spring striker assembly. Once the striker has operated to cause an associated switch to trip it must not be possible to reclose the switch until the operated fuse has been replaced. Thus a fundamental requirement is that it must not be possible to push the operated striker back into its housing. This is normally achieved by employing a ratchet-type serration on the striker pin which engages with a flat spring washer.

Typical explosive and spring-operated striker assemblies are shown in *Figure 34.10.*

With a HV fuse link it may not be possible to generate sufficient arc energy and consequent back-e.m.f. when operating under very low overload conditions to ensure that clearance will be achieved. It is thus common for HV fuses to have a specific minimum breaking capacity below which fuse operation cannot be guaranteed. Some other form of circuit-overload-protection device should be employed to cater for long-term overcurrents which may be of values between the fuse rated current and the fuse minimum breaking current. Such a type of fuse is referred to as a 'back-up' fuse and the fuse manufacturer will specify its minimum breaking current so that the user can ensure that coordination is achieved throughout the required operating range of the circuit-protecting devices.

HV fuses having very low minimum breaking currents can be produced, and are referred to as 'general-purpose' fuses. By definition, a general-purpose fuse is one that will operate at all values of short-circuit current or overload current down to a value of current which will cause the fuse to operate in 1 h. Even such fuses may have limitations in attempting to clear values of overload current between the rated current and the 1 h melting current. In consequence, there are now designs available which will clear all values of current to which the fuse may be subjected, even values below the rated current, if for example the fuse link is situated in a very high temperature environment. Such a fuse link is generally referred to as a 'full range' fuse link.

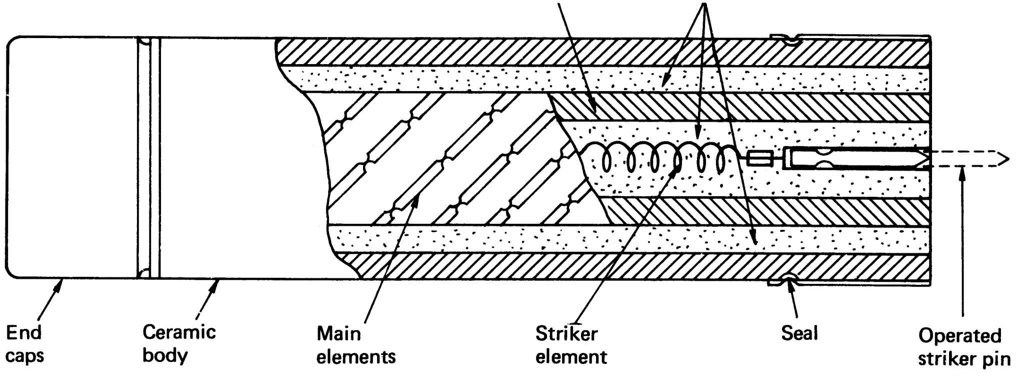


Figure 34.9 HV fuse construction

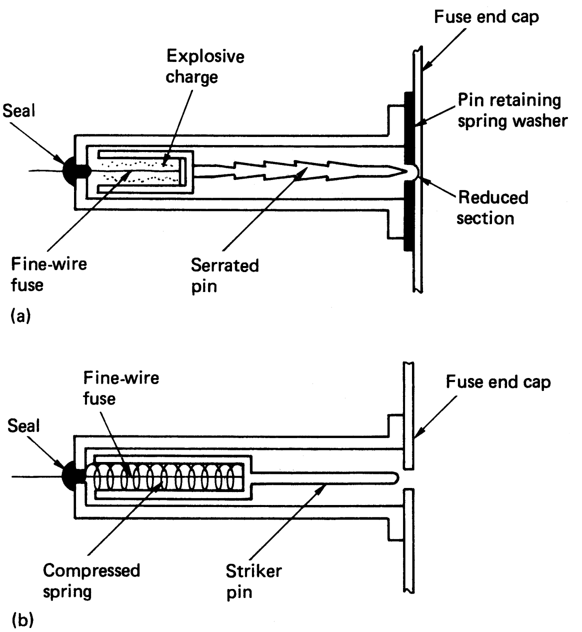


Figure 34.10 Typical explosive (a), and spring-operated (b), strikers

However, means of adequately verifying the performance of such a fuse is still under discussion within the appropriate international standards committees.

34.1.5.3 Fuse characteristics

A fuse exhibits a number of different performance characteristics of which the user needs to be familiar in order to ensure its correct application. The main advantage of a current-limiting fuse link is, as described, its ability to limit very rapidly the flow of fault current. Hence, in order to ensure adequate 'downstream' circuit design, it is necessary to know its cut-off current performance. This will vary depending on the current rating of the fuse and the value of prospective short-circuit current. A typical fuse cut-off characteristic is shown in *Figure 34.11*.

As the prospective short circuit is decreased for a given fuse rating, the proportional value of cut-off current to prospective peak current will increase until such a time as the cut-off current equates to the peak prospective current. Below this value of prospective current the fuse link will no longer exhibit cut-off.

A further important characteristic for 'downstream' circuit design is the total let through energy of the fuse link.

However, since energy is dependent on the total circuit resistance, an additional quantity is normally referred to, this being the 'joule integral', i.e. the square of the current over a given time interval:

$$I^2t = \int_{t_0}^{t_1} i^2 \mathrm{d}t$$

Specific values are quoted for each fuse rating and these are given as I^2t. The values generally quoted are pre-arcing time I^2t, i.e. the I^2t integral over the pre-arcing time of the fuse, and the arcing I^2t, i.e. the I^2t integral over the arcing period of the fuse. The sum of these two characteristics is referred to as the 'total operating I^2t'. Typical I^2t characteristics are shown in *Figure 34.12*.

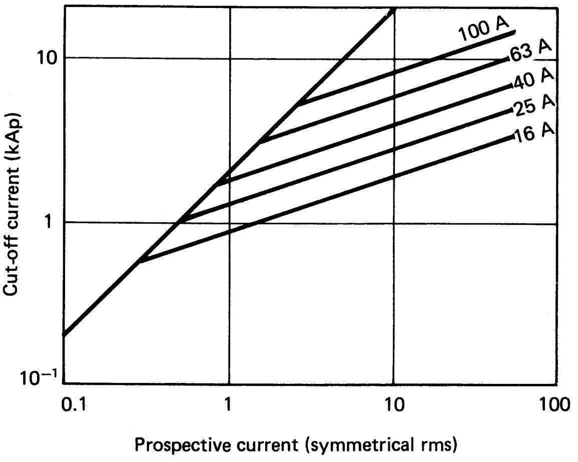


Figure 34.11 Cut-off characteristics

The total operating I^2t is important in designing the 'down-stream' circuit because its energy-handling capability must be in excess of this. Also, in order to achieve discrimination it is important to ensure that the total operating I^2t of the 'down-stream' fuse does not exceed the pre-arcing I^2t of the main fuse.

The I^2t characteristics are important parameters when assessing the circuit performance under high values of short-circuit current. With values of short-circuit current below the value which will cause the fuse to exhibit cut-off, it is then necessary to examine the fuse characteristics for these lower values of current. These characteristics are known as time–current characteristics and are usually presented in graphical form on log–log paper. Typical time–current characteristic zones are shown in *Figure 34.13*.

Each manufacturer's time–current characteristics will differ slightly for a given current rating, depending on the specific design features used. Thus specific time–current

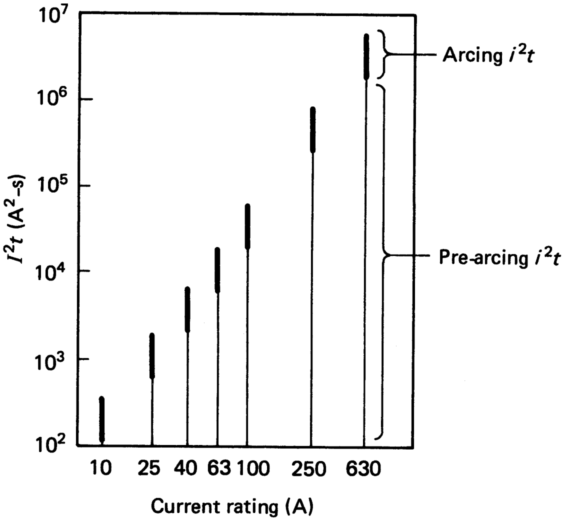


Figure 34.12 I^2t characteristics

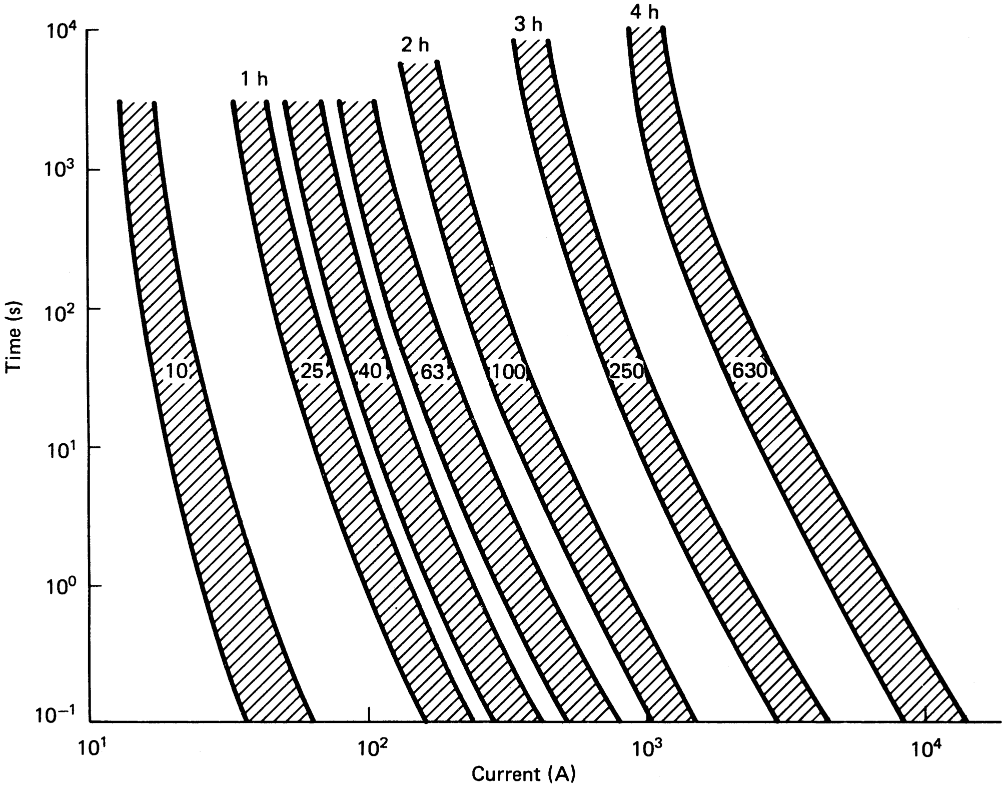


Figure 34.13 Time–current characteristic zones

characteristics cannot be used as the basis for circuit design as the designer may wish to be able to employ any particular manufacturer's fuse link. In order to overcome this problem, time–current characteristic bands are specified for LV fuse current ratings in the appropriate international fuse specifications.

34.1.5.4 Fuse applications

From examination of the fuse characteristics, it is evident that fuses are suitable for very many applications. Fuse designs have evolved around idealised characteristics for the protection of specific circuits. For example, recent changes within the international committees dealing with wiring regulations now require a close degree of protection of downstream cable circuits and protection against electric shock. The new regulations call for a discrimination ratio between major and minor fuses of 1.6 : 1. This requirement is met by modern LV fuses based on pre-arcing I^2t at a time of 10 ms.

Many fuses are used to protect motor circuits. With this application it is imperative that the fuse does not operate with the current surge present when starting a motor. For direct on-line starting this may require a possible current of six times the motor full-load current for some 10 s. For motor protection a steep time–current characteristic is desirable such that the fuse will clear quickly for high values of short-circuit current. For overload, these characteristics would not be ideal and it is common practice for separate overload protection to be provided. Such a characteristic is shown in *Figure 34.14*.

The motor fuse characteristics would not be ideal for protection of a transformer circuit where the fuse must be able to withstand the transformer in-rush current, typically $10 \times I_{\mathrm{FL}}$ for 0.1 s. It must also quickly clear faults within the transformer circuit and may have to discriminate with fuses on the LV side of the transformer. A shallow-sloped characteristic is more ideal (see *Figure 34.15*).

For HV fuses there are now separate international specifications covering both motor- and transformer-protection requirements. Specially developed LV fuses are available for protecting semiconductor devices. These must have a low I^2t let-through, low cut-off current and very low overvoltage on operation. In order to achieve the low I^2t such fuses tend to run very hot and the 'M' effect is generally not applied. They are not suitable for use in enclosed fuse holders and must be mounted so as to allow adequate air circulation. They generally have substantial terminations to assist in heat dissipation.

34.1.6 Fuse switches

A fuse switch is a switch which is connected in series with a fuse or, more precisely, the fuse is mounted on the moving contact system of a specially designed switch. An alternative device, a switch fuse, has the switch electrically connected to a series fuse to form a composite unit. This arrangement has disadvantages in that when changing a fuse the electrical circuit to which it is connected may still be live. Furthermore, if the switch is connected to the live side of the circuit the switch may not be capable of closing against

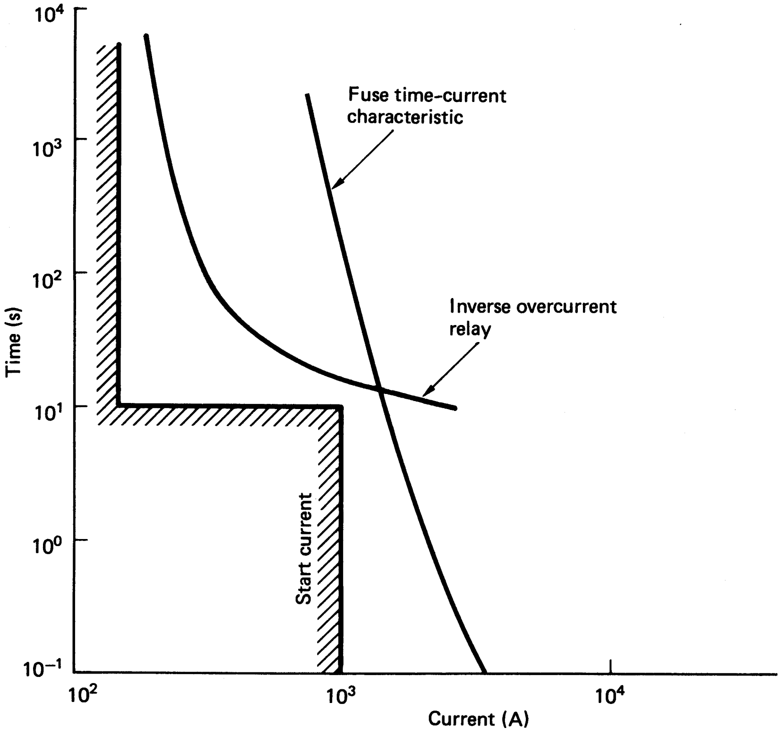


Figure 34.14 Motor starting fuse characteristics

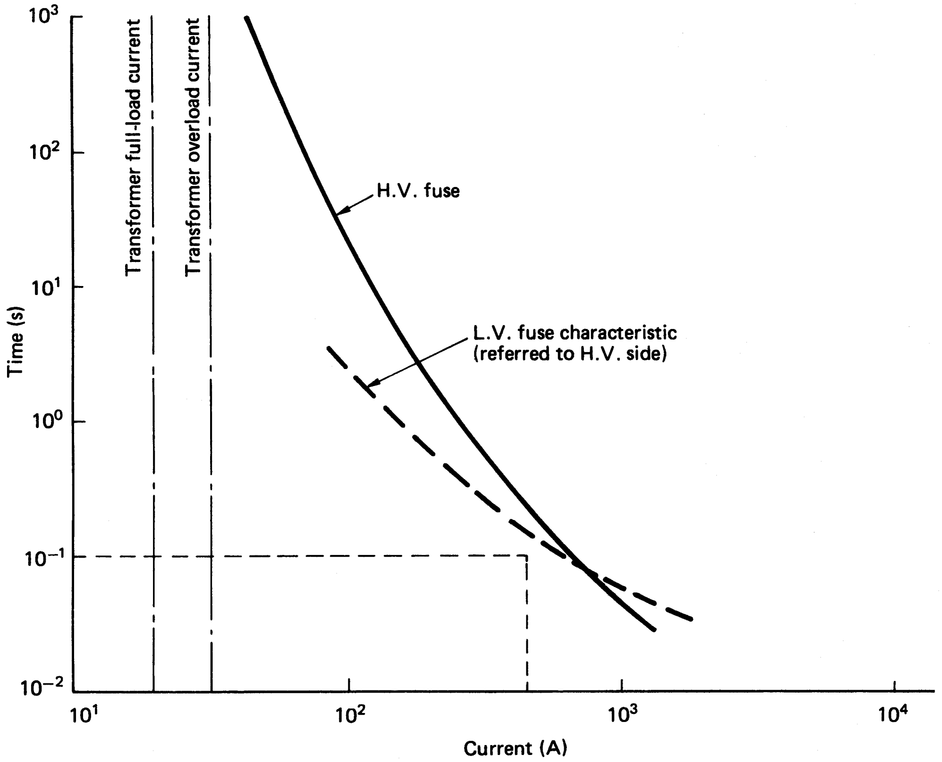


Figure 34.15 Transformer fuse characteristics

a fault between the switch and fuse. A fuse switch is inherently a safer device and is now widely used.

The advantage of the fuse switch is that it is an economical switching device that can safely make and break, by means of the fuse, short-circuit currents that may occur on the 'down-stream' system. Fuse switches are usually manually operated by means of an over-centre spring-operating mechanism which ensures a consistent and adequate opening and closing speed.

Fuse switches are perhaps the most commonly encountered mechanical switching device on industrial switchboards using three-phase LV circuits. A typical LV fuse switch is shown in *Figure 34.16*.

HV fuse switches operating at voltages of up to some 20 kV are widely employed on urban secondary distribution networks. In such an application three-phase operation of the fuse switch is necessary for any single-phase fuse operation and this is achieved by means of the fuse striker pin (see Section 34.1.5.2) operating the HV switch-trip bar. In addition, co-ordination is applied between the switch and fuse to ensure that for currents below the minimum breaking current, for a back-up fuse, the switch will trip satisfactorily, by fuse-striker operation, to clear an overloaded circuit. HV fuse switches use fuses mounted either in air or under oil, the latter being the more widespread application. When a fuse is mounted under oil the fuse design must incorporate appropriate sealing to ensure that oil does not enter the fuse body as this may result in fuse maloperation. A cross-section of an oil-filled fuse switch is shown in *Figure 34.17*.

34.1.7 Contactors

A contactor is capable of performing much the same switching duty as a switch and may also be capable of closing against a downstream short circuit when protected by an appropriate fuse. Its main difference is that it is capable of performing very frequent switching operations such as may occur, for example, with industrial processes where frequent stops and starts are required. In consequence it is necessary that it is capable of being remotely operated and, therefore, a spring-operated mechanism, for example, is unsuitable. The majority of contactor designs use a solenoid operating mechanism. This may be a continuously energised coil which holds the contactor closed, i.e. 'electrically held', or a short-time rated coil which closes the contactor via a mechanical operating mechanism and causes the mechanism to latch to hold the contactor in the closed position, i.e. a 'latched contactor'. Opening of a latched contactor requires the provision of a trip coil to operate a trip bar to release the latched mechanism. Most contactor-operating coils are energised from the source of supply; on loss of supply, all electrically held contactors would open and, unless appropriate protection were provided, would re-close on reinstatement of supply. For many industrial processes

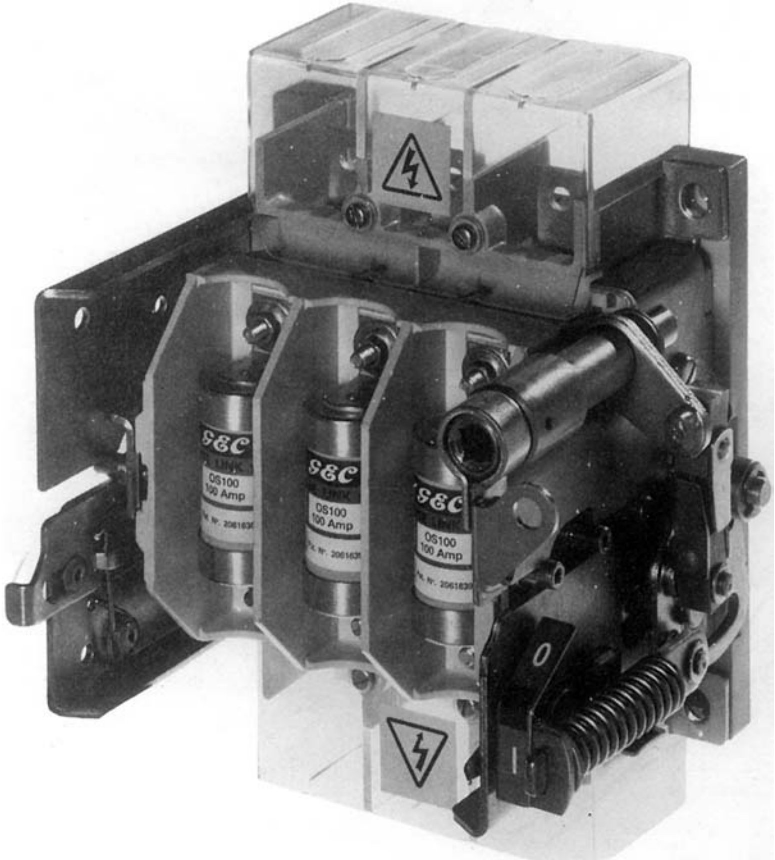


Figure 34.16 L.v. fuse switch

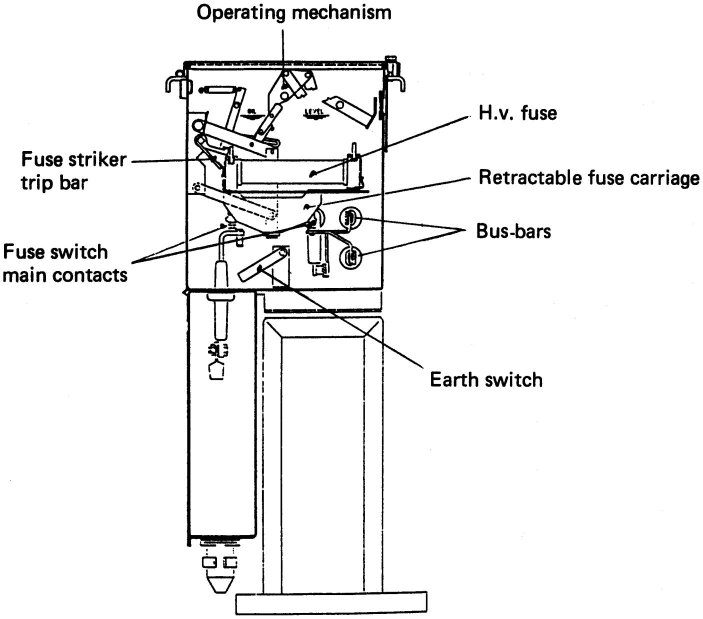


Figure 34.17 Cross-section of an oil-filled fuse switch

this feature may prove to be unsatisfactory, in which case latched contactors would be used.

A contactor, like a switch, needs to break inductive load currents and in order to achieve this within compact dimensions some form of arc control system is employed. This usually comprises a number of parallel disposed metallic plates separated by small gaps into which the arc is forced. These plates cause a number of separate arcs to be formed and at the same time they cool and extract energy from the arc to achieve arc extinction.

Early designs of contactors comprised a large open-clapper-type single-break construction. These have now largely been replaced by more modern, compact and economic double-break block construction contactors.

34.1.8 Circuit-breakers

A circuit-breaker is a more sophisticated mechanical switching device in that, in addition to making and breaking load and overload currents of the circuit, it is also capable of both making and breaking the full-rated short-circuit current of the system.

In order to break the full short-circuit current of the system, very sophisticated arc control mechanisms have been evolved over the years. The techniques of arc extinction were developed originally on a trial-and-error basis, and were considered to be more of an art than a science. However, with modern knowledge and experience, interrupter performance can more readily be predicted using advanced theory and computer techniques. Nevertheless, it is still necessary to perform actual short-circuit tests to verify the performance of a circuit-breaker.

In order to be able to clear a faulted circuit some means of detection of the fault is necessary and that information must be transferred into a signal to cause tripping of the circuit-breaker. For LV systems direct-acting overload coils may be employed or more sophisticated relays used. For HV systems it is necessary to ensure that the tripping signal is at LV, i.e. it must safely be segregated from the HV system. This is achieved by the use of current and voltage transformers. The output from these devices is fed into appropriate relays which operate under predetermined conditions to cause their contacts to make and energise the circuit-breaker tripping circuit.

Tripping supplies are usually obtained from continuously charged batteries operating at a nominal 125 V d.c. The most common method of satisfactorily achieving low tripping energy is to use mechanically held or latched operating mechanisms. This type of mechanism requires some form of prestored energy which can be released when required to cause the mechanism linkages to operate to close the circuit-breaker. Once the circuit-breaker has been closed and the stored energy dissipated, the mechanism must hold the circuit-breaker in the closed position by virtue of its interconnecting linkages, usually by some form of mechanical latching. The mechanism can then only be released to open the circuit-breaker by operation of a low energy trip bar operated from a trip coil plunger.

The most common circuit-breaker operating mechanism is the spring charged mechanism whereby the springs are charged either manually by means of a detachable spring charging handle or by means of an electric motor and geared drive system. The latter offers the advantage of remote charging. The charged springs can then be released by means of a small solenoid coil to close the circuit-breaker. An alternative closing energy source may be derived from a large solenoid coil but, whilst again this offers the facility for remote closing, it has the disadvantage of requiring a large d.c. supply.

This supply can either be fed from substation batteries or from a rectified supply derived from a voltage transformer. Either source of d.c. supply can be expensive and the motor rechargeable spring is now more widely employed.

For lower fault rated circuit-breakers the over-centre spring-operating mechanism, as used in most switches, is sometimes employed. Higher fault rated circuit-breakers, particularly at transmission system voltages, use either pneumatic or hydraulically operated mechanisms. These may take a number of different forms. For example, compressed air may be used to close a conventional mechanical operating mechanism to achieve mechanical latching with tripping being achieved by release of the latched linkages. Alternatively, the compressed air may be released through a series sequence of valves to operate a piston which in turn operates through direct mechanical linkages to the circuit-breaker contacts. A further alternative is where the compressed air is used directly to operate the circuit-breaker contacts.

Similarly with hydraulic mechanisms, a hydraulic pump may be used as a method of remotely charging the operating mechanism spring. This arrangement is seldom employed nowadays. Alternatively, the hydraulic system itself may be used to transmit the drive via a piston to operate the circuit-breaker moving contacts directly. With such an arrangement the stored energy required to operate the circuit-breaker is derived from an accumulator. This comprises a cylinder with a central 'floating' piston, one side of the cylinder being filled with a suitable compressed gas, usually nitrogen, and the other side of the cylinder being filled with hydraulic oil which is connected to the hydraulic pump. The pump operates to pressurise the system, the stored energy being contained within the pressurised nitrogen compartment of the accumulator. A series sequence of valves is then used to allow the stored energy to be released, via the hydraulic pipework to operate a piston to close the circuit-breaker.

Circuit-breakers have evolved in a number of different ways to suit particular applications and use a number of different arc interruption techniques. The types of circuit-breaker typically encountered are described in the following sections.

34.1.8.1 Miniature circuit-breakers

Miniature circuit-breakers are only used at LV, mainly in domestic or light-industrial or commercial applications. In general they are used in the same applications as semi-enclosed or cartridge fuses and offer an alternative for protecting radial or ring circuits. They are usually only single-phase devices and have a typical rated load current range of up to 100 A with a maximum short-circuit rating of 16 kA at 240 V. Manually operated over-centre spring-operating mechanisms are used. Miniature circuit-breakers usually employ a series overload coil for rapid short-circuit tripping and a bimetallic element for tripping on overloads. All miniature circuit-breakers operate on the air-break principle where an arc formed between the main contacts is forced, by means of an arc runner, and the magnetic effects of the short-circuit currents, into metallic arc splitter plates. These cause a number of series arcs to be formed and at the same time extract energy from the arc and cool it to achieve arc extinction. With some designs of miniature circuit-breaker this arc interruption process can be so rapid that current cut-off can be achieved in much the same way as described for a current-limiting fuse. The principle of a typical miniature circuit-breaker is shown in *Figure 34.18*.

Miniature circuit-breakers do not, however, provide rapid operation for low values of earth leakage current. Modern day wiring regulations require that very rapid operation is achieved in the event of an earth fault to minimise the dangers from electrocution. This requires operation for earth fault currents as low as 30 mA in a time of some 2–3 ms.

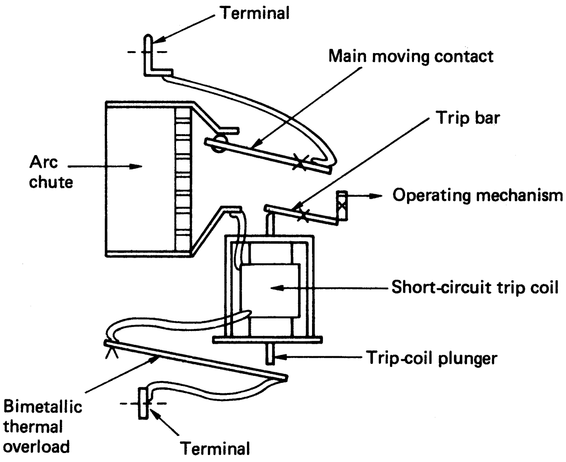


Figure 34.18 Principle of miniature circuit-breaker

A variation on the basic construction of the miniature circuit-breaker is used to achieve this requirement. Such a device is commonly known as an 'earth-leakage circuit-breaker', although the correct terminology is 'residual-current device'. Tripping at such low values of earth leakage current is achieved by passing both the feed and return conductors through an integral current transformer. Under normal conditions the resultant flux in the current-transformer core is zero. Under earth-fault conditions the feeding and return currents will not be of the same value and the current difference causes a flux to be generated within the current transformer core producing an output voltage at its secondary winding terminals which is used to energise the tripping circuit of the residual-current device.

Residual-current devices may be permanently wired into an installation or used as plug-in devices to protect domestic electrical appliances.

One feature of the miniature circuit-breaker and residual-current device is that their contacts are not maintainable and after a limited number of operations replacement of the device may be necessary. However, in practice, this is seldom a significant limitation and eroded contacts can usually be detected by overheating causing nuisance tripping of the device.

34.1.8.2 Moulded-case circuit-breakers

Moulded-case circuit-breakers are also only used for LV applications. They are basically an upgraded version of the miniature circuit-breaker and are invariably three-phase devices. They have typical current ratings ranging from 100 to 2500 A and may have rated short-circuit ratings up to 50 kA at 415 V. Some designs also exhibit cut-off similar to the current-limiting fuse and all are provided with inherent short-circuit and thermal overload protection devices. They may also be provided with earth-leakage protection.

There are two categories of short-circuit performance: P1 and P2. Category P1 requires the circuit-breaker to be capable of performing two short-circuit operations, an open (O), operation followed by a close/open (CO) operation. Subsequent to this duty the circuit-breaker may not be capable of performing its normal requirements and should be replaced. Category P2 requires the circuit-breaker to perform a further CO duty and subsequent to this test there must be no reduction in its current carrying performance.

In applying moulded-case circuit-breakers care must be taken to ascertain the circuit prospective short-circuit level and probable frequency of occurrence of faults. For circuits where the frequency of faults is high a category-P1 moulded-case circuit-breaker is likely to be inadequate.

Whilst some designs offer facilities for contact maintenance others do not, in which case erosion of contacts will only be detected by nuisance tripping resulting from overheating.

As the name implies, moulded-case circuit-breakers are invariably completely enclosed in a premoulded casing. There is usually an on/off operating toggle on the front of the unit with the three-phase terminals at the top and bottom of the unit. As for the miniature circuit-breaker, an over-centre spring-operating mechanism is usually employed.

Moulded-case circuit-breakers are generally used in similar applications to fuse switches, i.e. protection of large three-phase LV loads and for motor-starting applications. They do not, however, have the facility for remote or frequent operation and cannot be used to replace a contactor where frequent starting is likely to be a requirement. A typical moulded-case circuit-breaker is shown in *Figure 34.19*.

34.1.8.3 Air circuit-breakers

An air circuit-breaker uses atmospheric air as its interrupting medium. The arc is drawn between its contacts and extended via arc runners on to an arc chute where it is presented with a large cooling surface of arc splitter plates. These break the arc into a number of series arcs, the principle being to increase the resistance of the arc and extract energy from it via the metallic splitter in much the same way as described for LV switches, contactors and miniature circuit-breakers. A typical arc chute is shown in *Figure 34.20*.

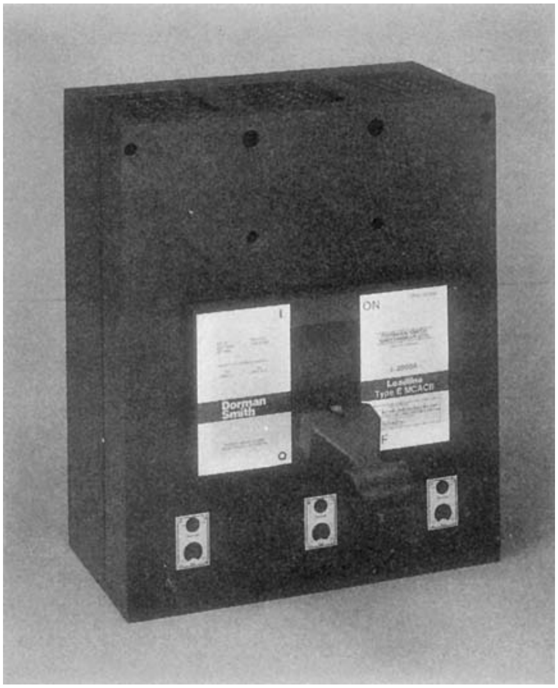


Figure 34.19 A typical moulded-case circuit-breaker

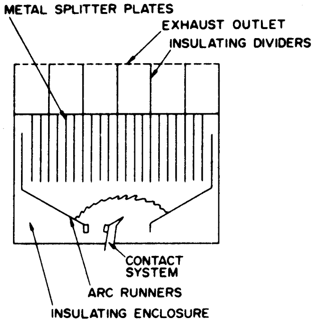


Figure 34.20 Arc chute interrupter used for air circuit-breakers

Free-air circuit-breakers are used for LV applications and HV applications up to some 20 kV. They can have very high rated currents of typically up to 4000 A and very high short-circuit interrupting capabilities of typically up to 90 kA at 12 kV. Their main application at LV is where an onerous performance is required in terms of load, number of operations and fault level. Mainly due to economic considerations, moulded-case circuit-breakers have replaced many LV applications where previously air circuit-breakers were used but where high performance, maintainability and long-term reliability are essential requirements air circuit-breakers are still used. A typical application is in generating-station LV auxiliary supplies.

The main application of HV air circuit-breakers has been in applications where the exclusion of flammable materials is a fundamental requirement. Again a typical application being generating-station HV auxiliary supplies. Such large high-rated air circuit-breakers are, however, extremely costly and their use is now tending to diminish in favour of high performance vacuum or SF_6 circuit-breakers.

A further application of air circuit-breakers is for use with d.c. supplies, this method of interruption still being the most suitable for d.c. circuits. D.c. circuit-breakers are widely used in traction applications where ratings of up to some 3 kV may be employed.

34.1.8.4 Air-blast circuit-breakers

Air-blast circuit-breakers use a blast of compressed air at a pressure of 25–75 bar which is directed across the arc path to cool and remove the ionised gas. Air-blast circuit-breakers are fast in interruption, which occurs usually at the first or second zero current, and arc lengths are short. The compressed air needs to be stored locally at the circuit-breaker within its own air receiver. Subsequent to operation the compressed air local storage needs to be replenished from a compressor system. This is usually a central system feeding all circuit-breakers via a suitable ring-main network.

Air-blast circuit-breakers are of two basic types, one where the interrupter contacts reclose subsequent to the blast of air, usually referred to as a ‘sequentially isolated circuit-breaker’ and the other where the interrupter contacts remain in the open position after the passage of the air blast and are continually pressurised to maintain the dielectric strength. This latter type is referred to as a ‘pressurised-head circuit-breaker’.

With the sequentially isolated type, air is admitted via a blast valve at the bottom of the vertical support insulator. A high pressure ceramic blast tube is necessary to transfer this blast of air up to the main interrupter contacts to cause them to open against a return spring. This type of circuit-breaker also incorporates a free air disconnector in series with its interrupters. During the period in which the circuit-breaker contacts are held open by the blast of air and, after arc extinction, the disconnector is automatically operated to open the circuit. Once the sequential disconnector is fully open the main blast valve recloses to cause the main contacts to reclose under their own return spring pressure. For closing, the circuit is made on to the contacts of the sequential disconnector. These contacts must be very robust as they must close against the full short-circuit current of the system.

Sequentially operated air-blast circuit-breakers are used typically at voltages of up to 420 kV. Their main application is at transmission voltages. Their application at distribution voltage is minimal. Ratings of 50 kA at 420 kV with normal current ratings of 4 kA are possible. The interrupting capability per interrupter is limited and it is necessary to employ a number of series interrupters per phase, 12 breaks being typical at 420 kV. It is necessary to ensure that the interrupting duty is shared equally across each break and parallel capacitors or resistors are often employed for this purpose. Sequentially operated air-blast circuit-breakers invariably employ pneumatic operating mechanisms.

With pressurised-head circuit-breakers the arc is extinguished in much the same way as for the sequentially operated air-blast circuit-breaker. The main difference is that this type employs a main operating valve on the exhaust side of the interrupter such that once interruption is complete the open contact system is pressurised, at the full pressure of the compressed-air system, to maintain the dielectric strength across the open contacts. The contacts are held open permanently whilst the circuit-breaker is in its ‘open’ position. Interrupter contacts are generally driven mechanically via insulated pull rods from the main mechanical operating mechanism normally operated from a compressed-air-driven piston. The circuit-breaker is usually held closed by mechanical latching of the mechanism and opened by operation of a mechanical release trip bar, although some designs employ a compressed-air mechanism which is held open or closed by pneumatically operated valves.

Whilst the pressurised-head circuit-breaker alleviates the need for sequential isolation its main advantage is that, because the pressurised air is already at the interrupter contacts, very rapid arc extinction can be achieved, typical total break times being of the order of 40 ms. Such fast operation is necessary in order to ensure stability of the supply system when subjected to major faults. The very severe duties encountered often require the application of an interrupter in parallel to the main interrupter. When the main interrupter opens a resistor is then inserted in series with the parallel interrupter. This has the effect of damping the very severe transient overvoltage that might arise on circuit interruption, thus ensuring satisfactory fault clearance.

Such circuit-breakers may be used up to 800 kV with fault currents of some 80 kA or more. A number of series interrupter heads is necessary to achieve this rating—at 420 kV, for example, 10 or 12 interrupter heads may be required per phase. In addition to the resistor interrupters, parallel capacitors are also necessary to ensure equal voltage distribution across the interrupter heads during operation.

One phase of a typical pressurised-head air-blast circuit-breaker is shown in *Figure 34.21*. Such circuit-breakers are both mechanically and electrically very complicated and are very costly. Nevertheless, until the advent of SF_6 circuit-breakers, they were the only satisfactory means of interrupting very high values of fault current at the highest system voltages.

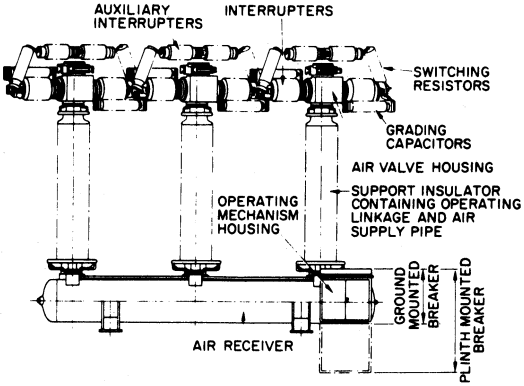


Figure 34.21 One phase of a pressurised-head air-blast circuit-breaker

Since all air-blast circuit-breakers expend, in a very short period of time, a large quantity of compressed air, they are very noisy in operation. To overcome this problem it is common to fit silencers to the exhaust systems of the circuit-breaker. The application of a silencer will reduce the noise level from approximately 120 to 90 dB.

Since the action of all air-blast circuit-breakers is dependent on arc extinction by a blast of compressed air, consistent fault-clearance times are achieved which are independent of the value of current interrupted.

34.1.8.5 Oil circuit-breakers

Oil circuit-breakers were the earliest devices used for satisfactorily achieving arc interruption. It was found that when an arc is drawn in hydrocarbon oil the arc energy decomposes the oil to form hydrogen (80%), acetylene (22%), methane (5%) and ethylene (3%), and the arc takes place in a bubble of gas surrounded by cooler insulating oil. The effect is that energy is extracted from the arc by chemical decomposition of the oil. Arc cooling is achieved mainly by the hydrogen gas which has a high thermal diffusion ratio; the surrounding oil also cools the arc plasma and the oil itself has a high dielectric strength when it flows into the arc path at zero current. Its good insulation properties also allow electrical clearances to be minimised.

The chief disadvantage of oil is that it is flammable. When expelled gases, formed due to decomposition of the oil on arc formation, mix with air these can be ignited as a result of a circuit-breaker maloperation to cause an explosion. For these reasons oil circuit-breakers are not used in generating stations where fire hazards are of paramount importance.

Early designs of oil circuit-breakers were of the plain-break design where contacts separated in oil without any specific means of controlling the arc. Performance was rather unpredictable, particularly under single-phase operating conditions and when drawing long inductive low-current arcs. For these reasons many plain-break circuit-breakers have now been withdrawn from service.

The problem was resolved in the late 1930s when the arcing process was controlled within an arc-control device. This device has the effect of constricting the arc, allowing a high pressure to be built up, which in turn assists in arc extinction. The gas, formed by the decomposition of the oil by the arc, is expelled through vents in the side of the arc-control device, allowing cool oil to flow across the arc path and cause rapid arc extinction (*Figure 34.22*).

Careful design of the arc-control device is necessary to allow the arc-interruption features to operate effectively. Each manufacturer has their own design of arc control device. All oil circuit-breakers now employ some form of arc-control device.

There are two basic forms of oil circuit-breaker: the bulk-oil circuit-breaker and the low-oil-content circuit-breaker. Both are widely applied with the former being traditionally used in the USA and the UK and the latter in Continental Europe.

The bulk-oil circuit-breaker comprises an earthed tank which contains the contacts and arc-control system. Connections are taken into and out of the tank by means of through-bushing insulators.

Operating mechanisms are external to the tank with drive linkage to the moving contacts taken via an oil-tight seal. For voltages up to 72.5 kV all three phases are contained in a common tank; above this voltage separate phase isolated tanks are usually employed. Up to 72.5 kV single-break contact systems per phase are used, although the major application uses double-break contact systems per phase.

Above 72.5 kV multibreak contact systems are necessary with six series breaks being used at 300 kV. Parallel resistors are usually employed to ensure equal voltage sharing across the breaks. Above 300 kV bulk-oil circuit-breakers become dimensionally unsatisfactory and uneconomical and are seldom employed.

Low-oil-content, or live tank minimum oil, circuit-breakers have only sufficient oil to surround the contact and arc control device and use phase segregated insulated enclosures. The main advantage of this design is that less oil is used and compact, economical designs can be achieved. The main disadvantage is that the small quantity of oil used per interrupter can soon become contaminated with carbon under frequent fault operations and the dielectric withstand capabilities across the oil and oil/insulator interface can deteriorate. More frequent maintenance is thus required than for an equivalent rated bulk-oil circuit-breaker. Low-oil-content circuit-breakers offer a convenient form of construction for transmission voltage open type switchgear in that the interrupters can be contained within a vertical hollow porcelain insulator and the assembly can be mounted on a support insulator.

Multiple interrupters can be connected in series using a modular construction to achieve short-circuit ratings of some 50 kA at 420 kV.

With dead-tank bulk-oil circuit-breakers, current transformers can be placed around the circuit-breaker entry bushings to provide an economical mounting arrangement. With low-oil-content circuit-breakers separate current transformers are required which must contain their own primary insulation. When current transformers are required on both sides of the circuit-breaker the low-oil-content concept can become a more costly arrangement than a dead-tank bulk-oil circuit-breaker with integral current transformers.

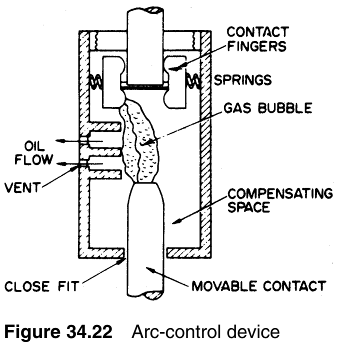


Figure 34.22 Arc-control device

34.1.8.6 SF_6 circuit-breakers

Virtually all current designs of circuit-breaker for use at transmission voltages now use sulphur hexafluoride (SF_6) gas both as an arc-interrupting and a dielectric medium. At distribution voltages SF_6 designs of circuit-breaker are also used, but here the market is still shared with vacuum and bulk-oil circuit-breaker alternatives.

Three basic types of SF_6 interrupter are commonly employed. The gas-blast interrupter, the puffer interrupter and the rotating-arc interrupter.

The gas-blast interrupter tends to have a higher performance capability than the other interrupters and is more commonly applied to transmission circuit-breakers. All gas-blast interrupters cause a flow of pre-pressurised gas across or along the opening circuit-breaker contacts. These interrupters may take a number of forms. Early designs used a store of high-pressure SF_6 gas kept separate from the main SF_6 used for dielectric purposes. On opening of contacts a valve operates to allow the flow of the high-pressure gas around the contacts to extinguish the arc. The gas is then re-compressed following an opening operation. Such a two-pressure system was used on early designs of SF_6 circuit-breaker. However, it has the disadvantage that the high-pressure gas, typically stored at some 15 bar gauge, liquefies at normal ambient temperatures and it is then necessary to apply heaters to the gas in the high pressure storage cylinder to ensure that it is retained in its gaseous state. Heater failure would require the circuit-breaker to be removed from service until the heater could be reinstated. In addition, the provision of a high pressure storage chamber makes the two-pressure circuit-breaker very expensive.

The problem was resolved by the development of the ‘puffer’ circuit-breaker. Here the gas is compressed during the initial part of the opening stroke and prior to separation of the circuit-breaker arcing contacts. This requires the circuit-breaker to have a long operating stroke and a powerful operating mechanism to precompress the gas. Nevertheless, it is this arrangement that is most commonly used on transmission SF_6 circuit-breakers. There are a number of variations of the puffer principle which determine the way in which the gas flows around the opening arcing contacts. These may be referred to as ‘mono-blast’, ‘partial duo-blast’ or ‘duo-blast’. A typical partial duo-blast interrupter is shown in *Figure 34.23*.

Attempts have been made to overcome the necessity for the provision of a large powerful operating mechanism on puffer circuit-breakers by using the heat of the arc itself to pressurise the surrounding gas and induce arc extinction. This principle is commonly referred to as the ‘self-pressurisation interrupter’. Whilst clearance at high values of fault current can readily be achieved, it is usually still necessary to apply a small piston to assist in arc extinction at very low values of fault current. This design of circuit-breaker is commonly used in distribution circuits and is now being employed in circuit-breakers for use at transmission voltages.

A further alternative, widely used at distribution voltages, is the use of the rotating-arc principle. Here the arc is induced to rotate very rapidly under the influence of magnetic fields set up by a series coil inserted into the current

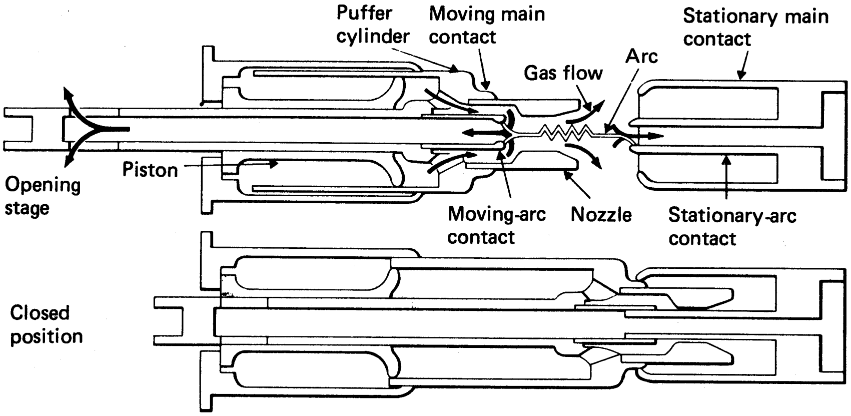


Figure 34.23 A typical partial duo-blast SF_6 interrupter

path during the opening of the circuit-breaker contacts. Very rapid movement of the arc causes a flow of cool SF_6 gas across the arc to achieve arc extinction.

The rotating-arc principle is shown in *Figure 34.24*. With this design more economical operating mechanisms can be achieved but, unlike the puffer circuit-breaker, the arc duration is largely dependent on the size of current being interrupted.

At transmission voltages, SF_6 circuit-breakers mainly use either pneumatic, hydraulic or spring operating mechanisms, whilst at distribution voltages spring operating mechanisms are now invariably used.

With SF_6 circuit-breakers it is imperative to ensure a leak-proof assembly. The number of enclosure joints needs to be kept to an absolute minimum and the main drive for the circuit-breaker moving-contact system should preferably be taken through only one point in the enclosure. Either rotary or axial drives may be used, but special provisions must be made to ensure adequate gas sealing. Maximum gas-leakage rates are typically specified as being not greater than 1% per annum in order to ensure adequate gas retention within the anticipated maintenance periods of the circuit-breaker.

Low-gas-density alarms are usually fitted to give indication of loss of gas and, in the event of rapid loss of gas, circuit-breaker immediate tripping or trip lockout systems are usually employed.

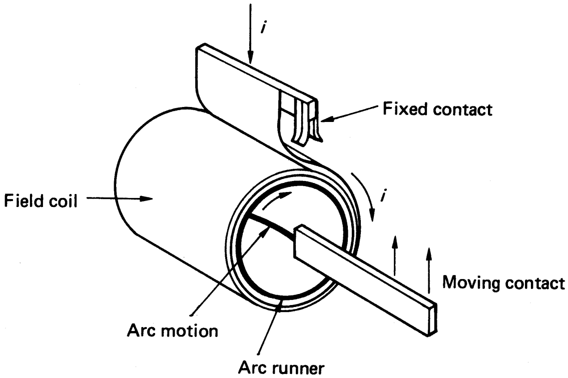


Figure 34.24 One type of SF_6 rotating-arc principle

The development of the SF_6 interrupter has been such that the interrupter capability has increased rapidly and it is now common practice to use only two interrupters per phase for a typical circuit-breaker rating of 55 kA three-phase at 420 kV. A single interrupter can achieve a typical rating of 40 kA at 420 kV.

The transient recovery voltage withstand capabilities are such that parallel resistor interrupters, as used on air-blast circuit-breakers, are not required. All that is necessary is simple capacitive voltage grading across both interrupters to ensure uniform sharing. Comparison with an equivalent air-blast circuit-breaker shows that the number of interrupters at 420 kV has been reduced from 10 to 2, parallel resistor interrupters and high pressure ceramic blast tubes are no longer required because standard porcelain insulators can satisfactorily operate at the SF_6 gas pressures required for interruption. This considerably reduces the number and complexity of components used and has enabled significant cost reductions to be achieved in the application of transmission switchgear.

34.1.8.7 Vacuum circuit-breakers

The possible use of vacuum as an interrupting medium has been studied since the 1920s. The major difficulty has been to produce a sealed device capable of retaining the full vacuum over an anticipated life span of some 25 years. It was not until commercially available high integrity vacuum devices, such as cathode-ray tubes, appeared that suitable techniques for ensuring, adequate sealing evolved. This then opened up the way for vacuum interrupters which became available in the late 1950s. Early designs were costly to produce and could not compete economically with the then widely used oil circuit-breakers. It was not until the late 1960s that mass produced economical designs became readily available.

In a vacuum interrupter, as the vacuum chamber is evacuated and the available ionisable molecules are reduced, the dielectric strength for a given gap rapidly increases. Pressures of some 10^{-8} T are typically used in vacuum interrupters. Pressures up to some 10^{-4} T would be satisfactory, but above this value ionisable atoms become available and the dielectric strength rapidly reduces to give the typical Paschan voltage–vacuum breakdown characteristic. Thus, at the end of the life of a vacuum interrupter, a pressure of some 10^{-4} T is still necessary. A typical contact gap in

11 kV vacuum interrupters would be some 15 mm. A major advantage of a vacuum circuit-breaker is that the very small contact travel required and the low mass of contacts have enabled very economical low output operating mechanisms to be used.

Very careful choice of contact material is necessary for vacuum interrupters. At high-current interruption, metal vapours are produced which assist in arc extinction and, therefore, for long contact life a hard contact material is the ideal choice. However, under low-current interruption conditions, hard contacts produce very little metallic vapour and very rapid arc extinction and 'chopping' of the current waveform results, which can produce high overvoltages. A softer contact material is better for low-current interruption, but erodes too rapidly at high currents. Contact material therefore must be a compromise between these two extremes and contact designs have evolved to ensure that the arc is kept in motion to minimise contact erosion. Important contact material criteria are vapour pressure, electrical conductivity, heat conductivity and melting point. Commonly used contact materials are copper–bismuth or copper–chrome alloys. Since the arc burns in an ionised metal vapour this vapour must condense on arc extinction and a vapour shield is usually provided for this purpose. This prevents the vapour from depositing on the insulating envelope and reducing its dielectric withstand capabilities. The shield also protects the envelope from thermal shock.

At or near zero current the arc extinguishes, vapour production ceases and very rapid re-combination and de-ionisation of the metal vapour occurs. The metal vapour products are deposited on the shield thus ensuring the clean conditions necessary for withstanding transient recovery voltage across the open contacts.

Means must be provided to allow moving-contact travel within the vacuum envelope. This is usually achieved by the provision of a stainless-steel bellows assembly. Design, construction and quality-control checks of the bellows are vitally important to ensure long term vacuum integrity.

Vacuum interrupters, by virtue of their construction, are single-phase devices. They are usually mounted in air in a three-phase circuit-breaker with air and solid insulation between phases. The maximum voltage rating of a vacuum interrupter is some 36 kV and maximum short-circuit currents may, in extreme cases, be as high as 100 kA, with rated currents of up to 4000 A.

The main use of vacuum interrupters is in circuit-breakers for use on distribution systems. More recently, with the higher ratings now achievable and for economic reasons, they are also being used in generating station auxiliary supply applications instead of free-air circuit-breakers. Vacuum circuit-breakers are seldom used at transmission system voltages as their maximum voltage limitation would require a large number of series connected interrupters which would be uneconomical.

The vacuum interrupter can also be used, in a much lighter construction, as a contactor for motor switching applications at voltages of up to 12 kV. A typical vacuum interrupter construction is shown in *Figure 34.25*.

34.2 Materials

34.2.1 Insulating materials

Many insulating materials are used in switchgear, they may be required just to provide insulation or to provide insulation and mechanical support, or to provide insulation and also assist in the arc interruption process. Where this happens some disassociation occurs and chemical recombination may be required.

They must also be capable, in addition, of withstanding temperature variations which may occur in service under both normal load and fault conditions. Under the latter they must withstand the extreme mechanical and electromagnetic forces that might result and sometimes also extreme pressure rises.

With insulation systems exposed to the atmosphere they must also be designed to withstand all anticipated environmental conditions of extreme temperature, rapid temperature changes, ice, snow, sun, wind, severe rain, solar radiation and lightning. All insulation systems must be designed to withstand transient overvoltages associated with lightning, switching surges and power frequency overvoltages. In some cases they may also experience d.c. trapped charges on circuit de-energisation or possibly d.c. with superimposed a.c. voltage.

Insulation systems must thus withstand many varied criteria and often with simultaneous combined phenomena, in addition, they must perform satisfactorily for many years with minimum of maintenance.

For outdoor insulation of overhead lines at transmission voltages for example, glass insulators have been commonly used, glazed porcelain insulators are also occasionally used. Whilst porcelain insulators are somewhat more expensive

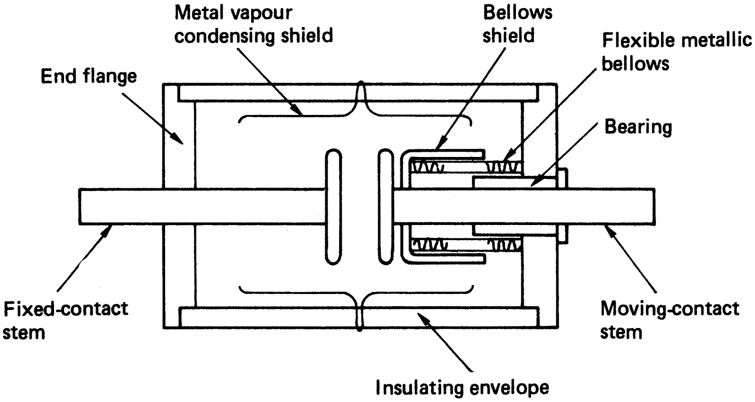


Figure 34.25 Cross-section of a vacuum interrupter

than glass insulators they generally show improved performance under polluted conditions. Porcelain insulators are widely used for distribution overhead lines and have almost universally been used for exposed terminations of substation equipment. In more recent years polymeric overhead line insulators have been used, these usually being in the form of EPDM or silicone rubber. Their performance is somewhat more complex than both porcelain and glass and very careful design is necessary. In addition, the composition of polymers is very critical and minor changes of constituents can cause significant performance changes.

Epoxy resin polymers have occasionally been used for outdoor insulation although these tend to deteriorate from solar radiation. Epoxy resin insulators are widely used in switchgear applications not exposed to external atmospheric conditions. These resins can be readily moulded to form complex insulator shapes which may not be possible with many other insulating materials. Such insulators have both exceptionally good dielectric and mechanical strength. Epoxy resin insulators will be encountered in most types of distribution switchgear and in gas insulated switchgear at transmission voltages.

A more economic form of insulation which is widely used in low-voltage switchgear is commonly referred to as dough moulding. This is a mixed polymer that rapidly sets at high temperature and pressure and is usually impregnated with numerous glass fibre strands to provide increased mechanical strength. Its mechanical and dielectric performance is usually inferior to epoxy resin systems and it is rarely found in distribution or transmission switchgear.

A further form of insulator which has been used for many years in switchgear is known as the ‘Synthetic Resin Bonded Paper’ insulator, often referred to as SRBP. This can only be used for straight insulators as may be required for example for circuit breaker bushings. The basic constituent is a thin, flat, electrical quality paper, commonly known as ‘Kraft’ paper. The paper, on a roll, is cut to the required length and then wound onto a mandrel, at the same time, a special electrical grade varnish is allowed to run in between the layers of paper. The mandrel may either be a metallic tube or in the form of a copper rod. Stress control throughout the section of insulator is important, this is usually achieved by winding in very thin aluminum foils known as stress control shields.The length of these shields tends to reduce from the HV conductor to the external earthed electrode to ensure uniform stressing both across the insulator section and along it’s length. On completion of winding the assembly is placed in an oven, often under vacuum, to allow the varnish to cure. The final insulator is mechanically very strong. Such a form of insulator is commonly used where an HV conductor has to pass through an earthed electrode i.e. a circuit breaker tank. The atmospheric side of the insulator is usually enclosed in a porcelain housing to protect it from the environment. The annular gap between the SRBP insulator and inner porcelain wall is filled with high quality insulating oil.

A somewhat similar form of insulator is referred to as the ‘Oil Impregnated Paper’ insulator, commonly known as OIP. This form of insulator is usually utilised where more complex insulation profiles may be required. The same type of paper as used for SRBP insulators is employed, but in a narrower strip form, such that it may be wound either by machine or by hand, over a complex profile. In some cases a crepe paper is used to give greater flexibility where even more complex profiles are required. Aluminum foils are also embedded for stress control purposes. On completion of winding the insulated assembly is placed in a vacuum chamber with gradual increasing temperature to remove both moisture and air which will be trapped within the winding. Finally, high grade insulating oil is slowly introduced into the chamber to fully impregnate the paper. Once impregnated in its final assembly the oil should not be removed otherwise moisture and air might be re-introduced and re-processing would again be required. OIP insulation is commonly used as the insulation system for transmission system current and electromagnetic voltage transformers.

Where insulated screens may be required, for example between the phases in air insulated distribution switchgear, these may be of either bakelised paper form, made in flat rigid sheets in a similar manner as described for SRBP bushings, or in the compressed fibre board form which comprises bonded layers of oil impregnated fibrous paper. Compressed fibre board has the advantage that it can be bent to form enclosures. Such bent assemblies of compressed fibre board are usually found, as insulating barriers, within the tanks of bulk oil circuit breakers.

A further form of solid insulating material which is commonly used for insulated mechanical drive rods within switchgear is compressed and impregnated plywood, the impregnation again being with a high grade electrical varnish. This material has a high breaking strength under both tensile and compressive loading.

For transmission circuit breakers where a much higher mechanical loading is required the drive rods usually take the form of wound, epoxy impregnated, fibreglass matting, either wound as a tube or a rod. Metallic end fittings can usually be attached either by crimping or by epoxy bonding.

For cable terminations to distribution switchgear bitumen compound has been used for many years to provide the main insulation between phases and to earth. This material has a low affinity to moisture and becomes fluid at temperatures of around 120°C such that when heated it can readily be poured into the cable box and sets hard on cooling. Care is necessary however to use the appropriate grade of compound as shrinkage, embrittlement and cracking may occur to produce voids which may latter result in dielectric breakdown. Bitumen compound is also commonly found in current transformer chambers of distribution switchgear.

In more recent years, heat shrinkable polymers have been used for insulation of cable terminations. The application process is somewhat critical as entrapped air can lead to dielectric failure. Correct stress control is also vitally important. As cable jointing is, by necessity, a site function, it is important that site personnel are appropriately trained in the techniques and that appropriate Quality Control procedures are applied.

Insulating mineral oil has been used for many years as a dielectric media and also as an arc interrupting media in oil circuit breakers. It is usually of a Napthenic or Parafinic base dependent upon its source. Both have been used successfully in switchgear but in more recent years parafinic based oils have been more widely used due to, limitations of supply of napthanic oils.

One disadvantage of insulating oils is that they absorb moisture which can lead to dielectric degradation. Hence it is necessary to ensure that moisture entry is severely limited by equipment design. A further major disadvantage is that oils are flammable and, dependent upon class, have flash points varying between 95°C and 140°C. When disassociated due to arcing, hydrogen is formed and under equipment failure conditions the hydrogen can ignite to result in severe explosion and subsequent oil fire. Whilst the failure rate of oil filled switchgear is extremely low, the number of equipment’s in service is very large (some hundreds of thousands of equipment's in the UK alone). Unfortunately failures and explosions do occasionally occur. It is for this reason that the application of oil filled switchgear has significantly diminished in recent years in favour of SF_6 insulated equipments.

In order to eliminate fire risks, synthetic insulating oils have been previously used, these were generically referred to as 'Askerals', these were based on polychlorinated biphenyls (PCBs). These have no flash point and are considered non-flammable. However, they are hardly bio-degradable and accumulate along the food chain, for this reason their production was banned in Europe in 1985. Strict regulations now apply to their use and most have in fact now been eliminated from service. Over the last 20 years considerable work has ensued on the production of less flammable liquids, the main alternatives are silicone and ester liquids, these are expensive and other performance parameters are limited. They are not generally used in switchgear.

Over the last 30 years or so the use of sulfur hexafluoride (SF_6) as an insulating and interrupting medium has slowly increased whereby today it has virtually replaced the use of oil in modern designs of switchgear.

Pure SF_6 is odourless and non-toxic but will not support life. It is 4.7 times heavier than air and tends to accumulate in low areas with the possibility of causing drowning. It has unique features which are particularly suited to its application in switchgear. It has a high electron attachment coefficient which allows it to have a high dielectric withstand characteristic. Its alternating voltage withstand characteristic at 0.9 bar(g) is comparable with that of insulating oil. A further advantage is that its arc voltage characteristic is low, hence the arc energy removal requirements are low. When subjected to the extremely high temperatures associated with arcing it tends to fragment into numerous constituent gases some of which are toxic, however, these recombine very quickly as the temperature falls and dielectric strength rapidly recovers. This process can occur in microseconds to allow several fault interruptions in quick succession. Solid arc products may be produced, these comprise mainly metal fluorides and sulphides, these are acidic and must not be inhaled. The gas performance is not impaired by the presence of these solids provided that it is maintained in a very dry state. At typical operating pressures of 4.5 bar(g) at 20°C SF_6 will remain in its gaseous form down to about -40°C, below this temperature it will liquefy and the equipment's performance may be impaired.

SF_6 is a very stable gas and if released into the atmosphere it may take over 3000 years to dissociate. It also has a high reflective index in the infrared spectrum which means it contributes to greenhouse warming. In fact it is the worst known greenhouse gas but at present the quantity in the outer atmosphere is small and its contribution to greenhouse warming is negligible. Nevertheless, it must no longer be deliberately released into the atmosphere and leakage and handling losses must be minimised. Contaminated gas can readily be re-processed to restore it to a state whereby it can safely be reused. It can also be safely disposed of at the end of its useful life.

34.2.2 Contact materials

Low-voltage high current contacts are used in LV distribution switchgear. They are usually of the bolted form or are intended to open and close to provide either circuit isolation, load interruption, motor switching duties or fault interruption.

Bolted contacts used to connect bus-bars for example are usually of silver or tin plated copper. The silver or tin plating being used to inhibit long term oxidation of pure copper contacts. Usually more than one fixing device is employed to minimise the possibility of a loose contact which could lead to over-heating or failure. The contact must also be capable of withstanding short circuit through current in the event of a system fault. A typical fault withstand capability being 43 kA for three seconds.

Disconnectors may be provided to facilitate a circuit open point, these are switched with the load removed such that their contacts have only to interrupt a few milliamperes. They are infrequently switched and copper contacts are usually silver plated to prevent oxidation of the copper surfaces. Switches or switch disconnector contacts have to break load current and may have to close against the system short circuit current. Such switches may be operated relatively frequently and contacts are generally unplated as the wiping action during closing or opening generally cleans the contacts and tends to remove any oxide formation. When fault closing is required pure copper contacts may well weld due to the softness of the copper. Such contacts usually contain a hardening additive within the copper, typically additives are beryllium or tungsten. For low-voltage circuit breaker applications where high fault current breaking is required the point of arc root during interruption must be of a very hard material and typically tungsten tips are welded to the ends of the contact fingers. It is important that the arc transfers to the contact tip and does not root on the normal current carrying surface of the contacts, failure would rapidly result if this should occur, i.e. the normal current carrying surfaces of the closed contacts should not be eroded as a result of fault interruption.

Circuit breakers at distribution voltages usually employ copper contacts with a hardening agent to prevent either cold welding in the normally closed position and high current welding resulting from the passage of through fault current. Copper/tungsten contact tips are invariably employed for arc rooting during short circuit interruption. Such contacts may be of the rod and multi-finger type, wedge and multi-parallel finger type or butt type contact.

Distribution switchgear bus-bar contacts are usually of multi-fixing point silver plated copper joints. Sometimes however, for lower rated current circuits, aluminum bus-bars may be employed. Jointing aluminum however requires more care than for copper as aluminum readily oxidises and it is necessary to wire brush joint surfaces with a stainless steel wire brush with a layer of grease applied to prevent re-oxidation. The joint is usually made with the application of a fine layer of jointing grease. Where aluminum bus-bars are used in outdoor substations it is necessary also to apply a proprietary jointing compound to inhibit moisture ingress.

For transmission circuit breakers similar contact materials are used but the copper contact fingers may have an additional additive, e.g. molybdenum, to provide spring resilience. In view of the generally higher current ratings of transmission circuit breakers cylindrical type contacts are often used to minimise the skin effects induced by the very high current.

For gas insulated switchgear (GIS) aluminum tubular bus-bars are invariably used. This is to provide a large diameter to minimise the surface dielectric stresses. Copper surfaces may be embedded onto the aluminum tube where bus-bars are jointed.

34.3 Primary-circuit-protection devices

Primary-circuit-protection devices are necessary to ensure that any system malfunction is quickly detected and steps taken to minimise its consequences. The main primary-circuit-protection devices encountered are current transformers, voltage transformers and surge arresters.

34.3.1 Current transformers

A current transformer is a current transducer that will give a current signal directly proportional in magnitude and phase to the current flowing in the primary circuit. It also has another very important function in that the signal it produces must be at earth potential relative to the HV conductor. The primary circuit of the current transformer must be insulated to the same level of integrity as the primary insulation of the system. For current transformers used on HV systems the primary-circuit insulation represents a very large proportion of the cost of the transformer.

The current transformer is the only current transducer widely used in HV networks. Recent developments of fibre-optic HV current transducers show promise but high cost and questionable reliability have limited their application. There is little doubt, however, that future current transducers will use fibre-optic technology.

A current transformer, as its name implies, is a transformer. It is almost invariably in the form of a ring-type core around which is wound a secondary winding.

The primary winding usually consists of a straight bar through the centre of the core which forms one turn of the primary winding. For low primary currents, typically below 100 A, multiturn primary windings consisting of two or more turns may be used in order to achieve sufficient ampere-turns output to operate the secondary connected equipment. For use at distribution voltages the core and secondary winding, together with the secondary terminations, are usually placed over a straight HV conductor bushing insulator which forms the segregation between the HV conductor and earth. An earthed screen is usually provided on the outer surface of the bushing and the current transformers are placed over this earth screen to ensure the limitation of HV partial discharge activity in the air gap between the bushing and the current-transformer winding. Current transformer secondary windings are generally connected to electromagnetic relays. These tend to require a high operating input which necessitates high-output (typically 15 V-A) current transformers. More modern protection is of the solid-state form and requires a much lower operating signal, thus enabling current-transformer designs and costs to be reduced. The secondary windings of current transformers tend to be rated at either 1 or 5 A, although other ratings are used at times.

Where long secondary connections are required between the transformer and the relay, 1 A secondary windings are advantageous in reducing the lead burden. Cold-rolled silicon-iron is usually used as the core material for protective current transformers but, where high-accuracy metering requirement is necessary, a very high grade alloy steel is used which is commonly referred to as 'Mumetal'.

For use at higher transmission voltages it is necessary to build integral insulation into the current transformer between HV conductors and secondary windings. This insulation is nearly always in the form of oil-impregnated paper, although SF_6 gas is occasionally used. The cost of providing the pressurised SF_6 gas enclosure usually makes SF_6 insulated current transformers uneconomical.

There are two basic forms of construction of transmission voltage oil-impregnated-paper insulated current transformers: the live-tank and dead-tank forms.

In the live-tank form the core and winding are placed at the same level as the primary conductor which passes through the centre of the assembly. The core and windings clearly need to be at earth potential. They are usually enclosed in some form of metallic housing which has a long vertical metallic tube through which the secondary winding leads pass to the base level. This housing and vertical metallic tube then have very many layers of paper wrapped around them to form the main primary insulation. Aluminium foil stress-control layers are wound in between the paper layers to ensure a uniform stress distribution from earth potential at the bottom end of the assembly to line potential at the top end.

The insulated current transformer assembly is then placed within an insulator housing having a metallic top assembly through which a primary conductor is passed. This conductor is electrically connected to the top assembly on one side and insulated on the other to prevent a current transformer short-circuited turn.

Before assembling the top cover, the whole transformer assembly is placed under vacuum for several days to ensure thorough extraction of moisture from the paper. The assembly is then filled under vacuum with a high grade insulating oil to prevent the formation of air bubbles. After filling the transformer to the top it is sealed. Some form of expansion assembly is incorporated to allow for expansion and contraction of the oil within its sealed compartment. This may comprise a bellows assembly or sealed nitrogen cushion. The current transformer may also incorporate an oil-level indicator to allow checking for loss of oil and a gas-detection system to allow monitoring for the production of gaseous products resulting from partial dielectric breakdown.

In the dead-tank version the current transformer core and windings are placed at the bottom, earth, end of the assembly and the insulation between the primary and secondary is in this case placed around the HV primary conductor rather than core and winding assembly. The centre section of the insulated HV primary conductor at which the core and windings are placed must be at earth potential. The HV primary conductor insulation must be graded on either side of the core and windings. Aluminium-foil wraps are inserted between the paper layers to provide the necessary grading from earth potential at the centre section to line potential at either end. To enable the HV primary conductor assembly to be accommodated in a vertical insulator, the assembly is bent in a 'hairpin' fashion. The insulated paper is in fact wound on to a conductor already formed to this hairpin shape. The legs of this insulated assembly are then opened up to allow the core and windings to be slipped over.

The completed assembly is vacuum processed and oil filled in a similar manner to that described for the live-tank current transformer.

Both live-tank and dead-tank forms of construction are very widely used. Both constructions are shown in *Figure 34.26*.

34.3.2 Voltage transformers

Voltage transformers are also, as the name implies, voltage transducers giving an accurate representation in magnitude and phase of the voltage of the primary conductors. They also require insulation to segregate the primary and secondary circuits. At transmission voltages they are always single-phase assemblies, whereas at distribution voltages they may be three-phase or single-phase assemblies. At distribution voltages the primary winding is always star connected with its neutral point generally insulated and unearthed.

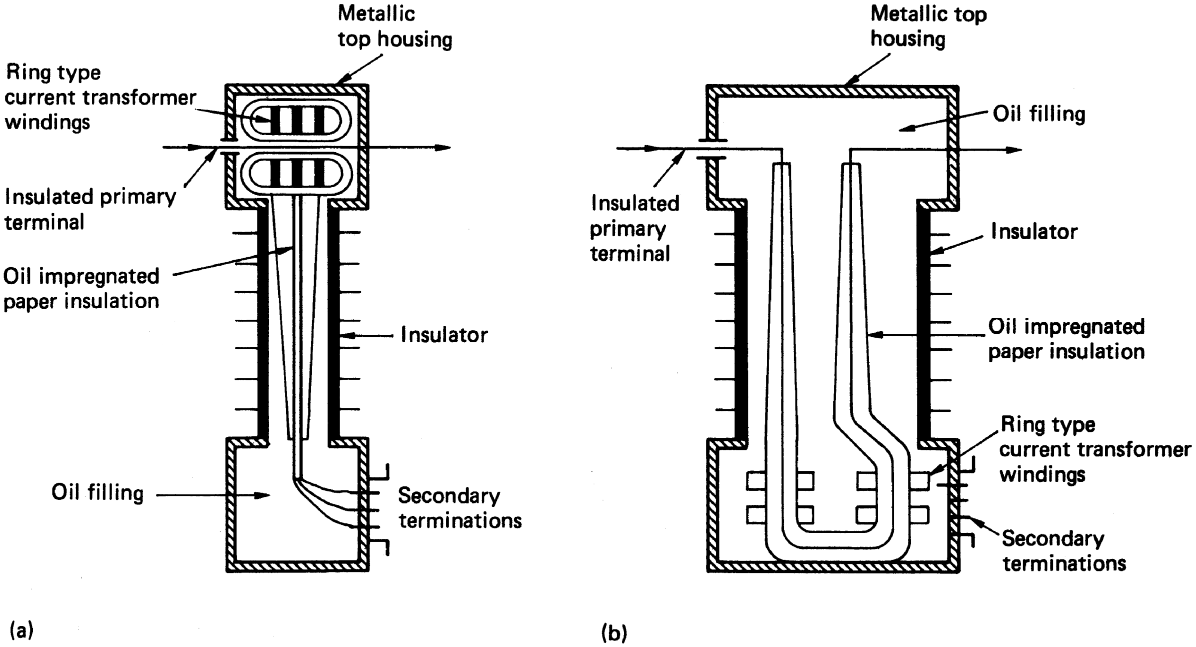

Metallic top housing
Ring type current transformer windings
Insulated primary terminal
Oil impregnated paper insulation
Insulator
Oil filling
Secondary terminations
(a)
Metallic top housing
Oil filling
Insulated primary terminal
Insulator
Oil impregnated paper insulation
Ring type current transformer windings
Secondary terminations
(b)

Figure 34.26 Cross-section of (a) live-tank and (b) dead-tank current transformers

The secondary windings are usually connected in a star arrangement to provide a standard phase-to-phase secondary voltage of 110 V, the individual phase secondary windings being rated at 63.5 V.

For certain applications a further secondary winding is sometimes provided which is connected in open delta. Under normal balanced voltage conditions the voltage across the open delta is zero. However, for earth-fault conditions balance will no longer be achieved and a voltage will occur across the open-delta winding. A three-phase residual voltage transformer uses a five-limb core to allow the zero sequence resultant flux to circulate. Such a system is generally used for residual earth-fault protection.

Three-phase assemblies at distribution voltage traditionally were of the oil-filled construction. More recent designs of three-phase voltage transformer employ a dry type cast epoxy resin insulated assembly enclosed in a metallic tank. Three single-phase cast epoxy insulated voltage transformers may also be mounted within an enclosure and connected to form a three-phase assembly.

Designs of electromagnetic voltage transformer usually employ an earthed electrical shield between the HV and the secondary winding. In the event of an HV breakdown, fault currents will flow to earth via the shield rather than through the secondary winding. A further useful feature of the shield is that it prevents h.f. coupling between primary and secondary windings as h.f. coupling in secondary windings and circuits could cause maloperation of secondary connected equipment. Such high frequencies may, for example, be generated by disconnector operation.

At distribution voltages, voltage transformer HV windings are usually protected by means of fuses to limit the energy that is fed into a faulted voltage transformer to prevent possible external disruption of it.

At transmission voltages, early designs of voltage transformers were also of the electromagnetic construction with the core and winding built in a conventional transformer manner and enclosed within a large oil-filled tank at earth potential. The HV connection was taken into the transformer via a large external bushing. Whilst some of these designs are still in service and are reliable, in the rare event of failure a large amount of energy would be expended within the transformer tank.

Since it is impracticable to protect such voltage transformers with fuses, to alleviate the problem a gas-detection device is usually fitted which will trip the associated circuit-breaker in the event of rapid build up of gas.

Electromagnetic voltage transformers for use at transmission voltages are also very costly and a more inherently reliable and economical device was developed which is referred to as a ‘capacitor-voltage transformer’. This concept uses one or more HV capacitor assemblies, each of which is enclosed within its own porcelain housing. The individual units are mounted on top of each other to form a series assembly of HV capacitors. The complete assembly is usually mounted on an earthed tank which encloses a further capacitor and an electromagnetic transformer/reactor assembly. The bottom capacitor forms the lower leg of a capacitor divider assembly and a voltage signal (typically 12–25 kV) is taken from the interface of the HV and LV capacitors. This signal is then fed via a reactor to a transformer which has a secondary winding giving an output of 63.5 V at the system rated voltage. The secondary winding may have tappings to enable correct setting and thus minimise the voltage-ratio error.

The reactor also usually has tappings to allow for minimum setting of the phase-angle error. The LV capacitor is usually of a large value, typically 20 000 or 30 000 pF. The value of the reactance connected to this capacitor is such as to allow resonance to occur at the required system frequency, e.g. 50 or 60 Hz. This arrangement allows a significantly larger secondary output to be taken from the capacitor-voltage transformer than would be available if it were purely a capacitor divider. Ratings of 200 V-A can readily be achieved by this means. Capacitor-voltage transformers are tuned devices and their accuracy and output will fall considerably at frequencies other than the tuned frequency.

Maloperation of a capacitor-voltage transformer is usually detected by variation in secondary output voltage and explosive failures are extremely rare. Unless a very high value of HV capacitor is used, external moisture or pollution contamination on the insulation can cause variations in the HV field around the insulator which will result in slight variations in output. These variations are of little significance for capacitor-voltage transformers used for protective purposes, but for high accuracy tariff metering such variations can be significant.

For these reasons capacitor-voltage transformers are seldom suitable for use in high accuracy tariff metering applications and it has been necessary to revert to the electromagnetic transformer for these applications. Early designs used a cascade arrangement of separately connected transformers in order to reduce physical size and to alleviate the consequences of dielectric breakdown. These arrangements have been used at both 420 and 300 kV and comprise three separate electromagnetic units vertically mounted on top of each other and enclosed within a large-diameter oil-filled insulator assembly. Cascade electromagnetic voltage transformers have proved to be very reliable, but are also extremely costly. Where large numbers of high-accuracy voltage transformers are required, for example at interconnections between adjacent electricity supply companies, then the cascade voltage transformer is no longer economical and it has been necessary to revert to a single electromagnetic unit enclosed within its own porcelain housing.

Because of the considerable reduction in output achieved by the use of solid state secondary equipment, smaller electromagnetic units can now be used. These designs nevertheless still comprise paper-insulated oil-filled units and are prone to the dielectric-breakdown problems of the earlier designs. Very careful design and strict manufacturing quality-control procedures are necessary to ensure the long-term integrity required of these equipments. It is still normal practice for gas-detection alarms to be provided which may also offer the facility for circuit-breaker tripping on sudden build up of gas.

34.3.3 Combined-instrument transformers

Where high-accuracy instrument transformers are required to be fitted retrospectively, space limitations in existing substations often present significant problems. The situation has been alleviated by the combination of both a current transformer and an electromagnetic voltage transformer into a common insulator housing. These devices are referred to as 'combined instrument transformers'. They are used at transmission voltages up to some 300 kV, but for higher voltages where very limited numbers would be required they are not economically viable and separate units are still used.

Figure 34.27 shows typical arrangements of a capacitor-voltage transformer, cascade electromagnetic voltage transformer and combined-instrument transformer.

34.3.4 Surge arresters

Electricity supply networks extensively use overhead lines to transmit the power from generating stations to major load centres. These allow interconnection of different parts of the system and, in rural networks, supply isolated consumers. All of these overhead-line circuits are prone to atmospheric disturbances. These may be physical in nature (i.e. wind, rain, snow, ice and pollution) or electrical (as may occur during lightning activity). Should an overhead-line circuit be struck by lightning the overvoltage surge generated on the line may be of the order of megavolts. This surge will travel down the line, possibly causing flashovers en route, and its magnitude and shape will be attenuated by the electrical parameters of the line. When it arrives at the terminations of electrical equipment it may cause the equipment to flashover externally or fail internally. The surge will be triangular in wave shape having a steep wavefront with a rate of rise of possibly up to 3 MV/μs and a long duration tail of typically 20–200 μs.

In addition, transient overvoltages can be generated by the switching devices themselves. These will also be triangular in shape but with a much less steep wavefront (250 μs) and much longer wave tail (2500 μs) than the lightning surge. Switching surges can also result in dielectric failure of connected electrical equipment. It is thus necessary to

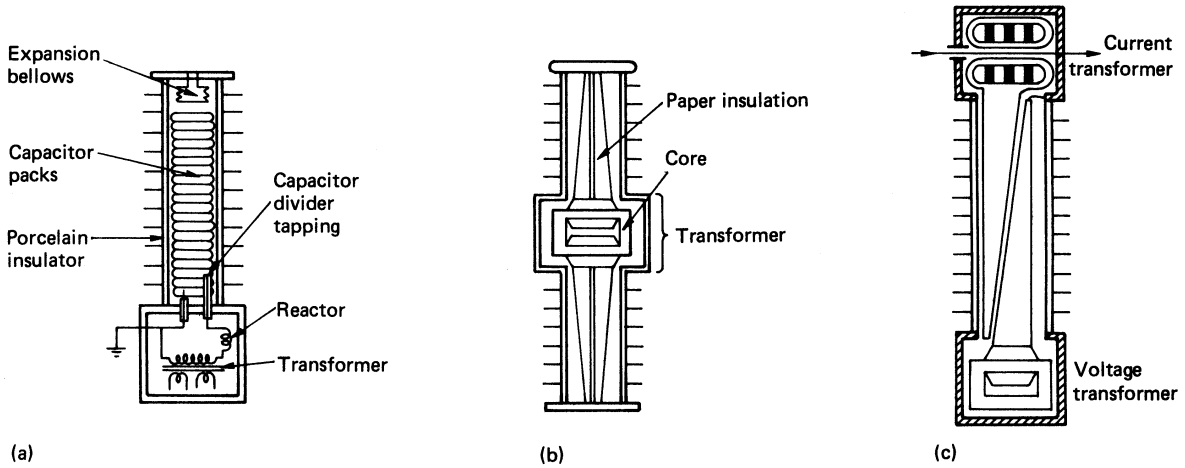


Figure 34.27 Arrangement of (a) capacitor-voltage transformer, (b) cascade electromagnetic voltage transformer, and (c) combined-instrument transformer

protect electrical equipment against these transient overvoltages. This is normally achieved by application of a surge diverter or, more correctly, surge arrester. The surge arrester is a device which is electrically connected between the conductor and earth and in close proximity to the equipment that it is to protect. Under normal power frequency conditions the current which flows to earth is negligible, but under transient overvoltage conditions it will detect the change in voltage magnitude and divert the surge to earth. Surge currents may be as high as 100 kA or more and of a few tens of microseconds in duration. Once the surge has passed, the surge arrester will very quickly reinstate itself to immediately withstand the power frequency system voltage and prevent power frequency fault current flow. The duration of the passage of surge current is so short that the system protection does not operate to cause circuit-breaker tripping.

Rod gaps are sometimes provided across insulators, or electrical apparatus also to divert the surge to earth but, in doing so, cause power frequency fault current to flow and result in circuit-breaker tripping.

Surge arresters provide the required characteristics by the use of non-linear resistor blocks which have a high resistance at LV and a very low resistance at HV Traditionally, silicon carbide blocks have been used as being the most suitable non-linear resistor material. However, they suffer from the disadvantage that even at LV a significant current will still flow which will lead to overheating and eventual failure. This problem was resolved by incorporating a number of series gaps in the arrester. The capacitance of these gaps causes insignificant current to flow through the blocks under normal voltage conditions and thus alleviates overheating problems. For transient overvoltages, the series gaps will break down and the resistor blocks will effectively be switched into circuit. These gaps will quickly reinstate after the passage of surge current.

Many lightning strikes comprise multiple strokes which occur within a few milliseconds of each other. For example, a low-probability strike might comprise six strokes within 100 ms. The application of each transient overvoltage to the terminals of the surge arrester will require the dissipation of a large amount of energy which will result in heating of the silicon carbide blocks. The arrester, however, must be capable of dissipating multiple lightning strokes without deterioration and the resistor blocks must be sized accordingly. The disadvantage of the silicon carbide gapped arrester is that, in order to overcome the high leakage current problems, series gaps are required and these make the arrester large and costly. Nevertheless, these arresters can readily withstand temporary power frequency overvoltages for long periods of time since the series gaps are unlikely to operate.

The problem of size, complexity and cost of the silicon carbide gapped arrester has been resolved in recent years by the development of the zinc oxide arrester. These arresters use resistor blocks manufactured from zinc oxide and a combination of a number of other additives and have the advantage that their resistance is such that, at normal operating voltages, only a milliampere or so of leakage current will flow. They have a sharp knee point at higher voltages which rapidly allows the flow of high current resulting in the very rapid dissipation of the surge.

A comparison of the characteristics of silicon carbide and zinc oxide resistor blocks is given in *Figure 34.28*.

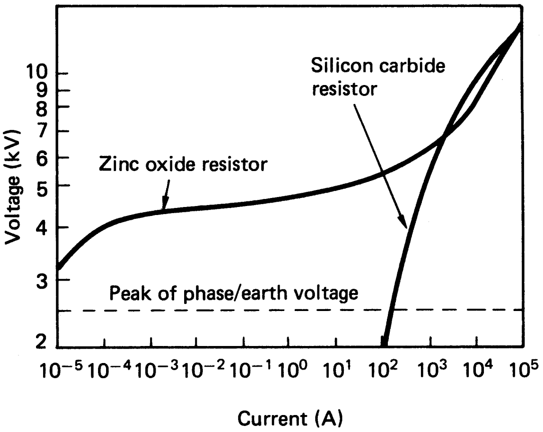


Figure 34.28 Comparison of (a) silicon carbide and (b) zinc oxide resistor characteristics

Since the leakage current for zinc oxide blocks is very low, series gaps are no longer required. In addition, the zinc oxide blocks are physically smaller for a given rating. This technology has allowed the rapid development of very economical high-performance designs of surge arrester and the virtual elimination of silicon carbide designs.

A further innovation has been the recent introduction of a polymeric housed, rather than a porcelain housed, arrester. This has significantly increased the mechanical strength of the arrester and elimination of the problem of shattered porcelain in the event of arrester internal failure. A typical porcelain housed silicon carbide gapped arrester and a polymeric house of zinc oxide arrester construction for the same voltage rating are shown in *Figure 34.29*.

34.4 LV switchgear

34.4.1 Fuse cut-outs

'Fuse cut-out' is a term commonly applied to the equipment placed at the interface between the supply authority system and a domestic consumer's system. It comprises facilities for terminating the main incoming cable, for incorporating a fuse link and for terminating outgoing conductors to the supply authority's meter. Since the consumer may have physical access to this equipment, safety is of paramount importance. All live connections must be enclosed to prevent danger and fuse holders containing the fuse links are sealed within the cut-out assembly.

Many thousands of these units are installed by the supply authorities and reliability is of paramount importance. Elaborate specifications and extensive type testing are necessary to ensure the suitability of the device. The cut-out must be rated for the consumer's load which is normally supplied from a single-phase circuit typically having current ratings of up to 100 A.

For higher loads three-phase assemblies must be supplied. The consumer's load can vary rapidly over short periods of time and the cut-out must be designed and tested under a very large number of repetitive cyclic loading conditions in order to ensure that long-term deterioration is unlikely to occur.

The fuse cut-out must also be capable of clearing a short-circuit across its terminals. This may be of the order of 16 kA at 240 V. The main purpose of the cut-out is to protect the supply authority's circuits in the event of malfunction of the consumer's installation. However, in performing this function it will also limit the energy fed into

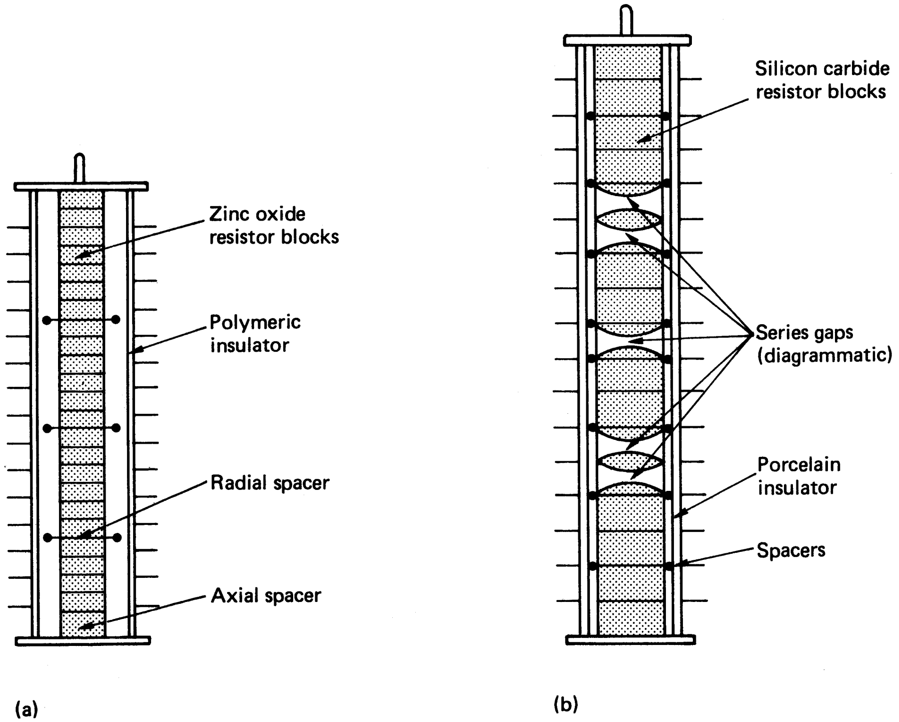

Zinc oxide resistor blocks
Polymeric insulator
Radial spacer
Axial spacer
(a)
Silicon carbide resistor blocks
Series gaps (diagrammatic)
Porcelain insulator
Spacers
(b)

Figure 34.29 Construction of silicon carbide and zinc oxide surge arresters

a consumer's faulted equipment. Removal of the cut-out fuse links will provide physical isolation between the supply authority's and consumer's equipment to allow work to be safely undertaken on the consumer's installation. As the cut-out fuse holders are sealed within their base their removal can only be undertaken by the supply authority.

A typical three-phase domestic fuse cut-out is shown in *Figure 34.30.*

34.4.2 LV fuse cabinets

For commercial or light-industrial consumers the connected loads are in excess of those that can be provided by the domestic fuse cut-out. Three-phase loads are generally required with possible current ratings of up to 1600 A at 415 V. Multiple circuit loads may also be required. These are usually provided by means of a LV fuse cabinet supplied directly from a HV transformer having ratings typically of up to 1000 kVA.

Fuse cabinets may be bolted directly to the LV terminals of the transformer. They comprise an incoming directly connected circuit from the transformer to three single-phase manually operated disconnectors. These disconnectors then control and supply each phase of a main three-phase bus-bar system from which up to five outgoing circuits can be taken. Each phase of the outgoing circuit is protected by a bus-bar-mounted LV fuse with terminations for the outgoing circuit cables provided on the outgoing fixed contacts on the fuse.

Fuse cabinets are generally of weatherproof construction suitable for installation outdoors. They are incorporated within the supply authority's substations and only authorised personnel can gain access. Since live manual fuse replacement may be necessary special training and

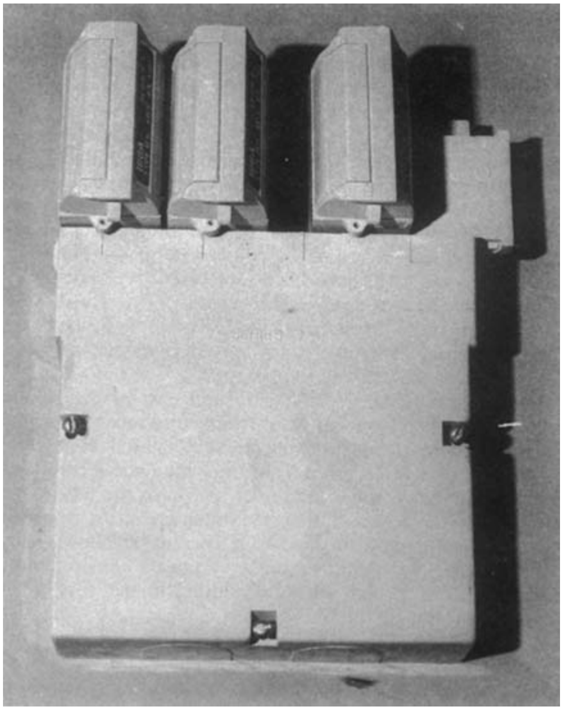


Figure 34.30 Domestic fuse cut-out

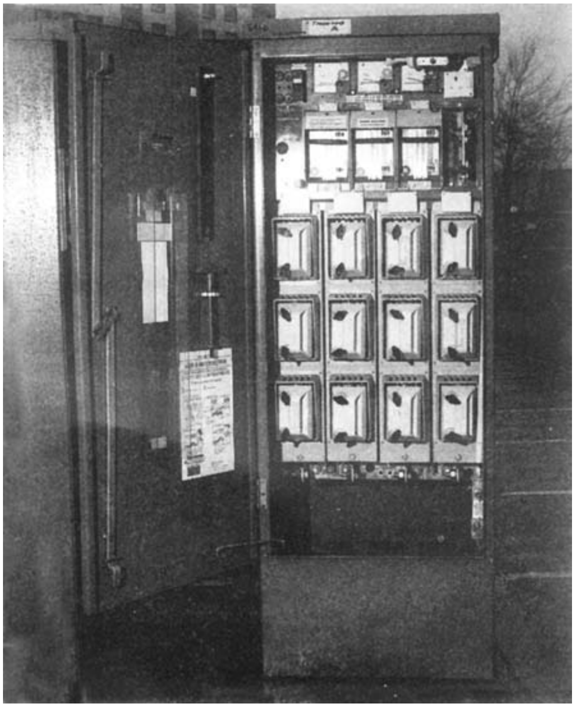


Figure 34.31 LV fuse cabinet

equipment is provided for authorised operators. Short-circuit ratings may be up to 43 kA at 415 V three-phase.

Figure 34.31 shows a typical LV fuse cabinet.

34.4.3 LV switchboards

For LV supplies to large commercial and industrial consumers a LV switchboard is often used whereby the incoming circuit-breaker is owned by the supply authority and is connected by cables to the LV of a suitable transformer. The incoming circuit-breaker will then feed on to a bus-bar system from which the load circuits can be fed. The load circuits may be protected by moulded-case circuit-breakers or fuse switches. The equipments are generally built into a metal-enclosed housing which provides glanding for terminating outgoing circuit cables. Similar switchboards are also used for LV distribution applications in large industrial complexes.

Various configurations of switchgear can be applied depending on the equipment importance. For example, all bus-bars, bus-bar tee-off connections, and circuit-protecting equipments may be enclosed in a common metal-enclosed housing. This arrangement has the disadvantage that any one faulted component within the switchboard will allow the fault to spread to other equipments within the switchboard such that for a major fault the complete switchboard could be rendered inoperative. For high-integrity switchboards, where reliability of supply is of utmost importance, metallic segregation is provided whereby the incoming circuit-breaker will be in its own metal-enclosed compartment as will the bus-bars and bus-bar feeder circuits, the outgoing feeder equipment and the cable terminations, thus preventing the spread of a fault from any one compartment to adjacent compartments. This latter construction is normally applied to high integrity industrial equipments such as may be used, for example, within generating stations. The switchboard can also incorporate combinations of fuse switches and contactors to form motor starter assemblies. A high-integrity LV switchboard is shown diagrammatically in *Figure 34.32*.

34.5 HV secondary distribution switchgear

34.5.1 Urban networks

Supplies to large numbers of consumers in urban conurbations are usually taken from a primary switchboard at between 10 and 20 kV. The switchboard usually comprises two sections of bus-bar with a central, normally open, bus-section circuit-breaker. The supplies are taken from a circuit-breaker on one section of the bus-bar via a ring cable circuit to a circuit-breaker on the adjacent section of bus-bar. The ring circuit may be some kilometres in length and secondary distribution substations can be situated at various points around the ring. The ring may be normally run open at some suitable point along its length.

The secondary distribution substation comprises a HV cable connected switchgear unit normally referred to as a ring-main unit, a HV/LV transformer having typical ratings up to 1000 kVA and a LV fuse cabinet of the type described in Section 34.3.2. The ring-main unit may be physically mounted on the HV terminals of the transformer with the LV fuse cabinet mounted on its LV terminals to form a complete transportable secondary distribution substation. Alternatively, the three items of equipment may be separately mounted and all cable connected. Secondary distribution substations are normally of weatherproof construction and used in outdoor environments. Such a substation is usually protected from the public by means of suitable fencing. The equipments may also be housed in a weatherproof, vandal-resistant housing in which case individual items of equipment may not be weatherproofed.

Very many thousands of secondary substations of the type described are in service and have, over a long period of time, given economical and reliable performance. The ring-main unit usually comprises two manually operated ring switches and a centrally disposed tee-off fuse switch, also manually operated. In addition to the ring switch, fault-making earth switches and suitable cable test facilities may be built into the cable side of each switch to allow the cable to be safely earthed, repaired if necessary, and re-tested. The ring switches, earth switches and test facilities may incorporate interlocking and padlocking facilities to prevent an operator from undertaking an incorrect switching sequence.

As described in Section 34.2, switches are not fault-breaking devices. However, there have in the past been occasions when a switch has inadvertently been closed on to a fault and the operator's reaction, on realising his error, is immediately to attempt to open the switch. In order to prevent this happening most switches now have a delay incorporated such that, subsequent to closing, the switch cannot be opened for a period of at least 3 s. This allows time for the protection at the primary substation to operate to trip the associated circuit-breaker to clear the fault.

The tee-off fuse switch may also incorporate a fault-making earth switch on its HV terminations as a safety precaution against possible LV back-feeds.

Careful selection of the fuse rating in the fuse switch is necessary as the fuse needs to fulfil a number of functions. It must operate to clear a fault within the HV winding of the transformer and in the HV cable circuit to the transformer.

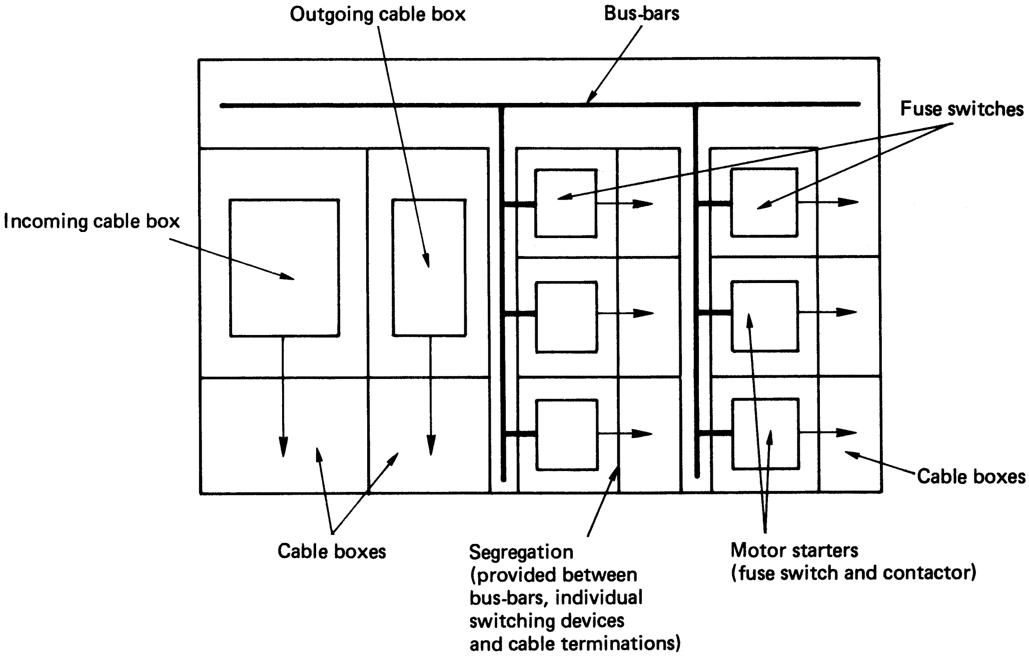

Outgoing cable box
Bus-bars
Fuse switches
Incoming cable box
Cable boxes
Cable boxes
Segregation (provided between bus-bars, individual switching devices and cable terminations)
Motor starters (fuse switch and contactor)

Figure 34.32 Construction of a high-integrity LV switchboard

It must be sized to discriminate with the LV fuses such that it does not operate for a downstream LV fault and it also must discriminate with the HV ring feeder circuit-breaker protection. It must not operate on transformer in-rush current and it must be sized to carry the rated current of the transformer or overload current of the transformer, which may be up to 150% of the transformer rating, for limited periods of time. It must also not deteriorate with age.

Most currently used designs of ring-main unit are oil-filled and use fuses under oil, although some may have fuses mounted in air.

SF_6 insulated designs of ring-main unit are now becoming available, their main advantage being the limited fire hazard in the event of a fault. However, they are relatively costly to produce since the enclosure must be gas tight. Relays or pressure gauges may be incorporated to give an indication of loss of gas pressure. HV fuses cannot be readily accommodated within the SF_6 insulated chambers and are usually mounted in air. Some manufacturers now use suitably rated vacuum or SF_6 circuit-breakers in place of the tee-off fuse switch. Whilst SF_6 insulated ring-main units can be made weatherproof for outdoor use, the general tendency now is to place them within a weather-protected housing. *Figure 34.33* shows an SF_6 insulated ring-main unit installed within a housing.

Modern safety legislation now requires equipment to be designed such that, even in the event of maloperation, it will present no danger to an operator. It is becoming common practice for new designs of equipment to be tested with a deliberately connected internal fault in order to demonstrate that any arcing products which may be expelled are directed away from where an operator may be standing whilst performing his normal duties. This is particularly pertinent for manually operated ring-main units.

Operating mechanisms which can be controlled remotely are available and provide the possibility for remote switching. However, pilot cable connections are seldom available for secondary distribution substations and this provision together with remote operational facilities can add significantly to the cost of a secondary distribution network where fault ratings are invariably low.

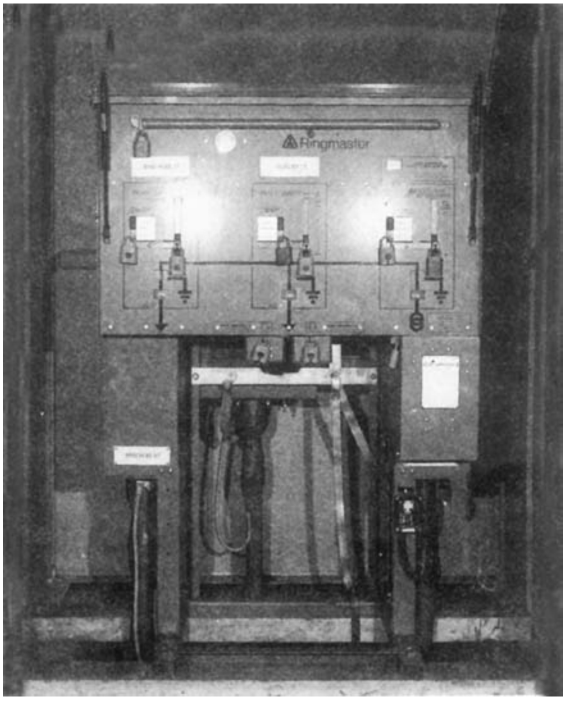


Figure 34.33 SF_6 insulated ring-main unit

34.5.2 Rural networks

Rural networks are usually fed from overhead-line systems operating at voltages of 10–36 kV. They may entail line lengths of some tens of kilometres which may only be lightly loaded.

The lines are fed from a primary switchboard and may comprise either a ring formation or radial network. The disposition of the load may be such that ring networks are not geographically practicable and the majority of circuits are radial with tee-off branches to consumers. The tee-off circuits traditionally have been protected by pole-mounted three-phase or single-phase expulsion fuse assemblies or, more recently, by automatic pole-mounted sectionalisers (see Section 34.1.5.1). The main line circuit is usually protected by a circuit-breaker incorporating an auto-reclosing facility.

Many overhead-line faults are transient in nature due, for example, to wind activity or lightning. The circuit-breaker will trip on fault initiation and the fault will then clear to allow the circuit-breaker to re-close automatically. Three or more reclosures may be permitted before it is assumed that the fault is permanent in nature when the circuit-breaker will then remain open. Under these conditions the whole of the overhead-line circuit will be out of commission with all consumers off supply until the faulted section can be located and repaired.

This situation can be alleviated by the use of pole-mounted auto-reclosing circuit-breakers, normally referred to as 'auto-reclosers', situated at suitable positions along the overhead line. The protection settings and number of permissible reclosures to lock-out can be pre-set such that only consumers downstream of a faulted section will be affected. Early auto-reclosers were operated by a falling-weight mechanism which needed to be re-charged subsequent to completion of a predetermined number of reclosures. More modern designs use solenoid coils operated directly from the HV feeding supply.

Auto-reclosers were traditionally oil-filled devices, but more recently SF_6 auto-reclosers have been available. A typical SF_6 auto-recloser is shown in *Figure 34.34*.

Sometimes a pole-mounted switch, often referred to as a 'sectionaliser' is used in conjunction with auto-reclosers, to allow manual reinstatement of circuits.

Rural consumers tend to suffer more and longer outages than urban consumers since the overhead line is more vulnerable to faults and longer distances are involved in operator travel to locate and repair the faulted section.

Where ring circuits can be employed, these offer the advantage of possible automation whereby a faulted section of line can automatically be isolated and all remaining sections up to the faulted section automatically reinstated. Signalling can be incorporated to allow the control engineer to direct the personnel directly to the circuit. Such systems are now technically viable and in use in some countries, but financial viability needs to be demonstrated.

34.6 HV primary distribution switchgear

HV primary switchgear feeding urban systems usually comprises cable-connected metal-enclosed switchboards housed within a brick-built substation. For feeding rural systems it may be in the form of open-type switchgear equipments, interconnected by air-insulated bus-bars, feeding overhead-line circuits. The metal-enclosed switchboard comprises disconnectors, bulk oil, SF_6 or vacuum circuit-breakers connected to a common bus-bar system and feeding via current transformers and voltage transformers on to a cable network. The bus-bars, circuit- breakers, current-transformer chambers and cable-box terminations may each be segregated within metal enclosures. Such equipment is referred to as being 'metal clad'. Live components within the enclosures may be air insulated or, more generally, insulated with solid material, usually cast epoxy resin, although more recently many different forms of cast resin materials have become available to the designer.

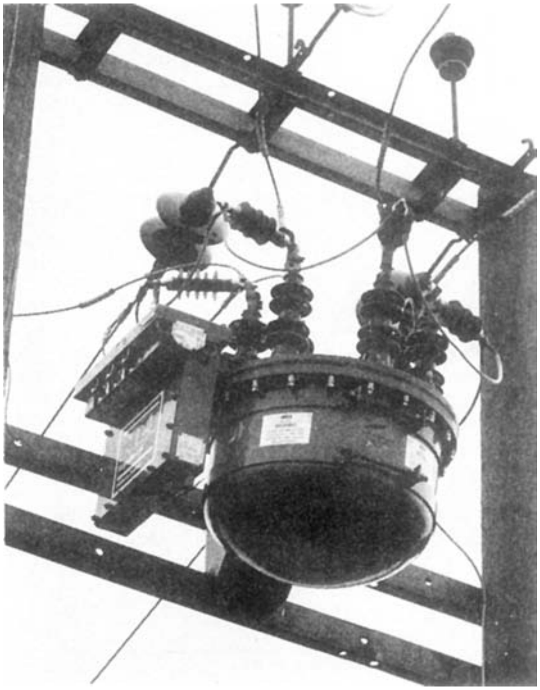


Figure 34.34 Pole mounted SF_6 auto-reclosing circuit-breaker

Bus-bars are usually segregated into two or more sections which can be coupled together by means of a bus-section circuit-breaker. A double bus-bar system may be provided for important installations where outages from a possible bus-bar fault cannot be tolerated. However, switchgear reliability is such that double bus-bar arrangements are now seldom employed at distribution-system voltages.

Outdoor substations at distribution voltages usually employ dead-tank oil, SF_6 or vacuum circuit-breakers incorporating ring type current-transformers around their bushings.

Manually operated disconnectors are usually employed. They are interconnected by copper or aluminium tubular bus-bars which are air insulated and supported on post insulators. The live equipment is mounted at a height of some 2 m above ground level to allow safe personnel access. Surge arresters may be employed at overhead-line entries into the substation.

34.7 HV transmission switchgear

HV transmission switchgear is used within substations to control the flow of large quantities of electric power within an electrical network. It may operate at rated voltages of 145–800 kV, with normal load currents of up to 4 kA and short-circuit currents of up to 80 kA.

The switchgear may be connected in the substation in various arrangements, the most common being the double bus-bar arrangement. Other commonly found alternatives are 1½ switch or mesh-connected arrangements. These are shown diagrammatically in *Figure 34.35*.

The switchgear may be of the open-type air-insulated construction basically similar to that described for open-type distribution switchgear or it may also be in the metal-enclosed form. The metal-enclosed form is a relatively recent innovation made possible by the use of SF_6 gas and is now widely employed, it is generally referred to as 'gas-insulated switchgear'. All live parts to earth are insulated with SF_6 and circuit-breakers also use it as their interrupting medium. Typical clearances between live parts and earth for 420 kV equipments might be some 250 mm which allows very compact substations to be built. A gas-insulated-switchgear substation, for example, may occupy only one-fifth of the ground area of an open-type substation. It will also have a low physical profile and is ideal for installation in urban networks where land availability and space is at a premium. Gas-insulated switchgear usually consists of phase-isolated assemblies at 420 kV and above but below this voltage the three phases may be housed in a common SF_6 enclosure. Cable compartments, bus-bars, disconnectors and circuit-breakers are generally segregated within their own gas enclosures. A typical 145 kV gas-insulated substation is shown in *Figure 34.36*.

In view of the very small electrical clearances involved between live conductors and earth, internal cleanliness is of utmost importance. For example, a small particle of a few millimetres in length can induce dielectric breakdown. Strict quality-control and inspection procedures are necessary during assembly. Such particles can induce partial discharge activity and most users now require monitoring for this activity in order to alleviate the consequences of dielectric breakdown. Partial discharges within gas-insulated switchgear excite resonances within the chambers. These resonances can be detected by the use of integral capacitive couplers and by examination of the frequency spectrum which may be generated. They are typically of the order of 1000 MHz. This recently developed technique is referred to as the 'ultra-high frequency partial discharge technique'.

34.8 Generator switchgear

It had long been the desire to be able to switch generator circuits directly at generator voltage, but circuit-breaker technology and testing techniques had not developed sufficiently to allow this. This situation changed some years ago and a few manufacturers now produce generator circuit-breakers capable of switching and clearing faults on the

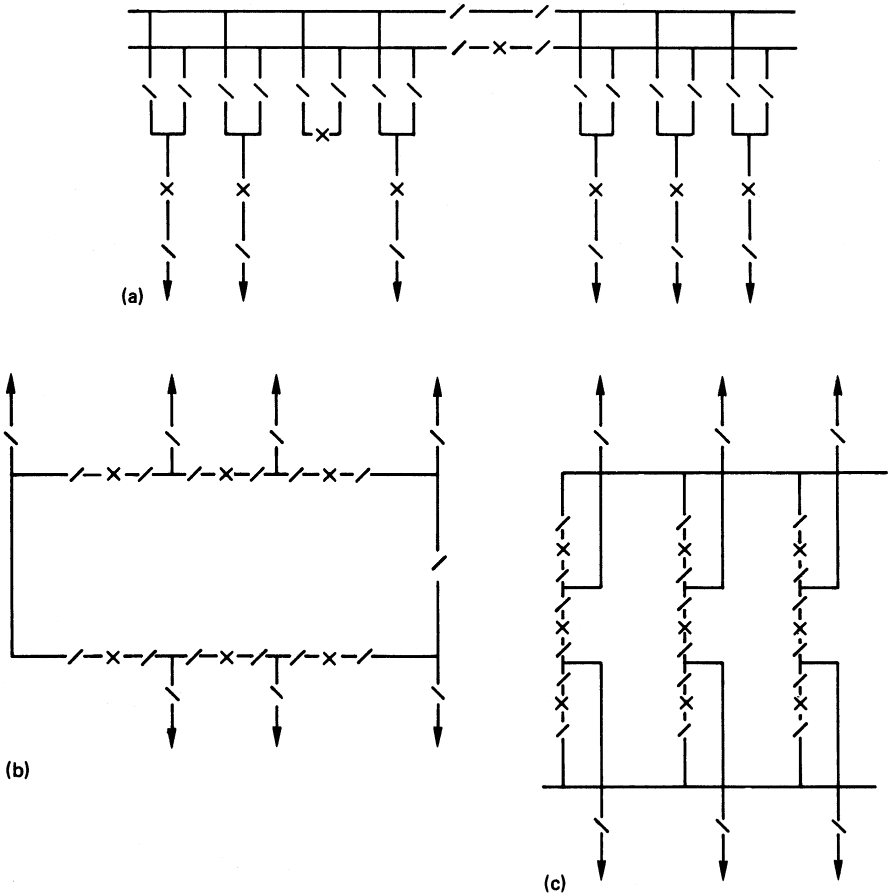


Figure 34.35 Typical substation arrangements: (a) double bus-bar; (b) mesh; (c) 1½; circuit-breaker

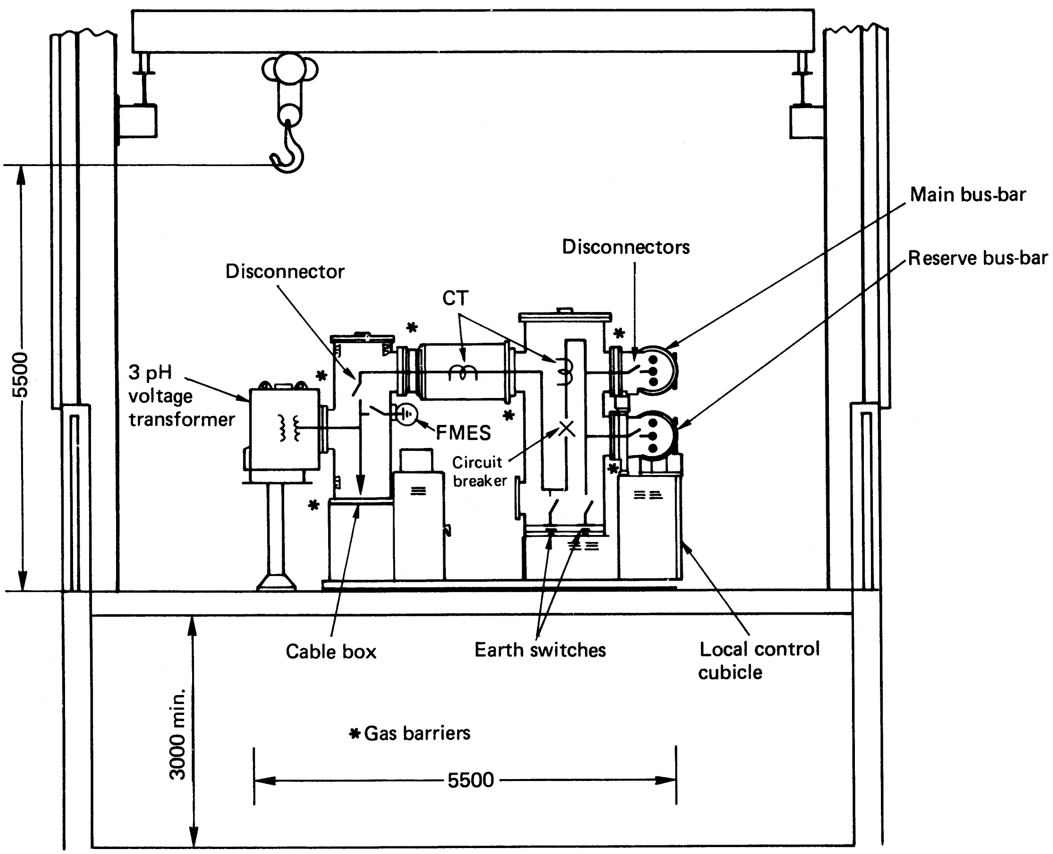


Figure 34.36 A 145 kV gas-insulated-switchgear substation. (FMES, fault making earth switch)

largest of generator circuits. Typical ratings may be 200 kA at 24 kV with rated load currents of 24 kA.

Early designs of generator circuit-breaker used air-blast technology, but more recent designs use SF_6 technology. The cross-section of a typical generator circuit-breaker is shown in *Figure 34.37.*

34.9 Switching conditions

Switching equipment must clearly be suitable for switching normal and, where required, abnormal conditions of the circuit in which it is situated. All circuits consist of series/parallel arrangements of resistance, reactance and capacitance and during switching these components will produce resonances of different forms. The switching equipments must therefore be capable of safely withstanding these resonant conditions during the switching operation. The switching conditions described in the following sections are regularly encountered.

34.9.1 Normal switching conditions

34.9.1.1 Shunt capacitor bank switching

When switching a shunt capacitor bank the load is purely capacitive. The source side of the circuit-breaker will include bus-bar capacitance to earth, this being small compared with the load capacitance, and series reactance. Since the capacitive current will be small compared with the circuit-breaker's rated short-circuit current, clearance of this current will often occur at the first available zero current. At this point of clearance a 1 pU charge will be left on the load-side capacitance. The source-side power frequency voltage will then swing to its opposite polarity to create a voltage across the opening circuit-breaker contacts of 2 pU. The circuit-breaker contact gap may be insufficient to withstand this voltage and breakdown across the gap might occur.

Thus the energy stored in the load capacitor is discharged to the bus-bar shunt capacitance and series inductance to cause a high-frequency oscillation. Large currents will flow across the opening contact gap and damage to arc control devices may occur. The term 're-ignition' is applied if this breakdown occurs within 0.25 cycles of arc extinction and 're-strike' if it occurs subsequent to this.

Circuit-breakers suitable for switching shunt capacitor circuits ideally should be designed and proven by test to be re-strike free.

34.9.1.2 Overhead-line and cable switching

Both overhead-line and cable switching conditions present problems similar to those of capacitor switching.

In the case of the overhead line there will be distributed capacitance between the phases, and to earth along the line,

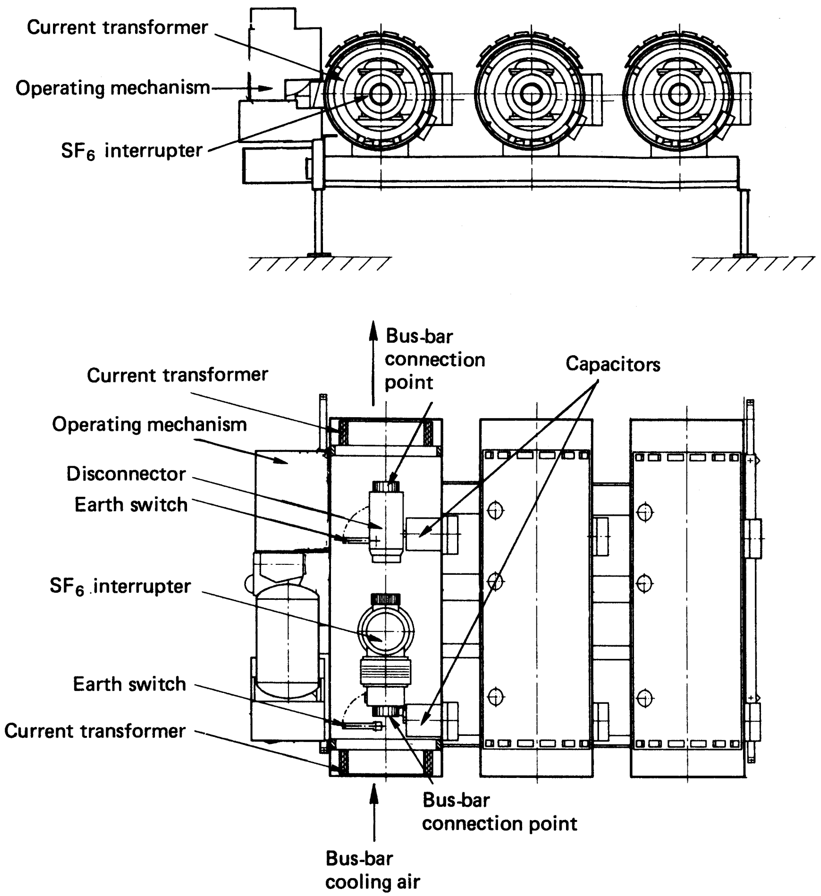


Figure 34.37 Cross-section of a typical generator circuit-breaker

interspersed with distributed line inductance. The load current is again small and its interruption is probable at the first available current zero. This will leave a charge on the line which may result in a travelling wave propagating along the line to the far end where it will reflect and return to the circuit-breaker.

The wave is generally attenuated when it reaches the circuit-breaker, but the circuit-breaker contact gap may be insufficient to withstand the voltage now appearing across its contacts and a re-ignition or re-strike may occur. Overhead-line switching may be a regular switching condition for a circuit-breaker and international specifications now require a circuit-breaker intended for overhead-line switching duties to be re-strike free.

A further problem may occur when switching an overhead line which is terminated at an unloaded transformer. The combination of line capacitance and transformer reactance then may cause high-frequency oscillations with phase to earth overvoltages in excess of 3.0 pU. Overhead-line switching regimes usually prohibit the switching of such circuits. If remote-end switching is necessary the overvoltage at the transformer can be limited by the application of surge arresters, but these would need to be rated to be capable of dissipating the energy associated with this switching condition.

With cable switching, the condition is similar to overhead-line switching except that capacitance values significantly higher than those associated with the overhead line will be encountered.

34.9.1.3 Transformer magnetising current switching

The arc-extinction methods described in Section 34.1.8 for circuit-breakers indicate that circuit-breakers generally clear on cessation of flow of current, i.e. at a natural zero current. When interrupting very low values of current this may not always be the case and the arc-extinction mechanisms may force the arc to extinguish before the current reaches its natural zero. This phenomenon is referred to as ‘current chopping’. Since the voltage appearing across the circuit is proportional to the rate of change of current times and inductance of the circuit, very high overvoltages may occur as a result of current chopping. This phenomenon occurs when switching low inductive currents, as for example may occur when switching the magnetising current of a transformer. The transient voltage will in fact oscillate at a frequency determined by the transformer magnetising inductance and winding capacitance. The oscillation will be damped by eddy and current hysteresis losses in the transformer. Overvoltages of the order of 3 pU may be generated.

Whilst vacuum circuit-breakers are particularly good at rapidly extinguishing low values of current, current chopping can occur when used for switching inductive currents such as transformers or motors.

34.9.1.4 Shunt-reactor switching

Shunt reactors are often used to compensate for the capacitance of lightly loaded lines. They may be frequently switched depending on the line-loading regimes. Shunt-reactor switching presents onerous switching conditions for a circuit-breaker and frequent maintenance may be necessary. At transmission voltages, shunt-reactor currents may be of the order of 400–600 A. The associated circuit-breaker will thus have to switch a very large reactive current with little parallel capacitance on the load side and very steep rates of rise of transient recovery voltage may result across the opening circuit-breaker contacts. These may cause multiple re-ignitions and, depending on the circuit-breaker type, may cause long arc lengths and arc durations. The dielectric withstand properties of the interrupter may rapidly deteriorate under these switching conditions.

34.9.1.5 Series-reactor switching

Series reactors are sometimes used to limit the short-circuit fault capacity between two connected sections of a supply system. Should a fault occur on the circuit side of the reactor this would cause a relatively high-frequency transient on the reactor side of the circuit-breaker and a low-frequency oscillation on the bus-bar side. A double frequency recovery voltage will thus appear across the opening contacts of the circuit-breaker and may result in a very high rate of rise of transient recovery voltage being imposed as the circuit-breaker contacts are opening. Thus circuit-breakers for series-reactor switching duties need to be tested for these specific conditions.

34.9.1.6 Disconnector switching

Whilst disconnectors generally have only to switch the small capacitive currents associated with a section of bus-bar between the disconnector and circuit-breaker, this can cause certain problems. Firstly, the disconnector must be capable of satisfactorily switching the maximum value of capacitive current likely to be encountered without causing contact or dielectric deterioration.

Secondly, the contacts of a disconnector are generally slow in operation and, when switching very low values of capacitive current, very many re-ignitions may occur. These will result in very high-frequency transients propagating along the bus-bar and may result in overvoltages of the order of 2.6 pU. The disconnected section of bus-bar may be left with a d.c. trapped charge which, in gas-insulated switchgear, may remain for some weeks. If, however, an electromagnetic voltage transformer is connected to the disconnected section of bus-bar, the trapped charge may dissipate through the primary winding of the voltage transformer and, depending on the relative circuit parameters, overheating or physical damage to the voltage transformer may occur.

When synchronising generators, a disconnector may be used to energise a short section of bus-bar from the generator circuit to an open synchronising circuit-breaker with the generator already excited to system voltage. If the open circuit-breaker has capacitive voltage grading, the voltage of the bus-bar between the open circuit-breaker and open disconnector may be very close to the system voltage since capacitance to earth will be small. This means that the open disconnector contacts have, on one side, a voltage fed from the system which is close to the system voltage, and on the generator side a voltage close to the system voltage fed from the generator. The generator frequency will not align with system frequency and the voltage across the disconnector contacts can rise to 2 pU. As the contacts close the contact gap may flashover at this 2 pU voltage which will leave a d.c. trapped charge superimposed on the a.c. voltage on the section of bus-bar between circuit-breaker and disconnector. The voltage across the disconnector contacts then may rise to as high as 3.5 pU.

The three disconnector switching conditions described above need to be proven by test, and international specifications covering these requirements are in preparation.

Disconnectors are also commonly used to transfer load current from one bus-bar to an adjacent parallel connected bus-bar. Under these conditions the disconnector may have to make and break full-load current with a voltage across its contacts equal to the impedance drop around the parallel circuit. Whilst this may only be some tens of volts, contact burning can result. Disconnectors require to be designed such that this type of operation does not adversely affect the dielectric properties of an open disconnector.

International specifications now call for disconnectors to be proven for switching this condition.

34.9.1.7 Earth switch/switching

Since earth switches are only applied to a de-energised circuit they should not, under normal conditions, be required to make or break current. However, when an overhead line is taken out of service and where work is required on the line, it is necessary to earth the line at both ends. In the case of a double circuit overhead line with one circuit earthed, the earthed line will have currents induced into it by means of magnetic coupling from the adjacent loaded line and inductive currents will circulate around the closed loop of the earthed line. The first earth switch to open will then have to break this inductive current which, at 420 kV, may be of some hundreds of amperes and the voltage appearing across the earth switch contacts may also rise to a few kilovolts. With the line then earthed at one end only, a current to earth will be induced in the line equal to the capacitive current flowing between the two parallel lines. When the second earth switch opens it will have to clear this capacitive current, which again for 420 kV may be some tens of amperes, but the voltage across the open contacts may now increase to some tens of kilovolts.

New international specifications will incorporate tests to verify earth-switch performance when operating under these conditions.

34.9.2 Abnormal switching conditions

The main difference of a circuit-breaker compared with other mechanical switching devices is its ability to satisfactorily clear abnormal circuit conditions as may occur, for example, when a short-circuit occurs on the downstream side of the circuit-breaker. Under these conditions the circuit is mainly inductive and the short-circuit current lags the system voltage by almost 90°. This means that when the current passes through a zero point the voltage across the opening circuit-breaker contacts is at or near its peak value. The arc-extinguishing process must therefore be such as to ensure that, at a suitable current zero, sufficient dielectric strength has built up to withstand this peak voltage. For an air-blast or gas-blast SF_6 circuit-breaker this will usually occur at, or before, the second available current zero. For oil circuit-breakers several current zeros may pass before arc extinction occurs. Unfortunately the process is more complicated than that described since during the arcing period the voltage across the circuit-breaker contacts is small and, at clearance, it must change from this small value to the peak of the system voltage in a very short period of time. In doing so the voltage across the contacts will tend to rise to twice the peak voltage which will set the downstream circuit into a damped oscillation at its natural frequency. This transient voltage is known as the 'transient recovery voltage'. For a given circuit the value of the fault current will depend on the impedance of the circuit up to the point of a fault, and fault currents will clearly be lower for longer distance faults. This means that a greater length of the circuit will be available to resonate on fault clearance. This has the general effect that the rate of rise in the transient recovery voltage, and its peak value, tend to be higher for lower values of fault current. This situation is covered in international specifications by calling for short-circuit tests at 100%, 60%, 30% and 10% of the circuit-breaker current with the transient recovery voltage parameters being specified for each value of short-circuit test current.

An additional complication arises where the ratio of inductance to resistance can be high, for example, for faults close to transformer terminals. This causes the a.c. current to contain a large decaying d.c. component of current. When the circuit-breaker contacts open, this decaying d.c. current may not have reached zero and the circuit-breaker will be required to clear an a.c. current off-set from the zero line. This is referred to as an 'asymmetrical short-circuit current'. This has the effect that, depending on the circuit-breaker type, clearance, for example at a minor loop of current, might be less onerous than clearance at a major loop of current. Again international specifications detail precise parameters to be met by a circuit-breaker when clearing these conditions.

A similar situation might arise when a circuit-breaker is required to close against a short-circuit when the system reactance to resistance ratio (X/R) is high. A large d.c. current in the initial part of the a.c. current wave may occur and the d.c. current peak may, in theory, reach $2 \times \sqrt{2}\ I_{\mathrm{rms}}$. For typical X/R ratios the peak current is, in practice, nearer to $1.8 \times \sqrt{2}\ I_{\mathrm{rms}}$ and it is this value which the circuit-breaker must satisfactorily close against. As large mechanical forces will result from the electromagnetic effects of the peak short-circuit current, the circuit-breaker operating mechanism must have sufficient energy to overcome these forces.

As described previously for disconnectors, a circuit-breaker may be required to close, and open, under conditions where it ties together two large power systems the frequencies of which may differ by up to 180° prior to circuit-breaker closure or during opening. Under these conditions the current that flows may be of the order of 25% of the rated short-circuit current of the system, but the recovery voltage across the circuit-breaker contacts may, in the extreme, be as high as 3 pU.

A further onerous condition may occur for a fault a short distance along an overhead line whereby the line impedance is such as to reduce the short-circuit current to some 75–95% of its rated value. This is termed a 'short-line fault condition'. At a current zero the voltage on the line-side terminal of the circuit-breaker is equal to the impedance drop along the faulted section of line. When the circuit-breaker opens to clear the fault this impedance voltage drop will cause a transient voltage to propagate up and down the short section of faulted line. This will result in a decaying 'saw tooth' wave on the line side of the circuit-breaker. The voltage across the circuit-breaker contacts will be the sum of this line-side voltage and the source-side power frequency voltage. This will result in a very steep rate-of-rise of the initial part of the transient recovery voltage wave thus presenting a very onerous clearance condition for the circuit-breaker. Again specifications require transmission circuit-breakers to be tested for this condition.

34.10 Switchgear testing

During short-circuit fault clearance an interrupter of a circuit-breaker will be required to dissipate a very large amount of energy. For example, higher rated circuit-breakers may be required to safely dissipate, over a very short period of time, something in excess of 1000 MW. The circuit-breaker must be capable of satisfactorily performing such a rated short-circuit duty at least three times, without danger to operators, or without danger to the system. Whilst present day theory can significantly assist in the development of interrupter design, it can still not satisfactorily ensure the safe operation of the complete circuit-breaker. It is thus necessary for the circuit-breaker performance to be proven under very specific short-circuit test conditions in order to verify its safe performance.

Whilst tests can be, and sometimes are, performed on actual systems, the system parameters cannot readily be modified to achieve the required test conditions. In addition, such testing may induce significant risk to system operation and is rarely undertaken nowadays.

Short-circuit testing is thus performed in specially designed short-circuit testing stations. These may entail the use of 1–4 large, specially designed, generators for producing the necessary output. The supply to the generator is usually disconnected immediately prior to the application of a short-circuit and a large flywheel may be attached to the generator shaft to maintain its inertia. The output of the generators may be paralleled to produce the necessary short-circuit ratings and they may also be connected via transformers so that a wide variation of test settings can be achieved. Reactors, resistors or capacitors may be added to both the source and load sides of the circuit-breaker in order to ensure the achievement of the specified test parameters. Such short-circuit testing is generally referred to as 'direct testing' and may be either single phase or three phase. For distribution circuit-breakers, for example, three-phase direct testing can nearly always be achieved. However, for large circuit-breakers the output of the short-circuit test station may be insufficient to allow direct testing of the complete three-phase circuit-breaker. Under these conditions a section or 'unit' of a circuit-breaker may be tested, i.e. one interrupter of a two-interrupter circuit-breaker. Test parameters on this 'unit' must be such as to reproduce the worst conditions that would apply to the interrupter when the circuit-breaker was operating to clear the specified three-phase fault condition. This method of testing is referred to as 'unit testing'.

Even with unit testing, however, the output of the short-circuit testing station may still be insufficient to meet the specified short-circuit levels. It is then necessary to revert to a more recently developed method of short-circuit testing known as 'synthetic testing'. With synthetic testing the current is still obtained from a short-circuit generator, but the recovery voltage is obtained from a separate supply circuit and is injected across the opening circuit-breaker contacts at an appropriate current zero. The recovery voltage circuit usually comprises a large precharged capacitor bank assembly which is switched, at the precise moment in time, via a combined inductance, capacitance and resistance network to produce the correct transient recovery peak voltage and frequency. Synthetic unit testing is now the standard method used for testing large transmission circuit-breakers.

International specifications have been produced which detail precise test requirements and methods of tests covering all the normally required short-circuit test conditions for circuit-breakers.

In addition to short-circuit testing there are other ‘type tests’ to which a circuit-breaker must be subjected in order to demonstrate its performance. These comprise a mechanical endurance test, generally of 1000 operations without maintenance, a temperature-rise test at rated current, a HV test and, for outdoor circuit-breakers, an environmental test.

The HV tests verify the dielectric performance of the circuit-breakers and generally comprise a power frequency overvoltage test, a lightning impulse voltage test, to simulate the effects of a lightning strike, and for higher voltage transmission circuit-breakers a switching surge on the system.

A further test is usually performed at a specified power frequency overvoltage to check for any partial dielectric-breakdown activity that might be occurring in faulted insulation. This is referred to as a ‘partial discharge test’.

Environmental testing on outdoor circuit-breakers entails subjecting the circuit-breaker to cyclic variations in ambient temperature, from its maximum working temperature of, say, 35°C to its minimum working temperature of, say, -25°C. It may also entail the application and verification of mechanical operations at various points during the temperature cycling and of checks for leakage of insulant.

The above tests are usually performed as type tests on one equipment of a specific design and are intended to demonstrate the suitability of that design of equipment.

Further works tests must be performed on production equipment and these are referred to as ‘routine tests’. These usually comprise a power frequency overvoltage test, measurement of resistance of primary circuits and a mechanical operations test. Similar tests are generally repeated once equipment has been erected on the site. Site tests would, in addition, include primary circuit injection tests to verify circuit-breaker operation via protective relays.

Whilst clear and precise erection instructions are necessary for all switchgear equipments this is particularly important for gas-insulated switchgear where even very small particles may, for example, induce dielectric breakdown. Site quality-control checks are of paramount importance and many users of gas-insulated switchgear now require the measurement of partial discharge activity as a site commissioning test and some also require on-line monitoring for partial discharge activity as an essential operational criterion.

34.11 Diagnostic monitoring

The concept of using maintenance-free switchgear whereby maintenance would not be required during the lifetime of the equipment is an ideal user objective. Indeed, with modern SF_6 switchgear, this ideal objective is now becoming more of a reality. On the other hand, current legislation may require equipments to be regularly maintained to be in a serviceable condition and, therefore, there is a need to be able to verify that equipment remains in a serviceable state without the need for physical dismantling. The application of techniques to monitor externally the internal state of equipment can reveal many incipient problems and such diagnostic monitoring techniques are currently being developed and applied. Continuous on-line monitoring and self-diagnosis using fibre-optic technology is now becoming a practical reality. Parameters that can be readily monitored are, for example, contact speed, contact engagement, moving system damping, circuit-breaker operating times, summated interrupted fault current, operating mechanism pressures/times, conductor temperatures, dielectric gas pressures, liquid levels and, for gas-insulated switchgear, partial discharge activity. In addition, system-fault currents and associated voltages can now readily be recorded and these data, in conjunction with the above on-line monitoring, can assist in ascertaining the internal condition of a circuit-breaker.

34.12 Electromagnetic compatibility

‘Electromagnetic compatibility’ refers to the susceptibility of electrical equipment to any electromagnetic interference phenomena to which it might be subjected as well as to electro-magnetic interference signals which the equipment itself may emit. Levels are currently being specified for maximum emissions from electrical equipment and also for levels of interference below which the equipment would not be affected. Switchgear equipment will be required to comply with the criteria.

Such interference phenomena may be caused, for example, by lightning, switching surges, high-frequency-discharge activity, voltage dips, signalling on lines, harmonics, and superimposed d.c. on an a.c. system. Some of the phenomena may be generated by the switchgear itself and may cause malfunction of associated electrical equipment, e.g. alarm circuits and relays. Typical phenomena in a metal-enclosed gas-insulated-switchgear substation may, for example, be disconnector induced transient overvoltages, these may be coupled either by conduction or by radiation into secondary connections and other associated equipments causing flashover or dielectric breakdown. It is then necessary for steps to be taken to limit the coupling of these signals. Typical steps might, for example, be: the provision of an earth screen between HV and LV windings of voltage transformers; short direct earth connections to a high-frequency earth mat; the use of screened secondary cables with the screening earthed at both ends of the cable; cables should be run together, preferably within a trench and parallel with main earth connection; and flow and return conductors should be within the same screened cable.

With ever-increasing use of modern electronic equipments this subject is gaining significant importance to the extent that in some countries not only are limiting values specified but legislation is being introduced to make compliance mandatory.

34.13 Future developments

In the last ten years or so since deregulation the Electrical Power Industry has been through major re-organisational changes. In the main, the way in which electricity is produced, transported and delivered to customers is much the same as it has always been. There is one significant change however, and that is the advent of combined cycle power plants which can be located nearer to load centres, within very short time scales, at lower costs and with reduced emissions compared with conventional coal fired plants. All of this minimises the requirements for transmission. If this trend continues, as it is likely to, then the way in which transmission systems operate will significantly change. In addition, there is also a trend for more and more generation to be embedded in the distribution systems so the way in which they operate will also significantly change. These factors mean that the requirements for switchgear will also have to adapt to the changing power systems. For distribution systems the switchgear will need to be re-locatable such that as new systems evolve it can be positioned to the most appropriate system node. It will also have to cater for power flows in both directions which means that protection systems will need to be modified. The switchgear should have improved reliability, require the minimum of maintenance and be capable of complete operation from a remote source. It must also be cost effective and in fact, may be leased to the utility and operated by an independent owner.

Similarly transmission switchgear will follow the system trends and many of the functions described for distribution switchgear will also be required.

Fortunately, many new technology developments are becoming available. There is a major breakthrough in polymers, for example, these can now be designed on a computer to formulate the appropriate molecular structure to meet the performance requirements. Similarly, there are major developments in magnetic materials, both conducting and insulating ceramics, superconductivity and solid state thyristors such that conventional power transformers may become a thing of the past. Developments in hard materials utilising diamond technology and carbon C60 will undoubtedly assist in the practical development of current limiting/solid state/superconducting circuit breakers. Already prototype equipment's are being developed and the concept of switchgear equipment's embodying all these new technologies, together with fundamental changes in the ways electricity is produced and provided to the customer presents many new and exciting challenges for younger engineers.

Acknowledgements

The author acknowledges the assistance provided by Scottish Power plc., in the preparation of this chapter and the assistance give by the following companies in the preparation of many of the figures: ABB Nitran Ltd, ABB Power Ltd, Dorman Smith Switchgear Ltd, GEC Alsthom Transmission Switchgear Ltd, GEC Alsthom Installation Equipment Ltd, GEC Henley Ltd, Hawker Siddeley Switchgear Ltd, Hawker Siddeley Fusegear Ltd, Long & Crawford Ltd, NEI Reyrolle Ltd, Yorkshire Switchgear Ltd.

References

BAXTER, H. W., *Electric Fuses*, London: Edward Arnold (1951)

BLOWER, R. N., *Distribution Switchgear: Construction Performance Selection and Installation*, London: Collins (1986)

BOGS, S. A., CHU, F. Y. and FUGIMOTO, N., *Gas Insulated Substations*, Toronto: Permagon Press (1986)

FLURSCHEIM, C. H., *Power Circuit Breaker Theory*, Stevenage: Peter Peregrinus (1982)

GILES, R. L., *Layout of EHV Substations*, Stevenage: Peter Peregrinus (1970)

HAGUE, B., *Instrument Transformers*, London: Pitmans (1936)

JACKS, E., *High Rupturing Capacity Fuses: Design & Application of Safety in Electrical Systems*, London: E & F. M Spon (1985)

KREUGER, F. H., *Partial Discharge Detection in High Voltage Equipment*, London: Butterworth (1989)

LAFFERTY, J. M., *Vacuum Arcs: Theory and Application*, Chichester: Wiley (1980)

LOOMS, J. S. T., *Insulators for High Voltage*, Stevenage: Peter Peregrinus (1988)

LYTHALL, R. T., *J & P Switchgear Book*, London: Butterworth (1979)

MATTHEWS, P., *Protective Current Transformers*, London: Chapman & Hall (1955)

RYAN, H.M. and JONES, G. R., *SF_6 Switchgear*, Stevenage: Peter Peregrinus 1989

SMEATON, R. W., *Switchgear and Control Handbook*, New York: McGraw-Hill (1987)

WHITEHEAD, S., *Dielectric Breakdown of Solids*, Oxford: Clarendon Press (1951)

WRIGHT, A., *Current Transformers*, London: Chapman & Hall (1968)

WRIGHT, A. and NEWBERY, P. G., *Electric Fuses*, Stevenage: Peter Peregrinus (1982)

35

Protection

A T Johns PhD, DSc, CEng, FIEE, SMIEEE, FRSA
University of Bath

Contents

35.1 Overcurrent and earth leakage protection 35/3
  35.1.1 Fault conditions 35/3
  35.1.2 Protective equipment 35/4
  35.1.3 Relays 35/6
  35.1.4 Solid-state equipment 35/7
  35.1.5 Overcurrent protection 35/7
  35.1.6 Fuses 35/9
  35.1.7 Earth leakage protection 35/10
  35.1.8 Balanced (differential) protection 35/11
  35.1.9 Miscellaneous equipment 35/14
  35.1.10 Efficacy of protection scheme 35/15
  35.1.11 Digital protection 35/15
  35.1.12 Artificial intelligence for protection 35/18

35.2 Application of protective systems 35/20
  35.2.1 Plant 35/21
  35.2.2 Feeders 35/23
  35.2.3 Motors and rectifiers 35/24

35.3 Testing and commissioning 35/25
  35.3.1 Commissioning tests 35/25
  35.3.2 Primary current tests 35/26
  35.3.3 Secondary injection tests 35/26
  35.3.4 Fault location 35/26

35.4 Overvoltage protection 35/27
  35.4.1 Insulation co-ordination 35/27
  35.4.2 Protective equipment 35/28

35.1 Overcurrent and earth leakage protection

The main effects of fault current in a power system are:

(1) disturbance of the connected load;
(2) overheating at the fault point and in associated plant;
(3) electromagnetic forces of abnormal magnitude, with consequent mechanical damage; and
(4) loss of synchronous stability.

The function of the protective equipment is to isolate the faulty plant from the running system by initiating tripping signals for appropriate circuit-breakers. It should therefore *discriminate* between faulty plant and sound plant carrying through-fault current. The whole process must be effected with the minimum of delay and disturbance.

Exceptionally, some faults may be allowed to persist: an example is an earth fault on a system earthed through an arc suppression (Petersen) coil, or one having an isolated neutral, where the fault may be located and alternative feeds provided before the fault is isolated.

Faults are due to insulation breakdown by deterioration, overvoltage, mechanical damage or short-circuit effects; they may be simple or complicated, and involve conductor breakage or short-circuited turns on transformers or generators.

The incidence of faults on cables and lines depends on the installation and the climatic conditions. For a typical system in a temperate zone the distribution of faults as percentages of the total is approximately: overhead lines, 60; cables, 15; transformers, 12; switchgear, 13. The causes of overhead-line faults, and the number of faults per 100 km of line per year, are, typically: lightning, 1.0; gales, 0.15; fog and frost, 0.1; snow and ice, 0.06; salt spray, 0.06. The total is 1.37 faults per 100 km per year: in tropical countries lightning faults may be markedly more numerous. In general, the higher the voltage the lower the number of lightning faults.

35.1.1 Fault conditions

Below are given circuit and phasor diagrams relevant to the behaviour of protective equipment under the simpler fault conditions.

35.1.1.1 Three-phase fault

A three-phase fault usually develops first as a phase-earth fault, and it may be unbalanced. Even when a circuit-breaker closes on to a three-phase fault, one phase may momentarily be faulted before the other two, a matter of importance in high speed protection. *Figure 35.1* shows the relevant phasor diagram. At the fault the voltages to neutral are zero; and if AN, BN, CN are the phase-to-neutral pre-fault voltages, the voltages at some distance from the fault (e.g. at a relay controlling a circuit-breaker) are A′N, B′N, C′N. With a symmetrical short-circuit the conditions are independent of the system earthing arrangements.

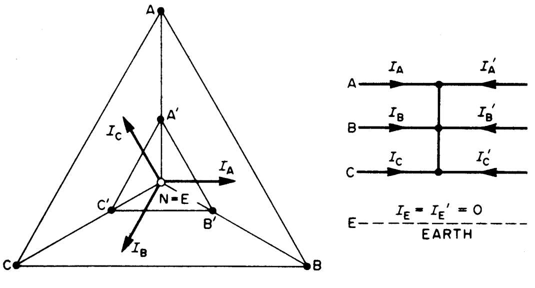


Figure 35.1 Three-phase fault

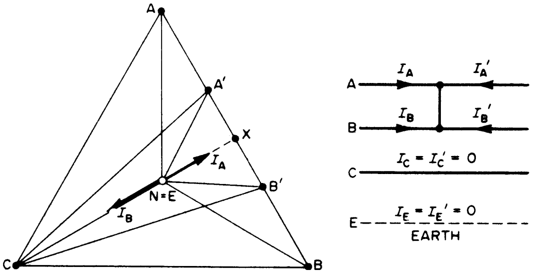


Figure 35.2 Phase-to-phase fault

35.1.1.2 Phase-to-phase fault

Two phases are faulted clear of earth, an unusual kind of fault even less likely on cables than on overhead lines. *Figure 35.2* shows AN, BN and CN as the normal voltages at the site of the fault. On occurrence of the fault, the voltages there are XN, XN and CN, and the voltages at some specified distance are A′N, B′N and C′N, with the locus of A′⇐and B′⇐on the line AB. The voltage to earth of the sound phase is not affected. The conditions are independent of the system earthing, but the fault current will be reflected into the further sides of associated transformers in accordance with their connections.

35.1.1.3 Single phase-to-earth fault

This type of fault is considered for a system with the generator or transformer having its neutral earthed (a) solidly, (b) through a reactor, (c) through a resistor.

Neutral solidly earthed The voltages to earth at the point of fault with phase A earthed are zero, B′E and C′E, as in *Figure 35.3*, and at some distance away the voltages are A′E′, B′E′⇐and C′E′.⇐ The locus of A′⇐of the line AN, and the voltages to earth of the sound phases are slightly changed to B′E and C′E by reason of the currents in these phases between the two infeeds. With the system earthed at one point only (apart from the fault), the total earth fault current divides between the two infeeds. If there are two (or more) system earth points and the infeeds are I_A and I'_A⇐,

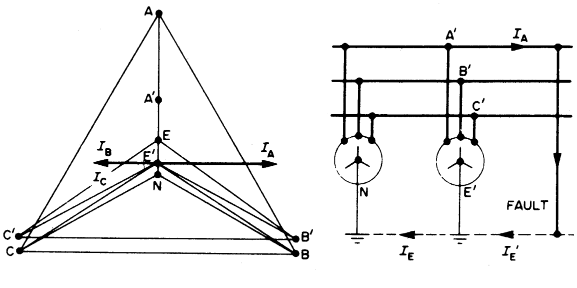


Figure 35.3 Single phase-to-phase earth fault (solid earthing)

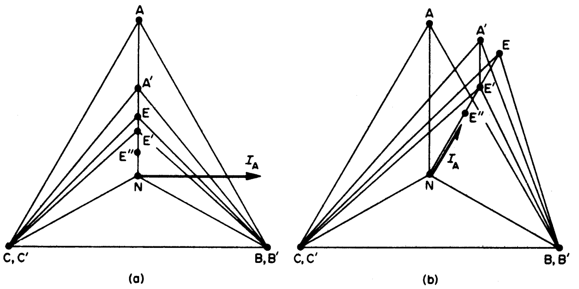
there may be currents in the sound phases of direction decided by the preponderance of zero-over positive- and negative-sequence impedance of one infeed compared with the other. In this case the residual or earth current on each side wil be $I_A + I_B + I_C = I_E$; the residual voltage will be 0 + B′E + C′E at the fault, and A′E′ + B′E′ + C′E′⇐some distance away.

Earthed through reactor With a low-valued phasor (small compared with an arc suppression coil), the earth point potential on the occurrence of a fault will be E (*Figure 35.4*(*a*)), and the drop in the reactor will be E′N. The fault position voltages are zero; voltages to earth of the sound phases are increased. If the reactor is an arc suppression coil, point E is raised to A, the sound phase voltages to earth become $\sqrt{3}$ times normal, and the residual voltage at the fault will be 3 times the normal phase-to-neutral value.

Earthed through resistor The conditions (*Figure 35.4*(*b*)) are essentially similar to those for reactor earthing except that there is unbalance in the voltages to earth of the sound phases. The volt drop across the resistor is E″N.

35.1.1.4 Double phase-to-earth fault

With *solid earthing* at one point only, the voltages to earth at a double phase-to-earth fault will, as shown in *Figure 35.5*, be zero, zero and C′E, and at some distance away will be A′E′, B′E′ and C′E′. The voltage to earth of the sound phase may be increased at the point of fault. The angle α between the phase currents I_A and I_B (or I_A and I'_B⇐) may be less than 60°, depending on the ratio of balanced to zero-sequence impedance; the angle varies considerably, according to the earthing arrangements, and in the limit with the neutral point isolated it becomes 180°.⇐ The residual current (i.e. the earth current) is the vector sum of the currents in the faulted phases and that (if any) in the sound phase. The residual voltage at the point of fault is the normal phase-to-neutral voltage, and it decreases with distance from the fault position.



(a) (b)

Figure 35.4 Single phase-to-earth fault: (a) through reactor; (b) through resistor

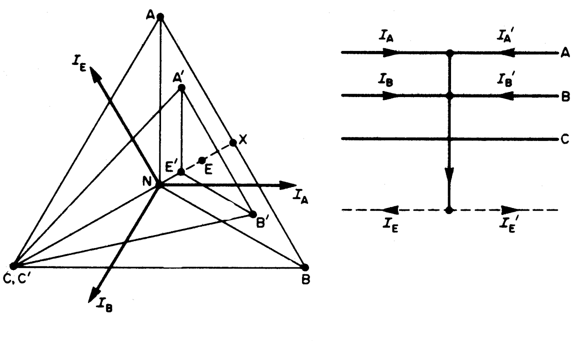


Figure 35.5 Double phase-to-earth fault

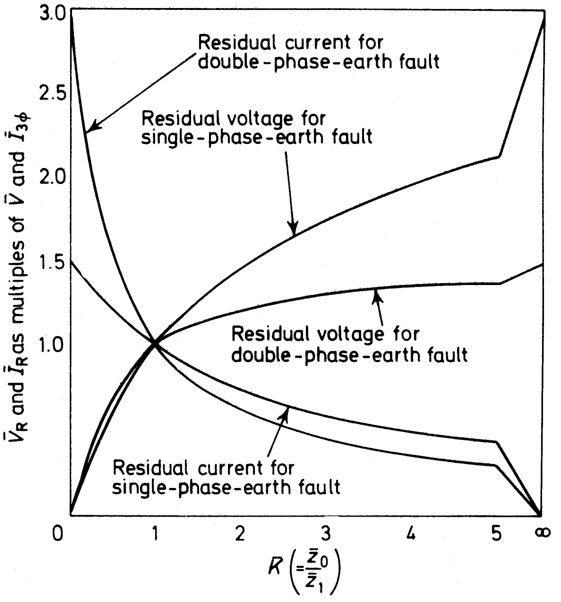


Figure 35.6 Variation of residual quantities at fault point

With the system neutral earthed through a low-valued reactor, the drop across the latter is NE′ and the voltage to earth of the sound phase is CE, increasing to a possible maximum of 1.5 times normal value if the reactor is an arc suppression coil.

With system earthing through a *resistor,* the point E in *Figure 35.5* is no longer on the line CX.

The system Z_0/Z_1 ratio is defined as the ratio of zero sequence and positive sequence impedances viewed from the fault. It is a variable ratio dependent on the method of earthing, fault position and system operating arrangement. When assessing the distribution of residual quantities through a system, it is convenient to use the fault point as the reference, as it is the point of injection of unbalanced quantities into the system. The residual voltage is measured in relation to the normal phase-neutral system voltage, and the residual current is compared with the three-phase fault current at the fault point. It can be shown that the character of these quantities can be expressed in terms of the ratio Z_0/Z_1 of the system; see *Figure 35.6.* The residual current in any part of the system can be obtained by multiplying the current from the curve by the appropriate zero sequence distribution factor. Similarly, the residual voltage is calculated by subtracting from the voltage curve three times the zero sequence drop between the measuring point in the system and the fault.

35.1.2 Protective equipment

The satisfactory operation of protective equipment (which extends from simple domestic subcircuit fuses to sophisticated electronic system devices) depends largely on *co-ordination.* The influencing factors include the network layout and characteristics, fault megavolt-ampere levels, earthing, the availability of pilot cables and the physical extent of the system. Generally, the higher the voltage and fault levels the more necessary is quick-acting discriminative protection covering lines and plant.

Protective equipment may be broadly classified into: (a) *restricted zone,* giving full discrimination; (b) *semi-restricted zone*; and (c) *unrestricted zone,* with no discrimination. Type (a) can be considered as applying to a network of such magnitude that the disconnection of each item of plant is necessary if an internal fault develops. Types (b) and (c) apply to less important installations and also as 'back-up' in restricted-zone schemes.

35.1.2.1 Modes of operation

An alternative classification based on the operating mode is as follows.

Overcurrent equipment operates if the current exceeds a pre-set value. Restrictions regarding the direction of the overcurrent are often included to improve discrimination.

Balanced (differential) equipment operates if an out-of-balance occurs between currents or voltages which under normal conditions are balanced.

Distance (impedance) equipment operates if the impedance (proportional to distance), as viewed at a circuit-breaker supplying a feeder, falls below a specified value.

Miscellaneous equipment is designed for special purposes.

35.1.2.2 Equipment elements

The chief items of plant making up a complete protective installation are as follows.

Current and voltage transformers These provide convenient levels of current or voltage proportional to the system values. Rated secondary values are normally 1 or 5 A (current transformer (c.t.)), and 110 V (voltage transformer (v.t.)).

Relays These, operating from c.t. and v.t. detect the presence of faults and energise the trip circuits of the appropriate circuit-breakers. Electronic (solid state) devices are now generally available as alternatives to the electromagnetic types, and give improved performance.

Pilot circuits These carry signals between points in the system for comparison, for feeder protection and other purposes. The circuits may be provided in a variety of ways:

(1) conductors specially installed for the purpose;
(2) rented telecommunication lines; the current and voltage limitations are 60 mA and 130 V peak, and the pilots must be insulated from all power system equipment for 15 kV;
(3) power transmission lines, using carrier signals at frequencies between 70 and 700 kHz;
(4) radio links using microwave transmission between line-of-sight terminals; and
(5) optical links, usually embedded in the earth wire.

The choice requires very careful consideration, particularly for long distances. Even apparently simple conductors, for which the resistance and capacitance between ends can be measured, often act as a transmission line operating in conjunction with relay equipment. The performance of the pilot circuit may be complex, and in operation is made even more so by inductive interference effects.

It is possible to obtain high-speed feeder protective schemes using voice frequencies of 600–3000 Hz, which are suitable for pilot-wire channels, or pilot-line carrier or radio channels.

The overall fault clearance time is made up of the signalling time, and the operating times of the protection relay, trip relay and circuit-breaker opening. This total time must be less than the maximum for which the fault can remain on the system for minimum plant damage, loss of stability, etc. In recent years practical minimum times have been achieved in reducing protection times from 60 to 20 ms, trip relay times from 10 to 3 ms, circuit-breaker times from 60 to 40 ms, leaving protection signalling times to be reduced from 70–180 ms down to 15–40 ms in the UK, and to 5 ms in certain parts of the world where system stability is critical.

Protection current transformers may have to maintain accuracy up to 30 per unit overcurrent, although often the error is less important than that c.t. characteristics should match.

Current transformers The errors are due to the exciting current, and they vary with the phase angle of the secondary burden. An increased burden demands a corresponding increase in core flux, and as the exciting current is a non-linear function of the flux, it is subject to a more than proportional rise accompanied by a greater harmonic content, so that the composite error is increased.

Protective equipment is intended to respond to fault conditions, and is for this reason required to function at current values above the normal current rating. Current transformers for this application must retain a reasonable accuracy up to the largest relevant current, known as the 'accuracy limit current', expressed in primary or equivalent secondary terms. The ratio between the accuracy limit current and the normal rated current is known as the 'accuracy limit factor'. Standard values are 5, 10, 15, 20 and 30. Protective c.t. ratings are expressed in terms of rated burden, class and accuracy limit factor (e.g. 10 V-A class 10P 10). Classes 5P and 10P are useful only for overcurrent protection; for earth fault protection, in particular, it is better to refer directly to the maximum useful e.m.f. which can be obtained from the c.t. In this context the 'knee point' of the excitation curve is defined as that point at which a further increase of 10% secondary e.m.f. would require an increment of exciting current of 50%. Such current transformers are designated class X.

Current transformers may have primaries wound or of bar ('single-turn') form for mounting on bushings. For high-voltage systems the c.t.s may be separately mounted with oil or SF_6 insulation.

Voltage transformers The main requirement in protection is that the secondary voltage should be a true reflection of the primary voltage under all conditions of fault. It is usual with electromagnetic v.t.s to apply additional delta-connected windings (*Figure 35.7*) to give a measure of the residual voltage at the point of connection. The three voltages of a balanced system summate to zero, but this is not so when the system is subject to a single-phase earth fault. The residual voltage in this case is of great value for protective gear practice as a means of detecting or discriminating between earth fault conditions. The residual voltage of a system is measured by connecting the primary windings of a three-phase v.t. between the three phases and earth, and connecting the secondary windings in series or 'open delta'. The residual voltage is three times the zero sequence voltage. To measure this component it is necessary for a zero sequence flux to be set up in the v.t., and for this to be

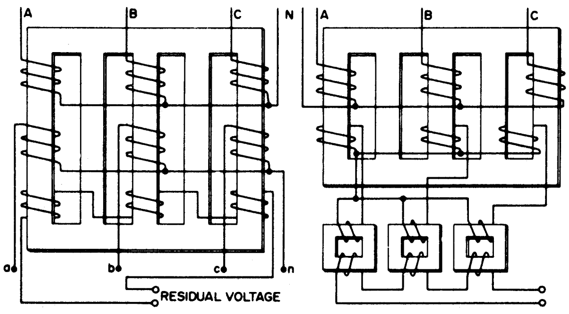


Figure 35.7 Voltage transformers with residual-voltage windings

possible there must be a return path for the resultant summated flux: the v.t. core must have unwound limbs linking the yokes (*Figure 35.7*). If three single-phase units are used (as is common for e.h.v. systems), each phase unit has a core with a closed magnetic circuit, so that the above consideration does not arise.

An alternative, avoiding the cost of a HV voltage transformer, is to use the secondary voltages from a.v.t. on the LV side of a power transformer, the voltage drops in which are compensated by the addition or subtraction of voltages developed by c.t.s in the delta of the power transformer. The v.t.s and c.t.s must be provided with tappings if the power transformer is equipped with tap-changing gear, and arranged for automatic selection with the main tappings.

For high-voltage systems the *capacitor voltage divider* gives a cheaper (but less accurate) device than its electromagnetic counterpart. A typical divider for 400 kV has a total capacitance of 1500–2000 pF, with about 34 000 pF between the tapping point and earth, to give about 13.5 kV across the primary of the intermediate transformer, the secondary of which gives 63.5 V (phase to neutral).

35.1.3 Relays

Relays may be classified according to the number of current inputs and their operating function. Types and constructional forms are available as follows:

(1) attracted armature
(2) moving coil
(3) induction
(4) thermal
(5) motor-operated
(6) mechanical
(7) magnetic amplifier
(8) thermionic
(9) semiconductor
(10) photoelectric
(11) digital

Of these main group types, (7)–(11) are commonly known as 'static' relays.

35.1.3.1 Single-input relays

In the simple *repeater* type a small signal can be multiplied by 10^3 or more to operate one or more secondary devices, e.g. the trip coil of a circuit-breaker. Operation is rapid (within 20 ms), although a time delay can be incorporated. The *magnitude indicator* type operates instantaneously (or after a fixed time delay) when the magnitude of the signal exceeds or falls below a specified value. In the *time-dependent* relay the operating time depends on the magnitude of the signal; the most usual characteristic gives an operating time inversely proportional to the magnitude.

35.1.3.2 Double-input relays

In the *amplitude comparator* one input, I_1, tends to operate the relay, while the other, I_2, tends to restrain it. Operation takes place when I_1/I_2 is less than the specified value. When drawn on a complex plane R–X, the characteristic is a circle (*Figure 35.8*(*a*)). Operation is independent of the phase angle between the currents and occurs within the shaded area.

In the *phase comparator* operation takes place when the phase angle between the currents lies within specified limits (*Figure 35.8*(*b*)).

35.1.3.3 Multiple-input relays

By using more than two inputs, various forms of operating zone can be obtained, as required for distance feeder protection.

35.1.3.4 Types of relay

Attracted armature This single-input relay comprises a small coil carrying the signal current: it attracts a magnetic armature with one or more pairs of contacts which control the secondary circuits, and is suitable for both d.c. and a.c. operation. The setting may be based on minimum pick-up or maximum drop-out current, with ranges usually selected by tappings to secure a constant value of operating m.m.f. Tapping changes the impedance and, therefore, the burden on associated c.t.s. The speed of operation is a function of the m.m.f., the length of armature travel and the inertia of the moving parts. The latter may be partly offset by counterweighting. When the relay coil is intermittently energised, it may be uprated to increase the operating speed: a set of contacts may be provided which, when energised by the operation of the initiating relay contact, gives a holding effect. The current in the coils of tripping relays is then usually broken by an auxiliary switch on the circuit-breaker. Magnetic sticking of the armature in attracted-armature relays is prevented by stops of non-magnetic inserts on the working face.

Rotary This is a rotary version of the foregoing. The armature is usually mounted on a vertical spindle, the rotating parts are light and well-balanced, and the relay is sensitive at a relatively low burden. Adjustment and accurate setting is by torsion head and helical spring control. In some types the armature carries a light silver contact, the coil operated contacts being energised when the silver contacts close. The relay can be made to operate with a few milliamperes, and both linear and rotary forms of the attracted-armature relay can be arranged for double-input working.

Induction The time-dependent induction protective relay is probably the most widely used of all. It employs essentially

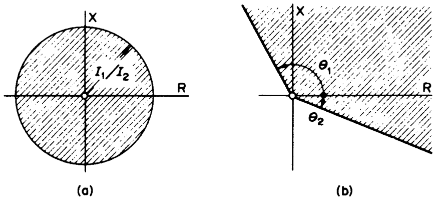


Figure 35.8 Comparator relay characteristics: (a) amplitude; (b) phase

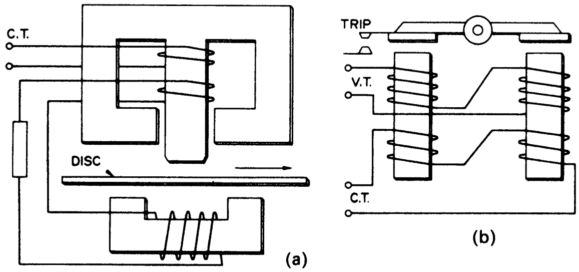


Figure 35.9 (a) Induction and (b) beam relays

the same construction as an integrating a.c. energy meter. For a single-input relay (*Figure 35.9*(*a*)) torque is produced by a phase difference between the flux of the main coil on the central limb of the upper magnet and that in the lower (secondary) magnet, an effect secured by impedance compensation in the secondary. The torque is proportional to the square of the exciting current and is unidirectional. The speed of the disc may be governed by a control spring and a permanent-magnet brake. Slots are cut in the disc to modify the operating time–current characteristic, and the shape of this curve at higher currents can be made almost flat if the magnetic circuit is arranged to saturate at two or three times the setting.

The driving disc usually has a vertical spindle, but one make of relay has a horizontal drum driven about its axis, the time–current characteristic being similar to that of the disc type.

The *single-input* relay is widely used for overcurrent protection. By feeding a separate signal to the secondary coil the relay becomes a *double-input* device with a torque proportional to $I_1 I_2$ sin α, where α is the phase angle between the currents. In this form it can be given directional properties, and is also applicable to certain forms of distance protection.

Directional Directional relays operate in a manner similar to that of an electrodynamic wattmeter or, more commonly, an induction energy meter. In the latter the torque on the disc is proportional to VI cos β, where β is the phase angle between the voltage and current inputs.

The maximum-torque angle is the angle by which the current applied to the relay must be displaced from the applied voltage to produce maximum torque. Although the relay may be inherently wattmetric, its characteristic can be varied by the addition of phase shifting components to give maximum torque at the required phase angle. Several different connections have been used, and examination of the suitability of each involves determining the limiting conditions of the voltage and current applied to each phase element of the relay for all fault conditions, taking into account the possible range of source and line impedances.

Beam A balanced beam with two electromagnets can be used as a simple double-input device and has been widely applied to distance protection (*Figure 35.9*(*b*)). It can also serve as a directional relay.

Definite time Devices employing a clockwork escapement released by the operation of an attracted-armature relay, or an excapement driven by a disc or drum, are available. More usually, however, induction relays are used because of the long periods of inactivity in service.

35.1.4 Solid-state equipment

Thermionic valve equipment has been in use for a half-century in carrier current protective signalling. The advent of solid-state devices (transistors and thyristors) has made possible alternatives to electromagnetic relays.

The exploitation of integrated circuits and microprocessors has made a major impact on protection in recent years.

Solid-state equipment differs markedly from conventional electromagnetic equipment in several ways, particularly in the reduction in overall size.

The flexibility and characteristics enable schemes to be devised that are difficult to achieve conventionally, such as the distance relay with a quadrilateral characteristic (*Figure 35.15*(*f*)). For most applications the transistor has the advantage of high speed, high sensitivity and simple amplifying and switching properties. Where heavier currents are necessary, e.g. in a tripping circuit, the thyristor is required. Solid-state devices can, if desired, readily be associated with electromagnetic relays if the time delays (10–20 ms) associated with the latter can be accepted; however, the lighter and faster (1–2 ms) reed relay is applicable in many cases.

Components are subject to damage by moisture, but encapsulation can obviate the trouble.

Static components have some resilience and very low mass; they can withstand mechanical shock and vibration, significantly reducing risk of damage during transport and erection.

All components, are sensitive to transient voltages, even of only a few microseconds duration. Such transients arise from switching, and can be injected into the protective equipment directly through current and voltage transformers, or by electric or magnetic field induction. Equipment must be impulse tested, e.g. by a 5 kV 1/50 μs impulse wave.

The burden imposed by static relay equipment on c.t.s is much less than that of comparable electromagnetic equipment. As the cost of instrument transformers is 20–40% of that of the complete protective scheme, smaller and improved types are being developed which may offset the higher cost of the static relay itself.

Even with conventional equipment, the use of 1 A secondary currents (instead of 5 A) is sometimes desirable, especially where long interconnecting leads are necessary. With static equipment even smaller currents can be used. The ultimate in this respect is a very small line mounted high-voltage current transformer feeding a laser system; the electromagnetic power (about 2 W) is converted to light power and transmitted to earth potential equipment by light-guides, the resulting power (a few microwatts) then being amplified and fed to static relay equipment.

The range of solid-state relays now available covers many applications hitherto dealt with by electromechanical devices. The static equivalent generally offers greater flexibility in the design to meet given protection functions. Typical characteristics are shown in *Figure 35.10* for static protection, which needs no auxiliary supply and provides a high reliability and accuracy.

35.1.5 Overcurrent protection

The most common overcurrent relay is the induction type, typical characteristics of which are given in *Figure 35.11*. The standard characteristic follows no simple law, but the

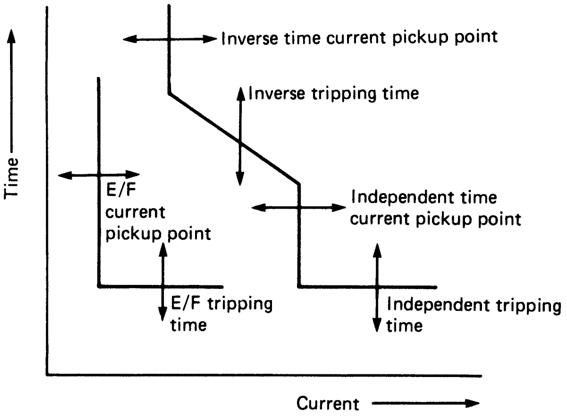


Figure 35.10 Static protection unit characteristics

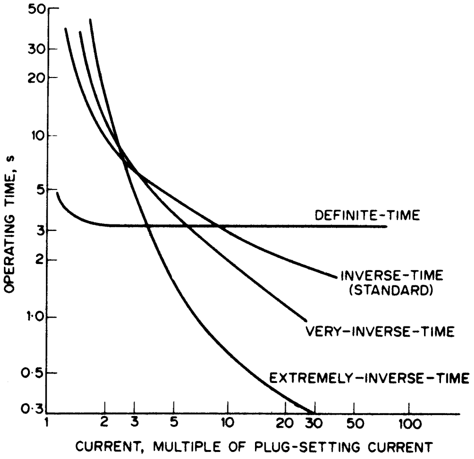


Figure 35.11 Overcurrent relay characteristics

very-inverse-time characteristic is $(I-1)t=k$; and the *extremely-inverse-time* law is $(I^2-1)t=k$.

The relay has two adjustable settings. The *plug setting* (current setting) comprises a tapped winding for the current coil, the tappings being brought out to a plug bridge on which, by insertion of the plug, current settings of 50, 75, 100, 125, 150, 175 and 200% of the full-load output of the current transformer (5 A) can be obtained; removal of the plug automatically short circuits the current transformer terminals. A *time setting* is achieved by varying the amount of travel of the contacts, giving time multipliers between 1.0 and 0.1. The times on *Figure 35.11* are for a time setting of 1.0 and the currents are multiples of the plug setting current.

35.1.5.1 Discrimination

Graded time lags On simple radial distribution systems, discrimination by graded time lags is effective, the lags of the relays being set successively longer for relays nearer to the supply source. The minimum time difference at all parts of the relay characteristic may be estimated as: error of relay 1 (near to fault), 0.1 s; circuit-breaker operating time, 0.1 s; over-shoot of relay 2 (nearer to supply), 0.04 s; error of relay 2, 0.1 s; total, 0.34 s. The greater low-current times of the extremely inverse time characteristic may be suitable for grading with fuses, and also are less likely to cause tripping on switching-in after a supply interruption, when there may be heavy overloads resulting from plant left connected.

Graded time lags and directional relays If the overcurrent relays embody a directional feature (DOC), discrimination on more complicated networks can be obtained. *Figure 35.12(a)* shows a *ring main*: the arrows indicate the current direction for which the relays will operate, and the figures show typical time settings. The directional feature ensures that time grading can be obtained in each direction around the ring.

For the two *parallel transformers* in *Figure 35.12(b)* directional overcurrent relays are employed at the transformer ends on the LV side to detect phase faults on the line. Thus, a fault at F would be cleared by relays at C, while those at D would restrain. For a fault on the LV system both relays at C and D would restrain.

A more complicated network is shown in *Figure 35.13*. Suppose that the interconnector and feeder 6 are open and that there is a fault F on feeder 5. Fault current will then flow from the supply infeed direction into feeder 5 through 2. Now if feeder 2 is provided at each end with current

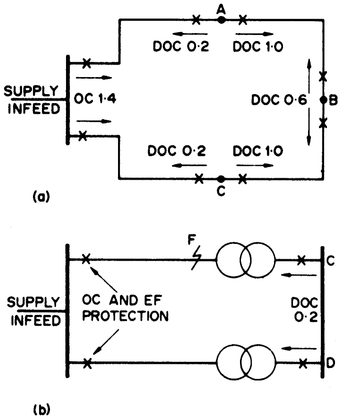


Figure 35.12 Directional overcurrent protection

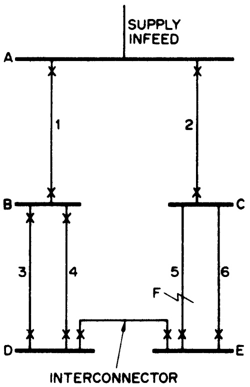


Figure 35.13 Network time-grading

operated relays having definite time delay, and if similar relays for feeder 5 are set lower than those for 2, then the former will operate first and clear the fault, after which the relays on 2 reset and leave the feeder still connected to the system. It will be apparent that a fault on feeder 2 will be cleared in a longer time than a fault on feeder 5.

Suppose now that feeder 6 is closed and that its relays are set the same as those of 5; then for fault F on 5 the relays on 5 and 6 both operate to trip both feeders. This unwanted result is avoided if the relays at the E end of 5 and 6 are set to a shorter time than those at the C end and provided with a directional feature; then for fault F the first operation will be the feeder 5 relays at E, and the second will be those at C, which are set in time to discriminate with the relays for feeder 2. Such a scheme is the simplest form of directional overcurrent protection, and at each point the relays provided may be (1) one overcurrent (o.c.) relay in each phase together with an earth fault (e.f.) relay, or (2) two o.c. relays and one e.f. relay, or (3) three o.c. relays only, where the system is solidly earthed. In the case of earthing through a resistor, the o.c. and e.f. relays would require different settings.

Consider the effect of closing the interconnector with a fault F on feeder 5. If the two radial systems are identical in length, feeder size and relay time settings, then, although fault current will divide between feeders 1 and 2, the first relays to operate will be those at E on feeder 5. However, when this occurs the fault is fed through 2, 6 and the interconnector; therefore the feeder 6 relays at E will operate first if their operating time plus that of the feeder 5 relays at E is less than the operating time of the feeder 5 relays at C. Furthermore, the interconnector cannot be proved with DOC relays satisfactorily for its own protection unless it is regarded as a 'loose link', in which case they would have to be set for rapid tripping to cut off the second infeed to a fault elsewhere on the system. If it is required that the interconnector should not trip, it must be provided with balanced protection and the o.c. relays set high for back-up.

A disadvantage of time graded systems is that the relays on feeders nearest to the supply point must have the longest operating times. As fault currents at such points are greatest, this is a drawback in the case of phase faults, though less so for single phase-earth faults, where fault current may be limited by earthing resistors.

Locking signals To avoid the long time lags associated with overcurrent equipment, a scheme involving transmission of *locking signals* from the remote end of a line may be employed. Consider *Figure 35.14*, in which only the locking signal pilot and relay contact circuits are shown: if the o.c. relays at, say, the left-hand end detect a fault current, they will operate to close the trip-coil circuit of the circuit-breaker at that end; if, however, the directional relay at the right-hand end detects a fault current going *out* of the feeder (indicating a fault on another section), it initiates a signal on the pilot circuit which energises the lock-out relays and prevents tripping at both ends. A failure of the pilot circuit will cause the sending end to trip on a through fault but will not interfere with operation on an internal fault.

Either private pilot wires, telecommunication pilots or carrier current over the line itself may be used, the equipment for the latter being similar to that for carrier current phase comparison.

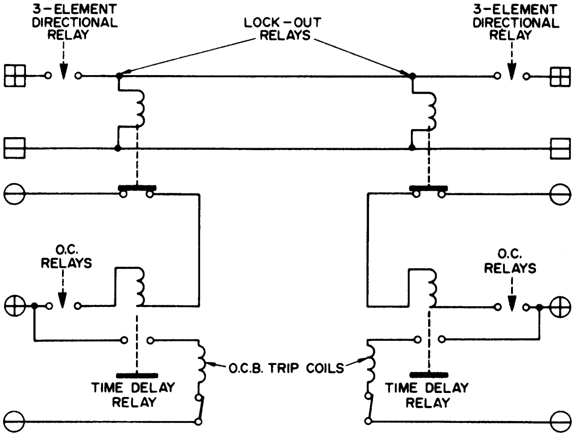


Figure 35.14 Locking signal protection using pilot wires

35.1.6 Fuses

The simplest overcurrent protection device is the fuse, which is used in vast numbers throughout the circuit operating voltage range 415 V to as high as 66 kV. The basic principle involves connecting a fuse directly in series with the protected equipment so that, when a given current is exceeded a metallic fuse 'element(s)' melts and thereby breaks the circuit. In this way, fuses both detect and directly isolate faulted equipment from the network.

The term 'fuse' is used in national and international standards to describe a complete assembly. In its simplest form, this consists of a piece of metal wire connected between two terminals on a suitable support; and at its most complex as a cartridge fuse-link mounted in a carrier and fuse base.

Modern cartridge fuse-links contain fusible elements mounted in rigid housings of insulating material. The housings are filled with suitable exothermal and arc-quenching powders, such as silica, and they are sealed by metal end-caps which carry the conducting tags or end connections. A typical fuse-link is shown in *Figure 35.15*. The metal parts, other than the fusible elements, are invariably of copper, brass, steel or composites and they must be capable of operating under the exacting thermal, mechanical and electrical conditions which may arise in service.

35.1.6.1 Fuse technology

A fuse must be able to carry normal load currents and even transient overloads (and the thermal cycling which accompanies them) for a service life of at least 20 years, without

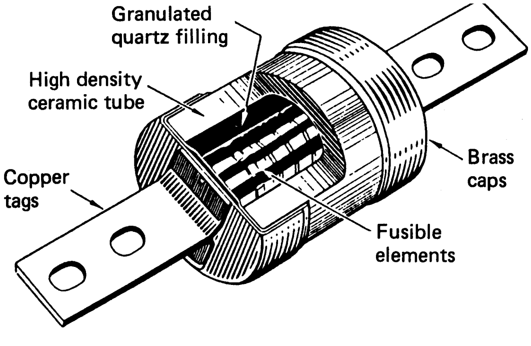


Figure 35.15 HRC fuse-link

any change of state that might affect its electrical performance. This property of 'non-deterioration' requires that the fusible element be both thermally and chemically compatible with the ambient media. It must also respond thermally to overcurrents by melting and subsequently interrupting its circuits.

The melting of an element is followed by a period of arcing during which the electrical energy input can be very high, its magnitude and the duration of arcing being dependent on the protected circuit. Successful fault interruption implies that the arcing is wholly contained within the fuse-link and the level at which this can be achieved is termed the breaking or rupturing capacity of the fuse-link.

The operating time of a fuse-link varies inversely with the level of an overcurrent and discrimination is obtained in networks by choosing fuses with the necessary time/current characteristics and current ratings.

The time/current characteristic is determined by the design of the fuse itself, in particular but not exclusively, the fuse material and the physical geometry of the fuse link(s). In practice, the fuse/current characteristic is chosen to ensure adequate discrimination with other fuses and/or overcurrent relaying devices around the network. *Figure 35.16* for example shows how, in general terms, a fuselink 'f' situated between protective devices 'u' on the source side and 'd' on the load side could be arranged to provide discriminative tripping on a radial feeder.

35.1.6.2 Fuse links with short operating times

All modern high-breaking-capacity fuse-links contain elements, usually with restrictions, of small cross-sectional areas connected between relatively massive end connections which act as heat sinks. To obtain rapid operation this principle is employed to a high degree. The extent to which the mass of the heat sink can be increased while reducing the length of the relatively thin element is determined by the requirement that the fuse should withstand the system voltage after the current has been interrupted i.e. it must not restrike. Considerable ingenuity has reconciled these two mutually incompatible requirements.

35.1.6.3 The M effect

The M effect, deriving from an exposition by Metcalf, refers to exploiting the thermal reactions of dissimilar metals in the control of time/current characteristics. The thermally most stable fuse element is a simple homogeneous metal. Such an element provides the highest degree of non-deterioration and reliability with adequate breaking capacity at higher overcurrents, but it may be insensitive to lower overcurrents. A lower melting temperature metal with high resistivity and, therefore, greater thermal mass, can be made to respond more sensitively to lower overcurrents but may be unreliable at higher currents.

The M effect is a means by which these extremes can be combined to produce a desired characteristic, but it needs to be used with care in design to avoid compromising non-deterioration properties. An element incorporating the M effect is shown in *Figure 35.17(a)*.

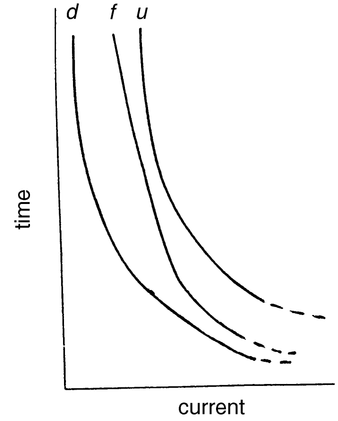


Figure 35.16 Use of fuse time/current characteristic to provide discrimination on radial feeder

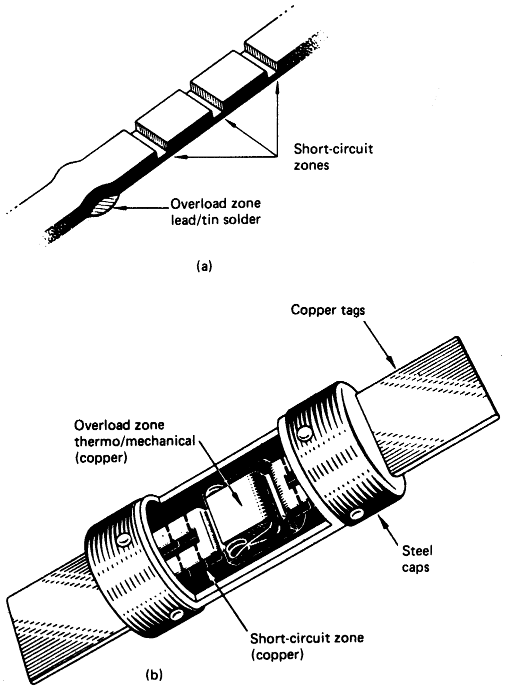

Short-circuit zones
Overload zone lead/tin solder
(a)
Copper tags
Overload zone thermo/mechanical (copper)
Steel caps
Short-circuit zone (copper)
(b)

Figure 35.17 (a) Fuse element (English Electric); (b) Dual-element fuse

35.1.6.4 Composite or dual-element fuses

Satisfactory operation throughout the overcurrent and short-circuit ranges is sometimes obtained in the same package by combining what are, in effect, two fuses connected in series in the same cartridge (*Figure 35.17(b)*). Typical of these is the so-called dual-element design common in the USA. The short-circuit zone is similar to the homogeneous element used in single-purpose HRC fuses. The overload zone may take the form of a massive slug of low-melting-point alloy, or some electromechanical device, e.g. two copper plates soldered together and stressed by a spring, so that when the solder melts the plate springs apart to interrupt the current.

35.1.7 Earth leakage protection

A current-to-earth on an HV system invariably indicates the presence of a fault: if large enough, it will operate the o.c.

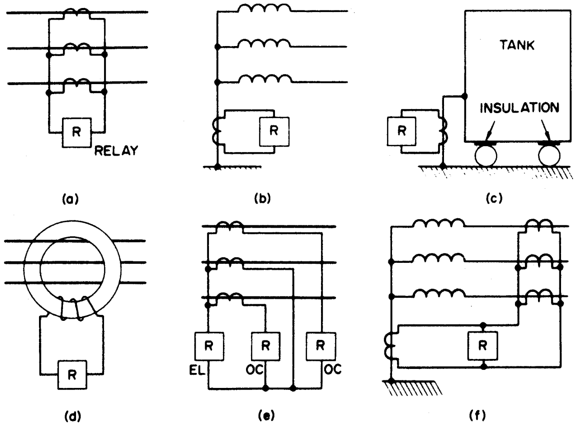


Figure 35.18 Earth leakage protective schemes: (a) residual c.t.s; (b) neutral c.t.s; (c) frame leakage; (d) core balance; (e) overcurrent and earth leakage; (f) balanced (restricted earth leakage)

relays, but it is necessary to detect earth currents even if they are limited by resistance earthing of the neutral point. Separate earth leakage (e.l.) protective equipment is thus usually desirable. The arrangements shown in *Figure 35.18* are available, usually with induction relays, or linear or rotary attracted-armature types for high sensitivity. (An overload conductor falling to earth of dry soil may produce no more than 5 A of earth fault current, but the associated potential field may be dangerous to human and animal life.)

Schemes (*a*)–(*e*) in *Figure 35.18* have no discrimination and operate for any fault beyond the equipment. If, however, the e.l. relays are given a time lag, any serious fault can be cleared discriminatively by o.c. or other protection. In (*c*) the transformer tank, switchgear frame or metallic bus-bar covering must be lightly insulated from earth: an insulation resistance of a few ohms is adequate, but it must never be short-circuited, e.g. by a length of conduit or a metal tool leant against it. Scheme (*d*) employs a 'core balance' c.t. in which the core flux is produced only by earth leakage currents. Scheme (*a*) is commonly extended to include two o.c. relays, as in (*e*).

Discrimination can be provided by balancing the e.l. currents at two points in the network. A common application is for the protection of the star-connected winding of a transformer (*f*) in which the e.l. relay will operate only for internal faults in the transformer.

A directional feature can be given to an e.l. relay, the second input being usually provided from open-delta voltage transformers.

35.1.8 Balanced (differential) protection

Balanced protection is based on the principle that the current entering a sound circuit is equal to the current leaving it, whereas, if faulted, the detectable current difference can be used to trip the appropriate circuit-breakers.

35.1.8.1 Circulating current schemes

The c.t.s in circulating current schemes operate with low-impedance secondary burdens, provided that the pilot wires are short, so that identical c.t. characteristics at the two ends are not difficult to achieve even under conditions of heavy fault; but a large d.c. fault current component may cause dissimilarity and spurious unbalance.

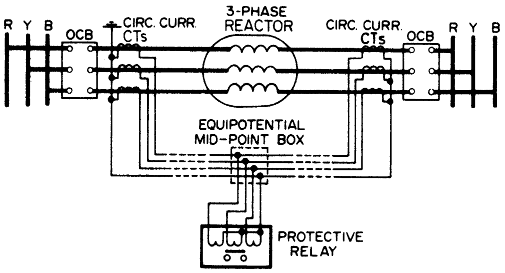


Figure 35.19 Circulating current protection on a reactor

The relays must be located at the electrical mid-point of the pilot wires. This introduces no difficulty where the ends of the protected zone are not far apart, but is less suitable for feeders (although some are in fact so protected). Short pilot circuits obviate the need for summation transformers to reduce their number, and the arrangement of *Figure 35.19* is typical. Circulating current protection is widely used for reactors, transformers, generators and bus-bars. In appropriate circumstances it has the advantages of simplicity, sensitivity and rapid response (e.g. 10–20 ms).

35.1.8.2 Opposed voltage schemes

As relays can readily be installed at each end of the protected zone, an opposed voltage scheme is suitable for feeder protection. Summation transformers or other means reduce pilot wires to one pair, but even so the pilot problems limit the practicable feeder length.

Per kilometre loop, the resistance of the typical pilot circuit with 7/0.075 mm conductors is 12.5 Ω and the capacitance is about 0.1 μF. The c.t.s operate, even on internal fault, with a high impedance burden; on a through-fault they are virtually on open circuit, and unless precautions are taken, the voltage between pilot wires can reach 1 kV, making magnetic balance of the transformers difficult and also giving rise in the pilot wires to capacitance currents which are likely to cause maloperation. Of the many schemes devised to overcome these difficulties, that mentioned below is typical.

This scheme employs induction relays so arranged that only two pilot wires are required, their capacitance currents being used to give a restraining effect. The pilot wire voltage is limited to about 130 V. In *Figure 35.20* consider a fault F between the R and Y phases: the currents in section (1) of the relay primary windings 11 and 11a induce e.m.f.s in

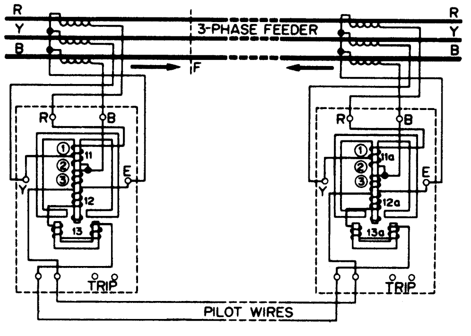


Figure 35.20 Biased balanced voltage protection

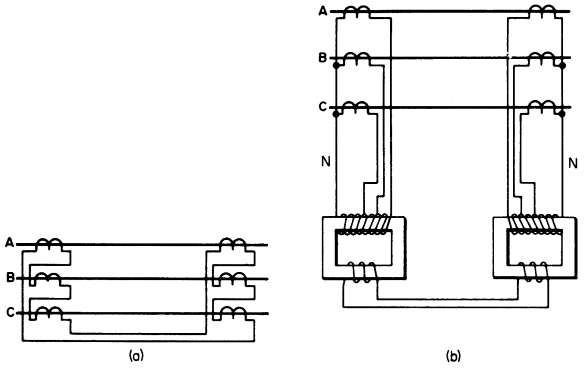


Figure 35.21 Balanced protection principles: (a) discriminating delta; (b) summation transformers

12 and 12a which, being now additive, circulate a current in the operating coils 13 and 13a and in the two pilot wires. Both upper and lower magnets are energised, and if the fault current exceeds the scale setting, the relays operate to trip their associated circuit-breakers. A fault between phases Y and B energises section (2) of windings 11 and 11a, while a fault between phases R and B energises both sections (1) and (2), the fault setting being one-half of that in the former cases.

In the event of an earth fault on line R, the resultant secondary current from the phase R c.t. flows through sections (1), (2) and (3) of windings 11 and 11a. A Y phase earth fault energises sections (2) and (3); and a B phase fault, only section (3).

Pilot wires It is often desired, especially for feeder protection, to reduce the number of pilot wires to two. One method is the *discriminating delta* connection (*Figure 35.21*(*a*)) in which the c.t.s have different ratios (e.g. N, 3 N/4 and N/2). A more common method employs a *summation transformer* (*b*). If the primary turns of the summation transformer are in the ratios 1 between A and B, 1 between B and C, and n between C and N, then the relative sensitivities for various types of fault are in the following relative proportions as secondary outputs:

Fault	A–N	B–N	C–N	A–B	B–C	C–A	A–B–C
Proportion	$2+n$	$1+n$	n	1	1	2	$\sqrt{3}$

Stability ratio High *sensitivity* is desirable to ensure operation of the relays on minor or incipient faults, while high *stability* is required to prevent operation on heavy through-faults. These requirements are in conflict and it is usual to restrain ('bias') the relays by an additional winding carrying a current proportional to the fault current, thus desensitising the response to heavy fault conditions. The ratio (maximum through-fault stable current)/(minimum fault operating current) can thus be made as great as 100.

Distance (impedance) protection This type of protection, used chiefly for overhead lines, involves the measurement of the impedance of the circuit as 'seen' from the relay location; if this impedance falls below a specified value, a fault is present. The impedance observed is approximately proportional to the distance of the fault from the relay and discrimination is obtained by introducing time delays for the more distant faults. Two schemes are available, *impedance time* and *steeped time,* the latter being more widely used on lines above 33 kV, on account of the shorter operating times obtainable.

35.1.8.3 Impedance time scheme

Relays having a directional characteristic give an operating time proportional to the impedance presented, i.e. proportional to the distance of the fault from the relay. By locating such relays at each circuit-breaker on a succession of feeders the characteristics of *Figure 35.22*(*a*) are obtained. It can be seen that should a relay, e.g. that at B, fail to operate, it will be backed up by that at A after a time interval. With feeders of different lengths the slopes of the characteristics must be adjusted to give correct discrimination; thus, the relay at A must have characteristic X (dotted) rather than Y, as the latter would give incorrect back-up discrimination for a fault near the end of section CD. The relays cater for fault currents flowing from left to right; a similar set would be required to cater for fault currents from right to left.

35.1.8.4 Stepped time scheme

Three separate relays are required for each direction of fault current at every circuit-breaker location. The relays are of the induction or (more rarely) of the beam type. In *Figure 35.22*(*b*) relay 1 of the group at circuit-breaker A is set to give 'instantaneous' operation for faults occurring within the nearest 80% of feeder AB, i.e. over distance X_1. Relay 2 operates, after a time lag, for faults occurring up to a point just beyond circuit-breaker B; it thus acts as a back-up for relay A over the distance X_1. Relay 3 operates with a still longer time lag for faults beyond the cut-off point of relay 2; it thus acts as a further back-up for relay A, and also for relay 2 over the last 20% of feeder AB.

35.1.8.5 Fault impedance

The impedance 'seen' by a relay depends upon the type of fault. Two sets of relays, differently connected to the system, are required (although a single set with a switching device is also practicable).

Earth fault For a simple earth fault fed with current I_e and phase-to-neutral voltage V_a of the faulted phase, the earth loop impedance presented to the relay is

$$V_a/I_e = Z_1 + Z_e = Z_1(1 + k)$$

where Z_1 is the normal positive-sequence line impedance from the relay to the fault, Z_e is the corresponding impedance of the earth path, and $k = Z_e/Z_1$.

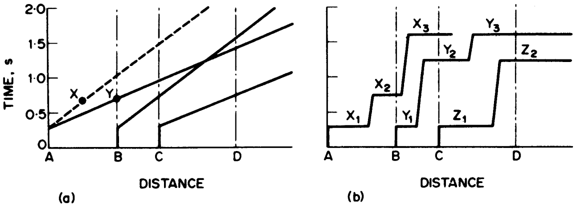


Figure 35.22 Distance (impedance) protection principles

Phase–phase fault The relay is fed with the phase–phase voltage and with the difference ($I_a - I_b$) between the faulted phase currents, so that the impedance is

$$V_{ab}/(I_a - I_b) = V_a/I_a = Z_1$$

Three-phase fault Relays connected as for the phase–phase fault see the same impedance Z_1, because

$$V_{ab}/\sqrt{3}I_a = V_a/I_a = Z_1$$

and consequently operate correctly with the same setting.

Earth current compensation If, as is common, the system is earthed at more than one point, a proportion of the earth fault current will return by a path other than that at which the relays are located. The impedance seen is thus in error: it will be too low. This can be corrected in two ways.

Sound-phase compensation A fraction $p = Z_e/(Z_1 + Z_e)$ of the sound-phase currents is added to the faulted-phase current, as shown in *Figure 35.23(a)*. Typical values for a 132 kV line are $Z_1 = 0.44\ \Omega$/km and $Z_e = 0.21\ \Omega$/km, so that $p = 0.34$.

Residual compensation A fraction q of the residual current I_r is added to the relay operating current, as in *Figure 35.23(b)*. If $q = Z_e/Z_1$, the measured impedance is Z_1, and the relays must be set for Z_1 instead of for Z_e to give correct operation. With the typical values above, $q = 0.52$.

Arc resistance If the fault has an arc resistance, the relay will 'see' the fault as located too far away, and errors in discrimination may result. Typical values of fault resistance lie between 0.5 and 3 Ω, the higher values for lower currents.

Power swing If a power swing occurs, i.e. phase swinging of the terminal voltages of the section of the system concerned, the impedance seen by a relay may fall to a low value even if there is no fault in or near its protected zone.

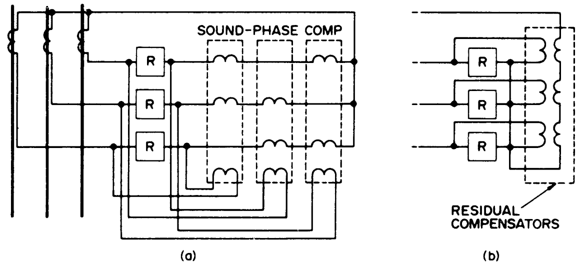


Figure 35.23 Sound-phase and residual compensation

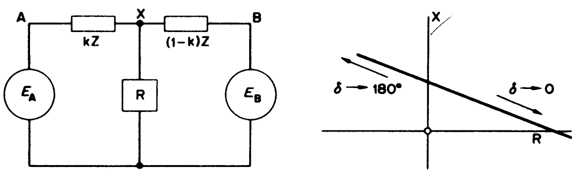


Figure 35.24 Impedance during power swing

In *Figure 35.24* the impedance seen by the relay at X and looking towards B is given by

$$Z_p = Z[A/(A-1) - k]⇐$$

where $A = (E_A/E_B)\angle\delta$, with δςrepresenting the phase angle between E_A and E_B, and Z the total impedance of the line. The expression gives approximately the linear impedance locus shown, part of which around δς= 90°⇐may well fall within the operating zone of the relay and cause false tripping unless special precautions are taken.

35.1.8.6 Relay characteristics

Characteristics for various relay constructions are given in *Figure 35.25*, the relays operating if the impedances seen by them fall within the shaded area. The impedance for which a relay is required to operate is given by OZ_1, OZ_2 or OZ_3 in *Figure 35.25*, so that this value should lie within the shaded area and any other impedance seen by the relay (due to load, arc resistance, power swings, etc.) should lie outside it. The quadrilateral characteristic (*f*) is desirable, but it can be obtained only by electronic means.

Typical arrangements for the three steps of a stepped-time scheme are also shown in *Figure 35.25*. At (*g*) the directional feature is given by the 'mho' relays used for the first two stages, while at (*h*) the mho relay for the third stage is used as a starting element and also prevents the reactance relays operating under load conditions.

Inaccurate operation can result if the relay voltages are low, owing to a high source/line impedance ratio, e.g. 30/1 or more. A polarising winding on the relay fed (1) through a memory circuit so that it retains pre-fault voltage or (2) from the sound phases through a phase shifting circuit, may be used. Method (2) is not effective for three-phase faults.

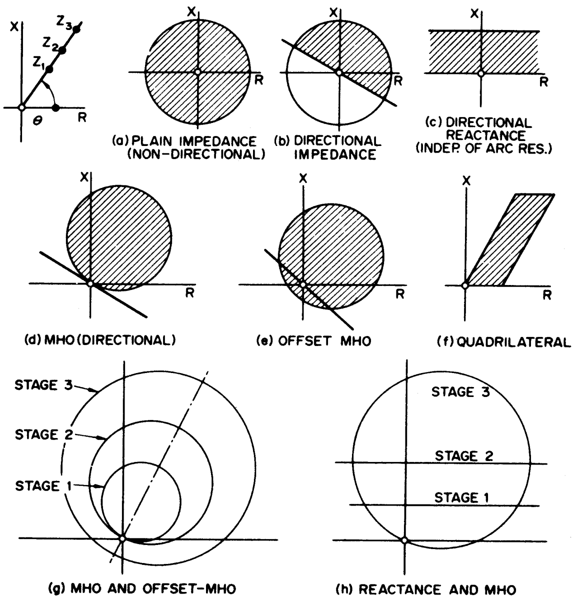


Figure 35.25 Distance relay characteristics

35.1.8.7 *Accelerated distance protection*

It can be seen from *Figure 35.22*(*b*) that faults near the remote end of the protected section AB will only be cleared after an appreciable time lag. Such a fault will, however, be cleared 'instantaneously' by the right-to-left equipment at B: and if this equipment is also made to transmit a signal to A over a pilot circuit (usually carrier over the line), it can initiate the immediate tripping of the circuit breaker at A, thus giving almost instantaneous protection over the whole line.

35.1.9 Miscellaneous equipment

35.1.9.1 *Negative sequence*

Any fault other than a symmetrical three-phase fault develops negative sequence currents, so that a network responsive to n.p.s. but not to p.p.s. components will indicate the presence of a fault and can be made to operate a relay. Two such circuits are shown in *Figure 35.26*. The cross-connection of the current transformers in (*a*) eliminates the z.p.s. components: the latter are earth leakage currents and are detected in other ways. The n.p.s. schemes illustrated employ impedance elements of resistance R and inductive impedance $Z = R\angle 60°$ but capacitive impedances $Z = R\angle -60°$ give the same results if the positions of R and Z are interchanged.

35.1.9.2 *Neutral displacement*

Displacement of the neutral point potential from its normal earth potential is indicative of a fault condition, and may be used to initiate tripping.

35.1.9.3 *Buchholz relay*

The Buchholz relay is used for the protection of large oil immersed transformers or shunt reactors with oil conservators, and is fitted in the pipe from the tank to the conservator. Any arcing fault causes the oil to decompose, generating gas which passes up the pipe to the conservator and is trapped in the relay. In the case of a large fault a bulk displacement of the oil takes place. In a two-float relay the upper float responds to the slow accumulation of gas due to the mild or incipient faults, while the lower float is deflected by the oil surge caused by a major fault. The floats control contacts—in the first case, to give an alarm; in the second case to isolate the unit.

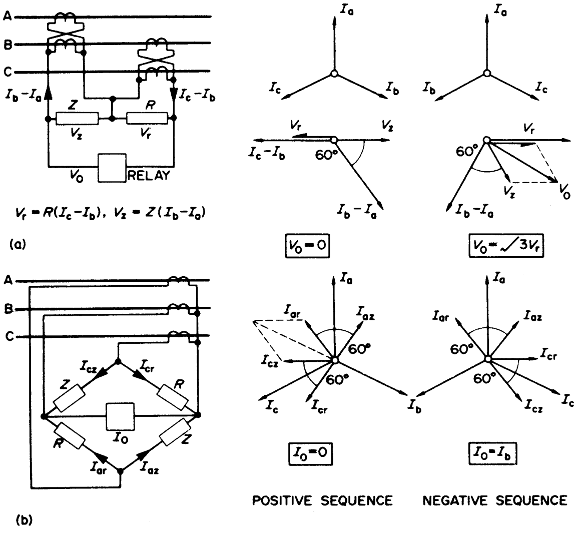


Figure 35.26 Negative-phase-sequence protection: (a) high-impedance relay; (b) low-impedance relay

Such relays also incorporate a petcock at the top for the removal of the gas; its subsequent analysis can show the origin and severity of the internal fault.

35.1.9.4 *Tripping of circuit-breakers*

Direct-operated trip coils The circuit-breaker trip coil may be energised directly from the current in the main circuit for lower voltage circuit-breakers, or from a current transformer with the trip coil shunted by a fuse which blows on overcurrent to cause tripping. With c.t.s it is possible to use the residual current to operate the tripping at a much lower earth fault current than for overcurrents. The advent of solid-state relays enables much better fault discrimination to be achieved, and also obviates the need for external auxiliary supplies (usually d.c.) for tripping.

35.1.9.5 *D.c. tripping and operating circuits*

The d.c. circuits brought into operation by protective relays are of importance, as they are responsible for circuit-breaker actuation.

Single tripping The tripping of one local circuit-breaker ('unit tripping') by the operation of relays is applicable to overcurrent or balanced protection of feeders having relays at each end. The essentials are shown in *Figure 35.27*(*a*); it includes a 'healthy trip circuit' indicating lamp with external resistance of such value that, if the lamp is short-circuited, the current is less than that required for tripping. The lamp also serves as a 'circuit-breaker closed' indicator, and proves the continuity of the trip circuit when the breaker is closed.

Intertripping By this is meant the necessary tripping of circuit-breakers identified with the unit protected; for example, the tripping between HV and LV breakers on a

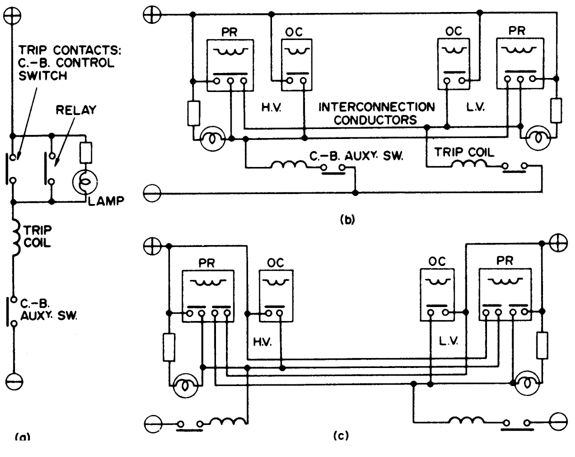


Figure 35.27 Tripping circuits: (a) single circuit breaker; (b) common d.c. supply; and (c) separate d.c. supplies for intertripping of transformer HV and LV circuit-breakers

transformer with (1) overall circulating current protection or (2) balanced earth leakage and overcurrent protection on each side, it being necessary to trip *both* breakers for a fault on either side. Intertripping may be local and relatively straightforward, or it may require to be performed on circuit-breakers at separated points in the network.

Local intertripping When two circuit-breakers are involved, use can be made of a standard 'three-point' scheme, either with a three-point relay, or by the addition of an interposing relay which gives a positive tripping supply to two independent circuits. *Figure 35.27*(*b*) shows an arrangement with three-point relays and a common d.c. supply, and *Figure 35.27*(*c*) the modification for separate supplies. In both cases the overcurrent relays, whether for 'back-up' or phase fault protection, trip their respective breakers only. Buchholz protectors or high-temperature relays, where used, should be connected so as to intertrip. When transformers are feeding distribution networks, three overcurrent relays on the HV side only with intertrip to the LV breaker are sometimes used, i.e. with no overcurrent relays on the LV side. The disadvantage of this arrangement is apparent, for the tripping of the LV breaker depends solely on the intertripping wire or cable. Also, the overcurrent relays are inoperative with the transformer charged from the l.v. side only, and with balanced earth leakage on each winding there would be no phase fault protection, obviously a dangerous condition. It is considered better practice to install overcurrent relays on each breaker so that, in the event of damage or bad connection on the interconnecting pilots, the 'back-up' value of overcurrent relays will trip each breaker independently, as well as giving complete protection with either side energised.

Distance intertripping is applicable to feeder transformer protection with breakers situated in different stations, or to special schemes in some forms of bus-bar zone protection. Basically, feeder transformer protection is treated as a combination of (1) feeder protection, generally a balanced protection employing either circulating current or balanced voltage principles; (2) transformer protection in the form of restricted earth leakage or circulating current, both sections being independent on the a.c. side, but the arrangements necessitate the additional tripping of the breaker at the remote end of the feeder in the event of a fault on either side of the transformer.

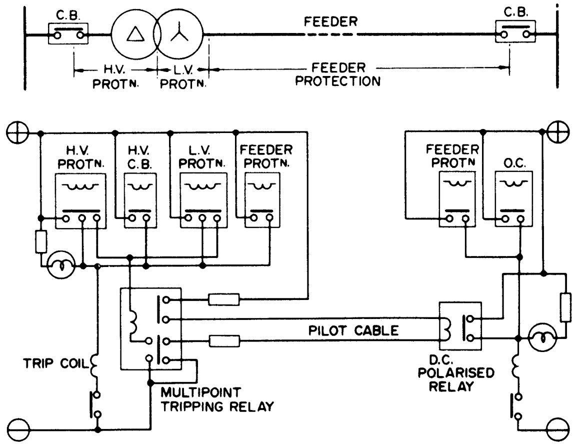


Figure 35.28 Intertripping of feeder transformer

Figure 35.28 shows an arrangement embodying a d.c. polarised relay which is energised over two pilot cables by the operation of the transformer protection relays. It is a 'dead pilot' system, as the pilots are connected to the d.c. supply through limiting resistors only when the protection operates—an obvious advantage. The scheme counters the induction of e.m.f.s in the pilots by earth fault currents in neighbouring main cables. Special relays, sensitive to d.c. but strongly biased against a.c., are also available for use in this arrangement.

Another method, applicable to certain types of balanced voltage protection, operates by interrupting one pilot cable and applying d.c. or a.c. injection across the break.

Alarm systems Modern protective relays are fitted with relay signals, but it is also necessary in attended stations to have audible and visual indication that a breaker has tripped. The methods are:

(1) Electrically operated alarm systems in which an alarm bell and lamp circuit are energised by a relay, manually reset or with self-holding contacts; in some makes a shunt or series coil closing suitable contacts is embodied in the protective relay. Separate contacts on a multipoint tripping relay, which are bridged when the relay trips, may be utilised.
(2) Mechanically operated free-handle type, in which an auxiliary switch is operated during the closing of the breaker, the alarm lamp and audible circuit being completed by an auxiliary switch on the circuit-breaker when it opens; this scheme gives the alarm not only when the breaker is tripped from protective relays, but also if the breaker slips or opens through mechanical vibration.

In all cases (with the exception of hand-reset alarm relays) cancellation of the alarm system is effected by turning the circuit-breaker controller to the open position.

D.c. supply In large stations it is the practice to employ a trickle-charged 'floating' battery with charging from the local a.c. supply through a rectifier. Where no charging supplies are available, a replacement routine must be employed.

To reduce fire hazard, modern switchgear is sectionalised. The d.c. control circuits should be similarly sectionalised and fed from the main d.c. panel by separate cables.

35.1.10 Efficacy of protection scheme

The true measure of the efficacy of a protective unit may be expressed in terms of the number N of operations, of which n are incorrect, as $(N-n)/N$. The number N includes both through-faults and internal faults; n includes failure to trip on faults in the protected zone and false operation under through-fault conditions. Many factors influence this efficacy, such as imperfect design, application or commissioning, or failure of the equipment from damage and other causes. Protective equipment must therefore be carefully selected and meticulously maintained.

35.1.11 Digital protection

A digital relay is distinguished from other static relays, largely by virtue of the fact that transduced voltages and/or current signals are sampled at regular intervals and converted, prior to further processing, to digital words (or numbers) representing the instantaneous level of the signals sampled. Once converted, the sampled values are used as input values to the protection algorithm, which effectively comprises a set of equations that are continuously evaluated for each set of new data. The protection algorithms are stored in the memory associated with a digital microcomputer or microprocessor, which in turn performs the necessary calculations on each set of incoming data so as to determine the state of the item of plant or line protected.

The basic arrangement of a digital relay is shown in *Figure 35.29*. Each input signal is passed through an analogue low-pass filter which limits the frequency content of each signal to at least half the frequency at which data are produced.

In digital systems, any frequencies in the sampled signal above one-half of the sampling frequency appear in the digital area as lower frequency components or 'aliases' and prefiltering of the analogue data so as to band limit its frequency content is necessary to avoid such data corruption. An analogue-to-digital converter then performs the necessary conversion to produce a train of digital data to the microprocessor which performs the necessary protection algorithm calculations. In practice, sampling and associated calculations are performed at frequencies which range from typically 200 samples/s to 4000 samples/s; the sampling rate is to some extent dictated by the performance requirements of the protection and, in particular, the required operating times. Digital relays commonly quantise the incoming data into at least 12 bits which in turn enables the incoming signals to be represented by signals ranging from 0 to 2^{11} (or ±2048) levels. The algorithmic calculations performed by the microprocessor will commonly involve 16-bit digital words (±32 768), which in turn generally enables the necessary degree of accuracy to be obtained. Where the algorithms are such that a high computational burden is imposed, due for example to its complexity and/or the need for a high sampling rate, it is common to use a high performance digital signal processor in preference to a conventional microprocessor.

The microprocessor is equipped with random-access memory (RAM) and read-only memory (ROM) for data and program storage. Communication with the outside world is needed to program the relay settings into the device and display status information. This is achieved by a display and keypad. Alarm and tripping signals are produced via the outgoing relays and, in addition, digital outputs supporting fault location equipment, communications modems linking other digital equipment and relays at other locations, etc., are often integrated.

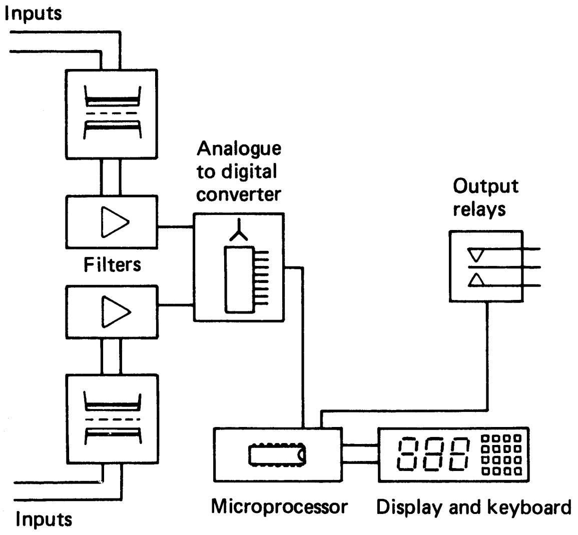


Figure 35.29 Block diagram of digital relay

Digital relays offer a number of benefits over more conventional devices and, in recent years, a number of relays have been marketed. The primary impetus for the emergence of digital protection is a general increase in demands from utilities for faster fault clearance times, better discrimination and satisfactory detection of difficult or contingency fault conditions which are not easily met by more conventional relays. Complex operating characteristics are readily programmed into digital relays which have been designed to automatically monitor themselves by executing automatic check programs which identify potential problems by comparing the response of specific circuit elements with that expected for given reference test conditions. Digital relays are generally much more flexible than more conventional types since they often include multiple characteristics and options to select any one of these. They can readily accept inputs from the digital devices and are directly compatible with digital communication systems for performing data transmission, alarm handling and supervisory controls. It is highly likely that, in the future, increasing integration of digital protection with 'electronic transducers', e.g. fibre optic voltage and current sensing devices, will occur.

Digital relaying can be applied in unit or non-unit form to protect specific items of plant and/or power feeders. At the present time, the major applications have been in the area of e.h.v. transmission line protection of the *distance, differential* and *directional comparison* types.

35.1.11.1 Distance protection

A digital distance relay has an algorithm which is designed to calculate the impedance between the point of measurement and the point of fault. This is done for each set of samples of voltage and current measured. There are a number of specific methods of calculating the impedances, the most common of which involve implementing an algorithm that is designed to evaluate the resistance and reactance of the fault loop. By reference to *Figure 35.30*, which is a simplified single-line diagram of a transmission line fault loop, the relationship between the voltage and current measured at the relaying point and the fault loop resistance R and inductance L is given by

$$v = \lessdot R + \lessdot L \, \mathrm{d}i/\mathrm{d}t \qquad (35.1)\Leftarrow$$

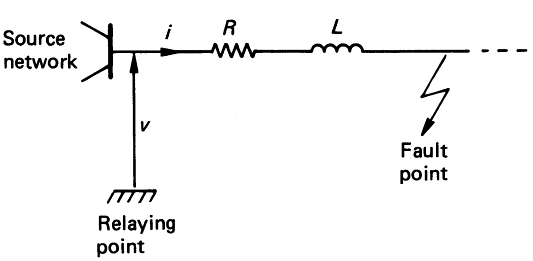


Figure 35.30 Typical distance relay fault loop

This equation applies to all samples processed. For example, consider two samples taken at times t_1 and t_2 (for which the voltage and current samples are denoted by v_1, i_1 and v_2, i_2, respectively). In this case the fault loop parameters can be obtained from an algorithm based on the equations

$$R \simeq (v_1 i'_2 - v_2 i'_1)/D \tag{35.2}$$

$$L \simeq (v_2 i_1 - v_1 i_2)/D \qquad (35.3)\Leftarrow$$

where i'⇐denotes a current signal rate of change ($\mathrm{d}i/\mathrm{d}t$) and $D = i_1 i'_2 \Leftarrow - i'_1 \Leftarrow i_2$

Estimates of the current differential terms together with the instantaneous values in the above two equations are obtained typically by utilising a succession of two or three sampled values and in this way, successive estimates of impedances measured are obtained. Such estimates are compared with the distance relay characteristic boundary, which is defined in accordance with the requirements of particular applications (see *Figure 35.25* and Section 35.1.8.6). Tripping is initiated when a defined number of successive samples of the fault loop impedance consistently lie within the defined characteristic boundary and tripping times typically as low as 8 ms are readily attainable for most applications.

35.1.11.2 Differential protection

The circulating current differential principle is at present that which is most commonly applied in digital form (see Section 35.1.8.1). *Figure 35.31* shows a functional single-line diagram of such a scheme applied to a plain feeder. In this arrangement, the currents at the two ends are sampled and transmitted to a digital differential relay located at only one end, though separate devices located at each end can be used. Each sample is converted to digital form and transmitted to a digital communications link which, in the case of *Figure 35.31*, comprises a light fibre optic link. Other digital communications channels, e.g. microwave links can equally be used. The current differential quantity, which is of the form given in Equation (35.4) is evaluated at each sample instant and compared typically with a bias quantity (Equation (35.5)). Internal faults are distinguished from healthy conditions by comparing the magnitude of successive estimates of the differential and bias quantities as determined by sampled digital values. In its simplest form, the algorithm checks that, for example, the differential quantity consistently exceeds the bias quantity for a number of successive samples before initiating circuit-breaker tripping. Where, as in *Figure 35.31*, the scheme is arranged on the 'master and slave' ends principle, a return fibre or digital communications channel can be used to transmit the signal from the measuring to the remote (slave) end.

$$i_P + i_Q \qquad (35.4)\Leftarrow$$

$$i_P - i_Q \qquad (35.5)\Leftarrow$$

35.1.11.3 Directional comparison protection

The basis of a directional comparison protection scheme is shown in *Figure 35.32*. In the particular arrangement shown, digital directional detectors measure from the voltage and current at each line end. Each directional measuring device is designed to determine the direction to the fault so that, for the faulted case shown, each measures the fault as being in the forward direction. The fault direction indicated is signalled to the equipment at the other end to produce a tripping signal. Conversely, any fault external to the line causes only one directional detector to indicate a fault in the forward direction and tripping is thereby inhibited. There are a number of alternative ways of utilising the directional indications issued, but the basic principle is the same as that outlined above. It is worth noting that only a single binary signal is required in transmitting the directional signal determined by the detectors, thus avoiding the necessity for multi-valued data communication channels.

It is common for digital directional comparison devices to measure from the fault superimposed voltages and currents. The superimposed measurands are in effect the difference between the actual value measured and the projection of the normal steady-state voltage (or current) existing at the point of measurement immediately before a fault.

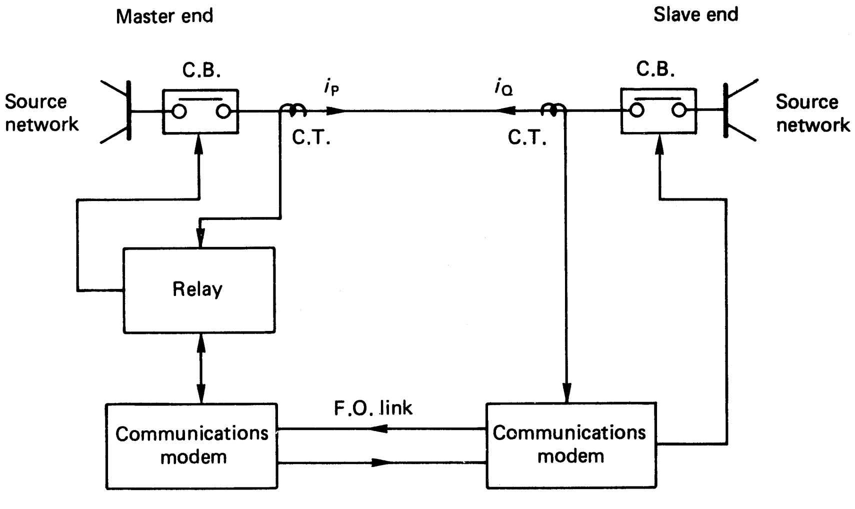


Figure 35.31 Digital difference

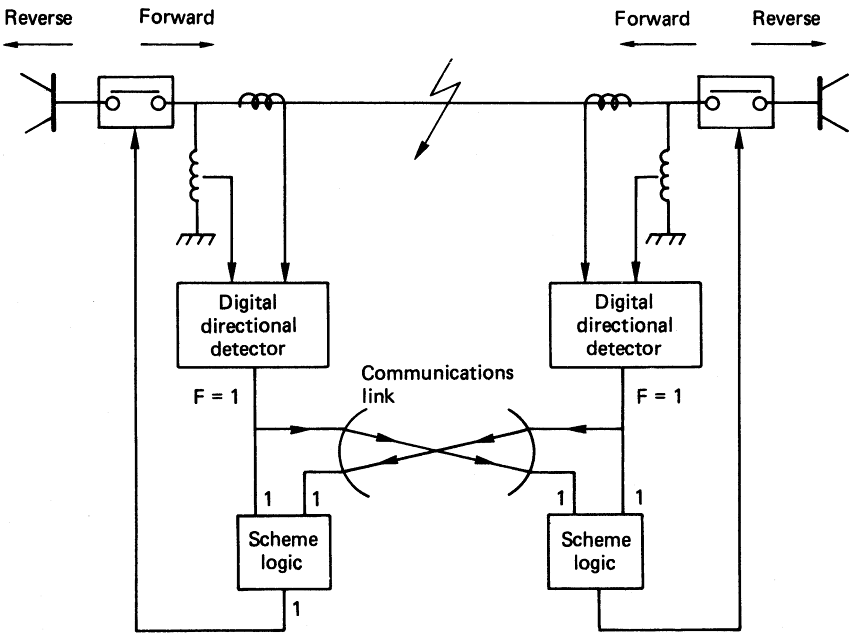


Figure 35.32 Basic directional comparison scheme

The superimposed voltages and currents can thus be considered as existing in their own right in an otherwise de-energised system, i.e. they propagate within a model of the system in which each voltage is hypothetically set at zero. For example, *Figure 35.33* shows the form of the resulting superimposed circuit for a transmission line subjected to a fault somewhere behind (or in the reverse direction with respect to) the directional measuring point at end S. *Figure 35.33* is a simplified superimposed model of a multiphase line in which the three-phase voltages and currents are combined into single voltage and current measurements. The means of achieving this network reduction is similar to that used to derive single measurements in distance protection (see Section 35.1.8.5) and further detailed information is given in the Bibliography at the end of this chapter.

The directional detectors derive superimposed measurements which are usually obtained by taking the difference between any sample on the incoming waveform and that derived from an integer number of power frequency samples previously; this differencing is equivalent to projecting the prefault steady-state variations forward in time and makes the resulting measurements consistent with a system superimposed model of the type previously discussed.

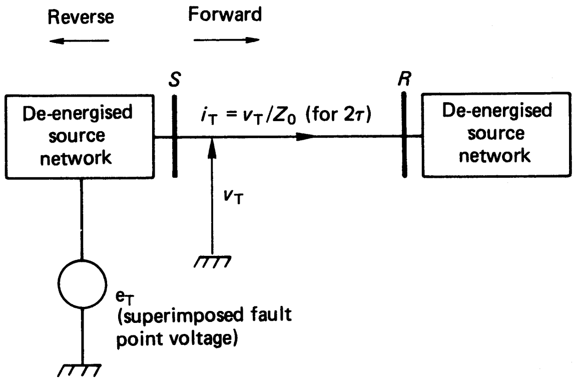

Reverse
Forward
S
R
De-energised source network
$i_T = v_T/Z_0$ (for 2τ)
De-energised source network
v_T
e_T
(superimposed fault point voltage)

Figure 35.33 Superimposed circuit for fault behind relaying point

With reference to *Figure 35.33*, the superimposed voltage and current measured at say end S (v_T, i_T) are related for all time up to twice the wave transit time (2τ) from *S* to *R* by $v_T = z_0 i_T$, where z_o is the line surge or characteristic impedance. Successive sampled values of v_T and i_T are used to form the directional signals

$$s_1 = v_T - z_0 i_T \qquad (35.6)\Leftarrow$$

$$s_2 = v_T + z_0 i_T \qquad (35.7)\Leftarrow$$

The directional signals given in the above equations are compared to determine the direction to the fault. For the reverse fault shown in *Figure 35.33*, the signal tends to zero whereas the signal s_1 attains a relatively large value which leads to twice the superimposed voltage measured at end s_2. Conversely, for a forward fault, the superimposed voltage and current are of opposite sign and in this case a reverse fault direction is indicated by signal s_1 exceeding signal s_2. Directional integrity is obtained by checking the relative size of signals over a number of post-fault samples, and the direction to a fault can be obtained in typically 2–4 ms when using a sampling rate of 3000 samples/s. The use of superimposed quantities brings a high degree of immunity to power swing and heavy circuit loading effects on account of their value being close to zero under normal steady-state conditions.

35.1.12 Artificial intelligence for protection

Conventional relays, including digital devices, rely heavily upon deterministic signal models and heuristic approaches for decision making and only a small amount of the total information available within the voltage and/or current signals used in the measuring process is utilised. Similar considerations apply in respect of relay settings and in recent years various artificial intelligence (AI) techniques have been investigated for use in the protection field. A vast amount of work has been done on the design of 'Intelligent relays'. In particular three mathematical AI tools lead themselves well to the protection field, i.e. expert systems, fuzzy logic and artificial neural networks.

35.1.12.1 Expert systems

The prime function of protective relays is the timely and discriminative clearance of system faults. In practice, a particular relay has to be set so as to ensure that its response is such that its operation is co-ordinated with that of other relays on a system. In this respect, the co-ordination of distance relays on an interconnected network is a longstanding problem. Consider for example the simple configuration of *Figure 35.34*, where relay R_x is ideally required to act discriminatively to clear faults on line P_L as well as act in a back-up mode for faults on lines S_L and Y_L.

The basis of setting distance relays is explained in Section 35.1.8.4 but it will be evident from *Figure 35.22*(*b*) that simple time grading of relay R_x in zones 1, 2 and 3 as indicated could for example result in that relay seriously underreaching in zones 2 and/or 3, where the infeed via line Y_L is significant enough to cause an 'apparent' increase in the impedance measured from relay R_x to a fault on line S_L. A relay or protection which operates to clear faults over a distance which is less than the desired or set value is said to underreach. The effect of such underreaching could for example be to increase the time of operation of relay R_x in circumstances where fast clearance of faults on line S_L is required under conditions where a circuit breaker on line S_L fails. The derivation of suitable settings for distance relays is a knowledge intensive problem that requires the experience of senior relay engineers. It is a time consuming task that is often complicated by the presence of a multiplicity of relays having different operating characteristics. Expert system based methods are particularly suited to the problem of deriving suitable settings. In essence expert systems are computer programs that contain sets of rules established using the experience of experts, which are applied to the problem in hand. The programs are thus built from explicit information derived from human experts using symbolic representation, inference and heuristic search techniques.

The basic elements of an expert system for deriving relay settings is shown in *Figure 35.35*. The inference engine commonly uses the AI software language Prolog which interacts with the user interface to accept network description from the user and supply the relay setting results.

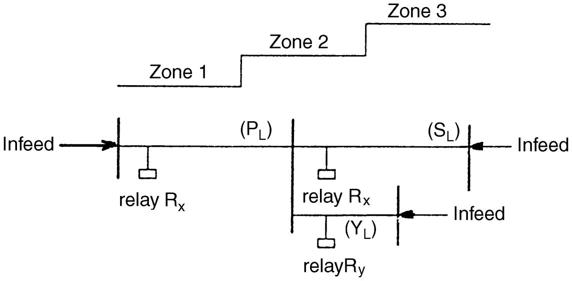


Figure 35.34 Distance relays applied to an interconnected system

35.1.12.2 Fuzzy logic

The majority of conventional protection techniques involve defining circuit breaker states by identifying the patterns of the measured voltages and/or currents. In practice however, there exists a significant degree of uncertainty and vagueness due to the complex relationships between the system response to disturbances and the resulting measurands. Examples of the factors contributing to such vaguenesses are signal transducer noise caused by electromagnetic interference and changes in load, generation or the network topology. Fuzzy logic has and is still being extensively investigated as a means of developing novel protection for power systems. It is essentially a method of readily representing human expert knowledge on a digital processor in particular where mathematical or rule-based expert systems experience difficulty. *Figure 35.36* shows the basic structure of a relay utilising fuzzy logic. The sensor data is converted to fuzzy data using a fuzzification process in which each variable is assigned a degree of membership of a particular fuzzy class. For example, *Figure 35.37* shows membership features for a particular voltage or current measured; here the process of classification involves defining the measurand as being very small (VS), small (S), normal (N), large (L), or very large (VL). Thus for example the vector showing the degree of membership for a voltage of magnitude V_1 would be [VL, L, N, S, VS] = [0, 0, 0, 0.2, 0.5] whereas for a voltage V_2, the corresponding vector would be [0, 0.7, 0, 0, 0]. A matrix of degree of membership is formed and the inference engine applies in effect a number of rules each of which generates a fuzzy parameter which is defuzzified to provide a crisp control signal which is channelled to the circuit breaker(s). There are a number of algorithms that are applied in the defuzzification process, the most common of which is the so called 'maximum algorithm' in which the element in the matrix with the largest membership value is chosen to define the required breaker control action.

35.1.12.3 Artificial neural networks

Artificial neural networks (ANNs) resemble the structure of the human brain but many experts are of the opinion that the resemblance is very superficial. In engineering terms, what is important is that the ANNs can be 'trained to perform' a required action. They are being developed for use in the protection field by using massive training sets of data for which the required response is known. *Figure 35.38* shows the topology of a simple ANN. The inputs are normally derived from measured voltages and/or currents derived from the power system via voltage and current transducers. Various features are extracted from the measured signals e.g. the magnitude of a harmonic in a voltage signal may form one input signal. It can be seen that the ANN is built up of a number of nodes (or artificial neurons) each of which summates weighted input signals and further weights the sum before transmitting the output to other neurons. In practice an ANN may have a number of hidden layers the number of which depends on factors relating to the nature and extent of the protection problem; the single hidden layer network is however by far the most common and has been found to provide a satisfactory performance in most protection applications. Each artificial neuron employs a non-linear weighting function of which a number are employed. In essence however all weighting functions

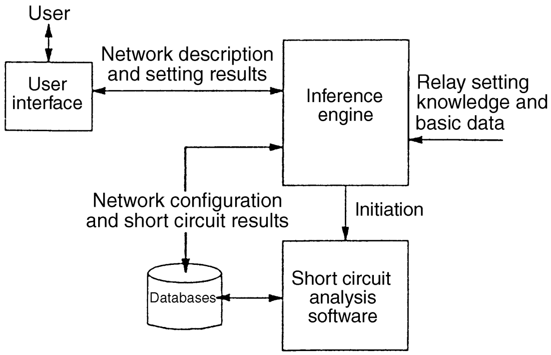

User
User interface
Network description and setting results
Inference engine
Relay setting knowledge and basic data
Network configuration and short circuit results
Initiation
Databases
Short circuit analysis software

Figure 35.35 Basic elements of a relay setting expert system

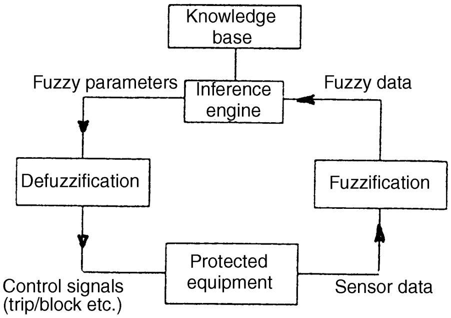

Knowledge base
Fuzzy parameters
Inference engine
Fuzzy data
Defuzzification
Fuzzification
Protected equipment
Control signals (trip/block etc.)
Sensor data

Figure 35.36 Basic arrangement of a 'fuzzy relay'

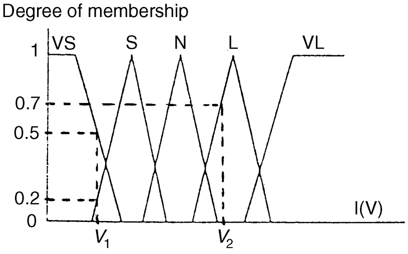


Figure 35.37 Typical membership functions

behave so as to produce an output level which suddenly switches from a low signal level to a high level when the summated input reaches a critical level.

Figure 35.39 shows the basic structure of a relay employing an ANN. As mentioned previously, the ANN needs to be trained. This is done using a training algorithm of which the 'backpropagation' algorithm is often used. In essence, the ANN is subjected to an input array derived via system simulation studies or actual recorded system data, for which the desired output is known. For example, the input may be derived from a faulted transmission line for which a tripping signal is required or conversely an input array associated with a healthy situation where the output should be zero. Training is performed iteratively by adjusting the weight of the input to each artificial neuron so as to obtain the desired output(s). In practice training has to be achieved using a training set which is large enough to ensure final convergence of the network weighing coefficients. Very significant progress has been made in the application of ANNs to protection and associated control functions such as single-pole autoreclosure. A number of commercial developments are underway and there is little doubt that ANN based protection and control will ultimately be applied. In this respect, it is likely that the first applications will be aimed at providing improved performance of protection in difficult applications and where it is necessary to cover contingency fault and systems conditions where more conventional protection, including digital measuring devices, often provide less than optimal response.

35.1.12.4 Hybrid artificial intelligence networks

Recent research indicates that AI based protection techniques may be significantly enhanced by the integration of expert systems, fuzzy logic and artificial neural network techniques. Fuzzy neural networks are actively being researched as a means of further enhancing next generation protection performance. Further integration of expert system techniques is also likely to provide a very important way forward in future generations of 'intelligent protection', though much more on-going work is required in this area.

35.2 Application of protective systems

There are usually several ways of protecting any given equipment, and the more usual zones of protection are

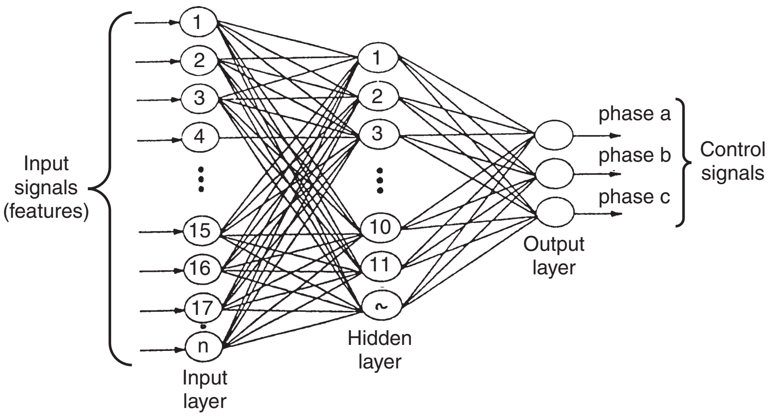

Input signals (features)
1
2
3
4
15
16
17
n
Input layer
1
2
3
10
11
~
Hidden layer
Output layer
phase a
phase b
phase c
Control signals

Figure 35.38 Rudimentary artificial neural network (ANN) topology

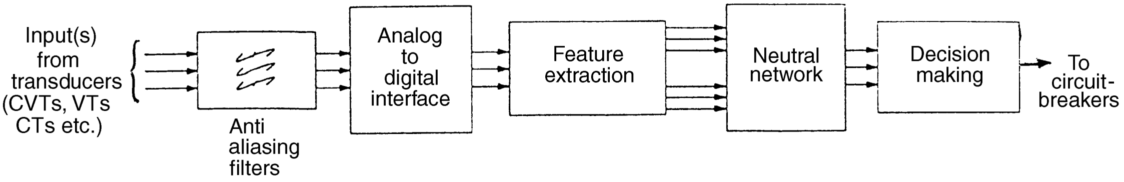


Figure 35.39 Basic structure of an (ANN) protection relay

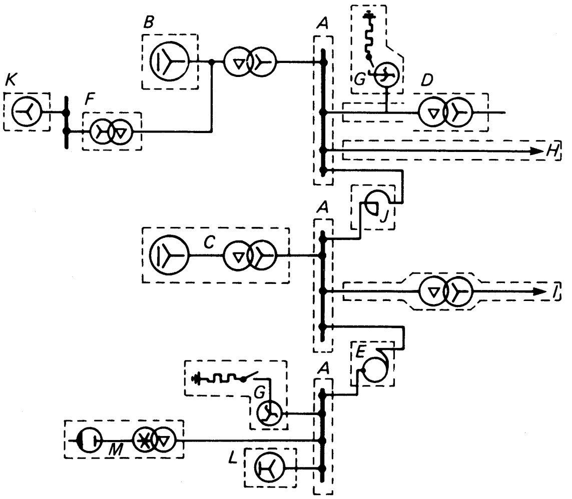


Figure 35.40 Zones of protection: A, bus-bars; B, generator; C, generator–transformer; D, power transformer; E, auto-transformer; F, unit transformer; G, earthing transformer; H, transmission line of feeder; I, transformer; J, reactor; K, induction motor; L, synchronous motor; M, rectifier

shown in *Figure 35.40*. One relay can often be used for several functions; for example, a triple-pole overcurrent relay can be used for both overcurrent and earth fault protection, and the following combinations are commonly found in practice: (1) inverse time overcurrent and earth fault; (2) inverse time with instantaneous high-setting overcurrent, with or without inverse time earth fault; and (3) thermal overcurrent, with instantaneous overcurrent and earth fault.

The power system neutral is considered to be earthed when the neutral point is connected to earth directly or through a resistor. Earth fault protection is applicable in all such cases. The neutral point is considered to be insulated when the neutral point is not connected to earth, or is earthed through a continuously rated arc suppression coil or through a voltage transformer.

Arc suppression coils are a form of protection applicable where the majority of the earth faults are expected to be of a transient nature only. They can be continuously or short-time rated (normally 30 s), with one phase of the system faulted to earth. The continuously rated coils can be provided with an alarm to indicate the presence of a persistent fault, and in some cases directional earth fault relays are used to indicate the location of an earth fault. The short-rated coil is short-circuited, either directly or through a resistor after a short time, so that persistent earth faults can be cleared by normal discriminative earth fault protection.

35.2.1 Plant

35.2.1.1 Generators

The core of an electrical power system is the generator, requiring a prime mover to develop mechanical power from steam, gas, water or diesel engines. The range of size extends from a few hundred kilovolt-amperes up to turbine driven sets exceeding 600 MV-A in rating. Small sets may be directly connected to the distribution system, while the larger units are associated with an individual step-up transformer through which the set is coupled to the transmission system.

A modern generating unit is a complex system consisting of the generator stator, the rotor with its field winding and exciters, the turbine and associated condenser, the boiler with auxiliary fans and pumps, and possibly an associated or unit transformer. Faults of many kinds can occur in such a system, for which diverse protective means are needed. The amount of protection applied will be governed by economic considerations. In general, the following faults need to be considered:

(1) stator insulation earth faults;
(2) overload;
(3) overvoltage;
(4) unbalanced loading;
(5) rotor fault;
(6) loss of excitation;
(7) loss of synchronism;
(8) failure of prime mover;
(9) low vacuum in condenser;
(10) lubrication oil failure;
(11) overspeeding;
(12) rotor distortion;
(13) differential expansion; and
(14) excessive vibration

The neutral point of a generator is normally earthed, with some impedance inserted in the earthing lead to limit the magnitude of earth fault current to values from a few amperes to about rated full load current. Phase–phase faults clear of earth are unusual, as are interturn faults; and they usually involve earth in a short time. Circulating current biased differential protection is the most satisfactory way of protecting a generator stator or a generator–transformer unit.

Generators on a large system under continuous supervision are not usually subject to prolonged accidental overloading and are often protected only by built-in temperature measuring devices on both stator and rotor. For the smaller machines inverse definite minimum time (i.d.m.t.) overcurrent relays are used (often forming a back-up feature) and may be voltage restrained to give better discrimination.

Power frequency overvoltages are usually dealt with by instantaneous relays with fairly high settings, and transient overvoltages by surge diverters and capacitors connected to the generator terminals if the incoming surges are not sufficiently reduced by transformer interconnections.

Any unbalanced load produces negative sequence components, the resulting reaction field producing double-frequency currents in the field system and rotor body, which can result in serious overheating. For turbo sets the negative sequence continuous rating is only 10–15% of the continuous mean rating value for positive sequence. The unbalance is detected by relays in a negative sequence network, providing first an alarm and finally tripping the generator load.

Rotor faults due to earthing of the winding, or partial short-circuited turns, can be detected by methods involving a potentiometer across the winding, or a.c. or d.c. injection. Failure of the field system results in a generator losing synchronism, running above synchronous speed and operating as an induction generator with the main flux produced by reactive stator current drawn from the system. In general, a 60 MW machine with conventional cooling will not be overheated by asynchronous operation at full load for 5 min; but a hydrogen cooled 500 MW set should not be run like this for more than 20 s. A generator may also lose synchronism because of a severe system fault, or operation at a high load with leading power factor and, hence, a relatively weak field. This subjects the generator and prime mover to violent oscillations of torque, but synchronism can be regained if the load is reduced in a few seconds; otherwise it is necessary to isolate the machine. Alternatively, the excitation may be removed so that the generator runs asynchronously, thereby removing the violent power oscillations. Re-closing the excitation at a low value will then cause the machine to resynchronise smoothly. The range of possible positions of saving curves makes detection by simple relays relatively difficult, but by the use of two impedance relays, pole slipping can be detected and corrected by varying the excitation.

Various faults may occur on the mechanical side of the set, such as failure of the prime mover, overspeed, boiler failure, loss of vacuum, lubricating oil failure, rotor distortion, etc. These can all be protected, the inclusion being a matter of economics.

Where generator–transformers are used, the differential protection will always need to be biased, using either harmonic bias or special attention to settings. Earth fault protection will cover the generator and the transformer primary winding; in addition, the transformer h.v. winding will usually be provided with restricted earth fault protection and Buchholz protection (see *Figure 35.41*).

The field of a faulty generator should be suppressed as quickly as possible. For machines above about 5 MV-A this is usually done by connecting a field discharge resistor of about five times the rotor field-winding resistance in parallel with the generator field when the field circuit-breaker opens. This reduces the field time constant, but it may still be more than 1 s.

35.2.1.2 Transformers

The power transformer is one of the most important links in a transmission system, and as it has the greatest range of characteristics, complete protection is difficult. These conditions must be reviewed before detailed application of protection is settled. The ratings of units used in transmission and distribution schemes range from a few kilovolt-amperes to several hundred megavolt-amperes. The simplest protection, such as fuses, can be justified only for the small transformers; those of the highest ratings have the best protection that can be afforded.

A fault in a transformer winding is controlled in magnitude not only by the source and neutral earthing impedance, but also by the leakage reactance of the transformer, and the fact that the fault voltage may differ from the system voltage according to the position of the fault in the winding. Star-connected windings may be earthed either solidly or through an impedance, presenting few protection problems; but delta-connected windings require close consideration, as

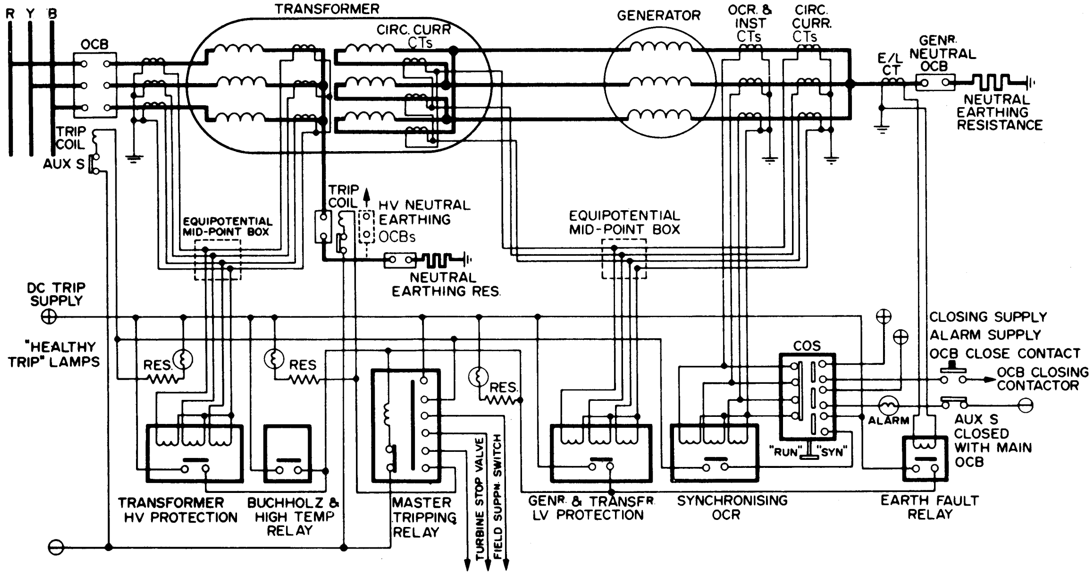

R
Y
B
TRANSFORMER
OCB
CIRC. CURR CTs
GENERATOR
OCR. & INST CTs
CIRC. CURR. CTs
E/L CT
GENR. NEUTRAL OCB
NEUTRAL EARTHING RESISTANCE
TRIP COIL
AUX S
TRIP COIL
HV NEUTRAL EARTHING OCBs
EQUIPOTENTIAL MID-POINT BOX
NEUTRAL EARTHING RES.
EQUIPOTENTIAL MID-POINT BOX
DC TRIP SUPPLY
"HEALTHY TRIP" LAMPS
RES.
RES
RES.
CLOSING SUPPLY
ALARM SUPPLY
OCB CLOSE CONTACT
OCB CLOSING CONTACTOR
COS
ALARM
AUX S CLOSED WITH MAIN OCB
"RUN"
"SYN"
TURBINE STOP VALVE
FIELD SUPPN. SWITCH
TRANSFORMER HV PROTECTION
BUCHHOLZ & HIGH TEMP RELAY
MASTER TRIPPING RELAY
GENR. & TRANSFR. LV PROTECTION
SYNCHRONISING OCR
EARTH FAULT RELAY

Figure 35.41 Generator–transformer protection

the individual phase currents may be relatively low. Faults between phases within a transformer are comparatively rare, and interturn faults in low-voltage transformers are unlikely unless caused by mechanical force on the windings due to external short circuits. Where a high-voltage transformer is connected to an overhead transmission line, a steep-fronted voltage surge due to lightning may be concentrated on the end turns of the winding, resulting in 70–80% of such transformer failures unless some form of voltage grading is employed. The bolts that clamp the core together are insulated to prevent eddy currents. Should this insulation fail, then severe local heating may damage the winding. As the oil is usually broken down, the gas produced can be used to operate a Buchholz relay. Tank faults are rare, but oil sludging can block cooling ducts and pipes, leading to overheating.

One of the major problems in the protection of large transformers is the phenomenon of magnetising current inrush when switching on. This is not a fault condition and the protection relays must not trip, although the inrush current appears superficially as an internal fault and may have a long time constant. To avoid long time delays and, hence, damage to important transformers, it is essential to clear all faults rapidly. Use is made of the fact that magnetising inrush currents contain a unidirectional component plus second, third and higher harmonics. The second harmonic is the most useful as a stabilising bias against inrush effects, and by combining this (extracted through a filter) with the differential current through a static device, a setting of 15% can be obtained, with an operating time of 45 ms for all fault currents of two or more times rated current. The relay will restrain when the second harmonic component exceeds 20% of the current.

Most of the usual transformer arrangements such as star/delta, auto-, earthing, etc., can be protected by differential relays, sometimes with restricted earth fault relays.

Sometimes transformers are connected directly to transmission lines without circuit-breakers. In such cases the transformer and feeder are both protected by differential protection plus intertripping of the remote circuit-breaker. An alternative to intertripping is to detect earth faults on a feeder connected to a non-earthed transformer winding by measuring the residual voltage on the feeder, using either voltage transformers or capacitors to detect the neutral displacement.

35.2.1.3 Reactors

The most common are the series and tie-bar reactors, for limiting overcurrents. On earthed systems with air insulated reactors the protection usually consists of differential relays or time-delayed overcurrent with time-delayed earth fault relays. For insulated-neutral systems the earth fault relays are unnecessary. Shunt reactors for compensating line capacitance are usually oil immersed and protected by time-delayed overcurrent relays, with instantaneous restricted earth fault relays where the system neutrals are earthed. Buchholz relays are often used on oil immersed reactors.

35.2.2 Feeders

Many schemes are available: therefore, only general guidance can be given. Scope still exists for adapting schemes to individual requirements. Fuses and/or overcurrent protection with graded time lags are commonly used for 11 kV and 33 kV distribution systems with balanced opposed-voltage schemes for the more important 33 kV lines.

The protection can be by either electromagnetic or static relays, both of which may be switched by fault detectors, and for earthed systems can incorporate earth fault relays.

The length of feeder frequently results in additional circuits being necessary between the relays at each end. Pilot wires are often used for this purpose, but for distances above about 20 km the pilot wires may behave as a transmission circuit, and may need shunt reactors to compensate for the circuit capacitance. Pilot wires are also liable to electrical interference, manual disturbance and limits as to overvoltage and current, so that many schemes are provided with continuous supervision to warn that the overall protection is unsound and relay adjustment may be necessary. By using the line itself as a link and injecting high-frequency signals (70–700 kHz), the necessary end-to-end protective coupling can be achieved. Information is impressed on the signal by a modulation process. Attenuation of the signal carrier makes pilots unsuitable for transmitting the amplitude of a measured quantity; however, the phase can be transmitted satisfactorily. Frequency modulation can convey all the characteristics of the modulating quantity but involves frequency–amplitude conversion. The best use of carrier involves simple on/off switching: this method of modulation has been extensively used in conjunction with polyphase directional relays by transmitting a locking signal over the line for through-faults.

With extensive power systems and interconnection to ensure continuity of supply and good voltage regulation, the problems of combining fast fault clearance with protective gear co-ordination have become increasingly important. To meet these requirements high-speed protective systems suitable for use with automatic reclosure of the circuit-breakers are under continuous development and are already very widely used. The distance scheme of protection is comparatively simple to apply and offers high speed, primary and back-up facilities, and needs no pilot wires; but by combining it with a signal channel (such as carrier) it is particularly suitable for use with high-speed auto-reclosing for the protection of important transmission lines.

Since the impedance of a transmission line is proportional to its length, it is appropriate to use a relay capable of measuring the impedance of a line up to a given point. The basic principle of measurement (Section 35.1.8) involves the comparison of the fault current seen by the relay with the voltage of the relaying point, whence it is possible to measure the impedance of the line up to the fault. The plain impedance relay takes no account of the phase angle between the current and voltage applied to it, and is therefore non-directional. It has three important disadvantages:

(1) it is inherently non-directional and therefore needs a directional element to give it discrimination;
(2) it is affected by arc resistance; and
(3) it is highly sensitive to power swings, because of the large area covered by the impedance circle.

The *admittance relay*, by addition of a polarising signal, combines the characteristics of the impedance and directional relays. It is satisfactory on long lines carried on steel towers with overhead earth wires, where the effect of arc resistance can be neglected, but for lines on wooden poles without earth wires the earth fault resistance reduces the effective zone to such an extent that most faults take longer to detect. This problem can usually be overcome by using either reactance or fully cross-polarised admittance relays for the detection of earth faults. In theory, any increase in the resistive component of the fault impedance has no effect upon a reactance relay: however, in practice, when the fault resistance approaches that of the load, then the relay characteristics are modified by the load and its power factor and it may cover either more or less of the line. This can be overcome by the use of a fully cross-polarised circuit which needs two input signals for a ±90° comparison angle.

A distance protection scheme comprises starting, distance measuring, auxiliary, zone timer and tripping relays. To cater for the economic and technical requirements of a particular power system, schemes are available using either a plain distance measurement with several steps of protection, or a combination of distance measurement and a high-speed signalling channel or power-line carrier to form a unit system of protection over the whole of the protected line and to provide its own back-up protection to the adjacent lines. Full-distance, or switched-distance, schemes are applied according to the system voltage and the importance of the lines to be protected. The main difference between the two arrangements is that the full-distance scheme uses six measuring units (three for phase and three for earth faults), whereas the switched-distance scheme uses only one measuring unit for all types of fault, this being switched to the correct fault loop impedance by means of a suitable set of starting units of the overcurrent or underimpedance type.

One of the main disadvantages of conventional time stepped distance protection is that the instantaneous zone of protection at each end of the protected line cannot be set to cover the whole of the feeder length. It usually covers about 80%, leaving two end zones in which faults are cleared instantaneously at one end of a feeder but in much longer time at the other end. In some applications this cannot be tolerated; either the delay in fault clearance may cause the system to become unstable, or, where high-speed auto-reclosing is used, the non-simultaneous opening of the circuit-breakers interferes with the auto-reclosing cycle. These objections can be overcome by interconnecting the distance protections at each end by a signalling channel to transmit information about the system conditions at one end to the other end of the protected line. It can also initiate, or prevent, tripping of the remote circuit-breaker. The former arrangement is a ‘transformer trip scheme’, while the other is a ‘blocking scheme’.

If two circuits are supported on the same tower or are otherwise in close proximity over the whole of their length, there is mutual coupling between them. The positive and negative sequence coupling is small and usually neglected. The zero sequence coupling cannot be ignored. Types of protection that use current only (e.g. power-line carrier phase comparison and pilot-wire differential systems) are not affected, whereas types using current and voltage, especially distance protection, are influenced by the phenomenon and its effect requires special consideration.

The protective schemes described previously for the protection of two-ended feeders can also be used for multi-ended feeders with load or generation at any terminal. However, the application of these schemes to multi-ended feeders is much more complex and requires special attention.

As transient faults, such as an insulator flashover, make up 80–90% of all faults on HV and e.h.v. systems, and are cleared by the immediate tripping of one or more circuit-breakers, they do not recur when the line is re-energised. Such lines are frequently provided with auto-reclosing schemes, and result in a substantial improvement in continuity of supply and system stability. The choices of dead time, of reclose time, and of whether to use a single- or a multi-shot scheme, are of fundamental importance. With a three-phase scheme the tripping of all three phases on the occurrence of a fault means that no interchange of synchronising power can take place during the dead time. If only the faulty phase is tripping during earth fault conditions (which account for the majority of faults on overhead lines), synchronising power can still be interchanged through the healthy phases, but this entails the provision of circuit-breakers provided with tripping and closing mechanisms on each phase, and the reclosing circuitry is more complicated. In the event of a multi-phase fault, all three phases are tripped and locked out. It is important to ensure simultaneous tripping of the circuit-breakers at all ends of the faulty section, and distance protection imposes some difficulties in this respect, so that a signalling channel is often used between ends. On highly interconnected transmission systems where the loss of a single line is unlikely to cause two sections of the systems to drift apart and lose synchronism, three-phase auto-reclosing can be delayed for 5–60 s with an increase in successful reclosures.

35.2.2.1 Bus-bars

Bus-bars have often been left without specific protection for one or more of the following reasons.

(1) The bus-bars and switchgear have a high degree of reliability and are often regarded as intrinsically safe.
(2) It was feared that accidental operation of bus-bar protection might cause widespread dislocation which, if not quickly cleared, would cause more loss of supply than the very infrequent bus-bar faults.
(3) It was expected that system protection, or back-up protection, would provide sufficient bus-bar protection if needed.

The reasons are applicable only to small indoor metal enclosed stations: with outdoor switchgear the case is less clear, for although the likelihood of a fault is higher, the risk of damage is much less. In general, bus-bar protection is required when the system protection does not cover the bus-bars, or when, to maintain system stability, high-speed fault clearance is necessary.

To maintain the high order of stability needed for bus-bar protection, it is now an almost invariable practice to make tripping dependent on two independent measurements of fault quantities. Methods include: (a) two similar differential systems; (b) a single differential system (*Figure 35.42*), checked by a frame earth system (*Figure 35.43*); (c) earth fault relays energised by c.t.s in the neutral earth connection; (d) overcurrent relays; and (e) a frame earth system checked by earth fault relays. Separate current transformers are normally used with separate tripping relays series-connected in pairs to provide a single tripping output, so that independence is maintained throughout. The essential principle is that no single occurrence of a secondary nature shall be capable of causing an unnecessary trip of a bus-bar circuit-breaker.

35.2.3 Motors and rectifiers

35.2.3.1 Motors

Protection is required against (a) imposed external conditions, and (b) internal faults.

Category (a) includes unbalanced supply voltages, undervoltage, single-phasing, and reverse-phase-sequence starting; (b) includes bearing failures, internal earth faults and overloads.

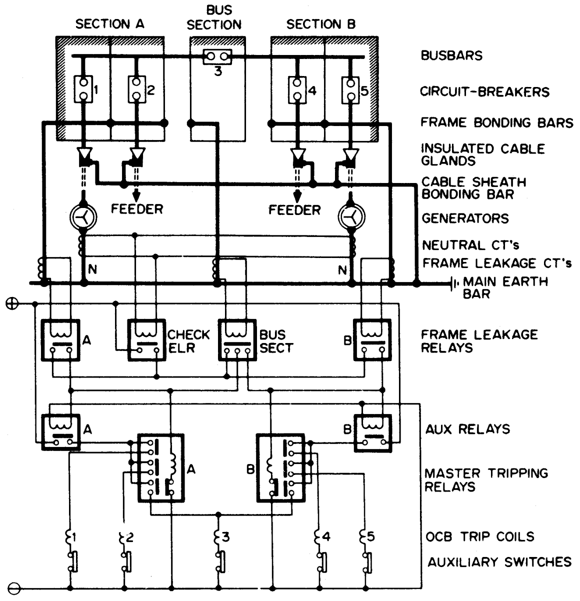


Figure 35.42 A single differential system for bus-bar protection

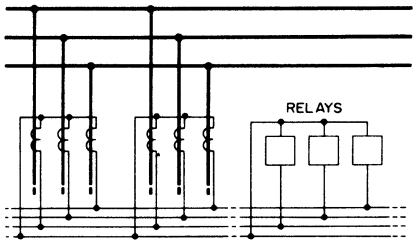


Figure 35.43 A frame earth system

Motors up to about 400 kW usually have ball or roller bearings which may fail very quickly: the only remedy is to disconnect the motor as rapidly as possible to avoid the overcurrent damaging the windings. Larger machines usually have sleeve bearings, incipient failure of which may be detected by a temperature device embedded in the bearing. As the current increase may be only 10–20%, it cannot be detected by the overload relays.

The majority of winding failures are caused by overload, leading to deterioration of the insulation followed by electrical fault in the winding.

The wide diversity of motor designs makes it impossible to cover all types and ratings with a simple characteristic curve. Motors with fluctuating loads where shutdown would affect the whole process might be left running by giving the overload relay a higher current setting. Motors on a steady load can be tripped more quickly, as overload will probably be due to a mechanical fault.

A temperature compensated static relay can follow changes of the working temperature of the motor more accurately, so that it will not be shut down on overload unless it is overheated. By including a number of alternative operating time–current characteristics it is possible to cover a wide range of motor designs and applications. Protection of motors operating on unbalanced voltages is provided by separating the positive and negative sequence components of the line currents by means of a sequence filter, arranged to allow for the greater relative heating effect of negative sequence current components.

35.2.3.2 Rectifiers

The majority of rectifier duties are now performed by semiconductors: these have special demands for protection. Faults can be classified as: (a) cell overload and failure; (b) operational faults such as backfires; (c) d.c. faults on the bus-bars and cables; and (d) a.c. faults on the supply transformer and cables.

Because of their low thermal mass and consequent low fault-current withstand, the series- or parallel-connected cells need to be individually protected, usually by fuses shunted by a small indicating fuse which, on blowing, operates a microswitch to trip the supply or give an alarm.

Semiconductor cells are susceptible to overvoltage failures caused by switching or lightning surges, arc voltages on fuse clearance, chopping of load current, etc. Protection usually takes the form of surge diverters or resistor/capacitor networks connected to earth (or between phases) and across the d.c. output. Rectifier transformers are often provided with earthed screens between the windings to limit the surge transfer.

35.3 Testing and commissioning

The testing of protective schemes has always been a problem, because protective gear is concerned only with fault conditions and cannot readily be tested under normal system operating conditions. The problem has been aggravated in recent years by the complexity of protective schemes and relays.

Protective gear testing may be divided into three stages: (1) factory tests, (2) commissioning tests, and (3) periodic maintenance tests.

The first two stages prove the performance of the protective equipment during its development and manufacture, and in its operational environment. The last stage, properly planned, ensures that this performance is maintained throughout its life.

The relay manufacturer must provide adequate testing of protective gear before it is accepted and commissioned. The tests performed are (a) tests in which the operating parameters of the relays, etc., are simulated; (b) conditions such as temperature range, vibration, mechanical shock, electrical impulse withstand, etc., which might affect the operation in service. In some cases tests of both groups are conducted simultaneously to check performance.

With the advent of static relays the use of semiconductors and other electronic components has posed new problems to the manufacturer, because the production of such devices is not within his control. Quality control testing procedures have been established with the object of identifying the device that may fail early in its life, the usual method being to determine which critical parameters will show a substantial change when subjected to accelerated life testing and to examine these.

35.3.1 Commissioning tests

The object of the commissioning tests is to ensure that the connections are correct, that the performance of current transformers and relays agrees with the expected results, and that no components have been damaged by transport or installation. This performance test includes correct current transformer ratio, correct calibration of relays and tests proving that the tripping, intertripping and indication of the scheme are in order.

Although the details of every protective scheme vary, the main points to be checked on every scheme are: (a) stability of the system under all conditions of 'through' current; (b) that the sensitivity of all relays with reference to the primary current is correct.

Prior to making these final checks on site a very careful preliminary check on the protective gear scheme should be made. Such tests would include:

(1) examination of all small wiring connections, and insulation resistance tests on all circuits;
(2) identification of all pilot cables—insulation resistance tests and loop resistance tests on balanced pilot-wire schemes;
(3) polarity and ratio check (if possible) on all current transformers;
(4) tests on d.c. tripping, intertripping and operation of all breakers connected with the unit, including alarm circuits and relay signals;
(5) checking of calibration of all relays by secondary injection; and
(6) miscellaneous tests depending on special details of a particular scheme.

35.3.1.1 Saturation curve

Considerable use on site may be made of a c.t. saturation curve using a LV local supply; it compares closely with results obtained by primary current. An LV a.c. supply is fed to the secondary winding through a control rheostat, and a curve of voltage–current for the secondary is plotted. The equivalent primary current is inferred from the turns ratio. The test is useful for comparing the performance of c.t.s required to have matched characteristics.

Such preliminary tests ensure that components are correct. The commissioning process thereafter must depend upon site facilities. The following notes give a general outline.

35.3.2 Primary current tests

A *generator* is isolated with the unit under test and by means of primary short circuits and earth faults stability figures, on balanced systems, under full load conditions and operation tests with internal faults may be carried out. It is important that the makers of the generator be consulted before it is used for steady unbalanced conditions, e.g. testing with one phase earthed, as the distortion of flux combined with armature reaction, if prolonged, may result in excessive heating. Earthing resistors and transformers should be short circuited or bypassed during current testing, as they are normally short time rated. An *inverted transformer* is a means of stepping-up testing current where a transformer is available between the unit under test and the source of test supply. For example, if the transformer has a normal step-up ratio of 6.6 kV to 66 kV, then by connecting the testing supply (isolated generator or low-voltage source) to the 66 kV windings and supplying the test current from the output of the 6.6 kV windings a 10/1 step-up of current is obtained. This 'inversion' can readily be carried out by flexible cabling when the connections to the transformer are made through open-type bushings, and is applicable to any power transformer.

A heavy current *testing transformer* may be designed with an input suitable for a general range of low-voltage supply. The types available include three-phase and single-phase units with resistance or induction regulator control. For the purpose of producing primary current the output terminals are connected across the primary of the circuit under test, to test windings or test bar primaries embodied in the current transformers. On metal clad switchgear, connection to the primary can be made through the circuit spouts with testing plugs; potential transformer spouts may also be used, but the current-carrying capacity should be investigated.

When using externally mounted current transformers around cables, e.g. in core balance types of protection, flexible cables may be used for threading through the transformers with current supplied from a heavy current testing transformer. The cable should be as central as possible, as small errors are introduced if the cable is asymmetrically located. For balancing purposes it is noted that if the relative position of each cable in each transformer is the same, the same error will be introduced and the effect neutralised.

35.3.3 Secondary injection tests

The preceding methods utilise primary current, but use can be made of secondary injection. The effect of primary current in a c.t. is to develop a secondary voltage; in secondary injection, voltage is applied to the secondary terminals, usually from an injection transformer, so applying the 'output' voltage. The method has application for checking relays, calibration, commissioning and maintenance. It is not in itself sufficient for balanced schemes, and must in such cases be supplemented by tests on load; however, it is possible in most cases to simulate through-faults to earth, and phase–phase conditions, by rearrangement of the secondary c.t. connections. One limitation is that the 'load' current is dependent on the load on the unit, usually determined by network conditions, and on balanced systems the stability check may be at a low primary current.

The various types of injection transformers generally have a tapped primary winding, with a secondary resistance controlled output covering a range of voltage. The method gives fine control for calibrating and enables a low power input to be used.

When using injection transformer for timing tests on induction-type relays with inverse time characteristics, the waveform characteristics should be carefully ascertained. Such relays are very sensitive to wave form (having an inherent, saturable iron circuit), and a more satisfactory and reliable method for checking timing is to supply the test current from a low voltage with resistance control. In this way the saturating feature of the relay is absorbed by the controlling resistance, the source of supply being treated as a transformer of relatively high capacity. It must be noted that the saturation of the injection transformer itself is important.

35.3.4 Fault location

Rapid location of a fault, together with some idea of its nature, is clearly an essential preliminary to its quick repair. With overhead lines visual inspection either from the ground or from a helicopter may be possible, but if this is not practicable, and almost always in the case of underground cables, inspection of the flag indicators on protective relays and simple Megger or continuity tests will usually provide useful evidence and enable a suitable method of more detailed investigation to be selected.

35.3.4.1 Loop tests

If the fault is of low resistance, a low-voltage (car) battery may suffice for the test supply. For higher-resistance faults about 500 V from, say, a Megger may be required, while for very high resistance, where it is necessary to break down the fault, an HV rectifier set may be used.

35.3.4.2 Other tests

Fall-of-potential, capacitance and pulse reflection tests are more sophisticated methods. Such tests, with loop tests, may be accurate to a location within 20 m or so of the fault point. In the case of underground cables a more precise location is desirable before excavation. Typical are induction and discharge methods.

35.4 Overvoltage protection

Lightning, switching and other less common phenomena produce overvoltages on transmission and distribution systems. Precautions against consequential damage and system outage must be taken, by (a) preventing overvoltages from being impressed on the system, and (b) protecting vulnerable equipment from voltage surges.

Insulation levels are based on combinations of short-duration power-frequency overvoltages, and impulse voltages arising from lightning or switching. Protective devices are (1) non-linear resistor surge arresters, (2) expulsion surge arresters (up to voltages not exceeding 36 kV), and (3) spark-gaps. These do not provide the same degree of protection, and the choice between them depends on such factors as the relative importance of the plant to be protected, the consequence of an outage, the system layout, the probable lightning activity and the system voltage.

The breakdown of plant insulation varies with the time for which an overvoltage is maintained. *Figure 35.44* indicates typical voltage–time relations for plant insulation and overvoltage protection devices. A surge arrester has a characteristic similar to that of the plant, and is usually arranged to spark-over at a voltage slightly lower. A spark-gap has a more nearly hyperbolic characteristic, so that the plant insulation may break down first on overvoltages of short duration; but if, to avoid this, the spark-gap breakdown voltage is reduced, many unnecessary outages may result.

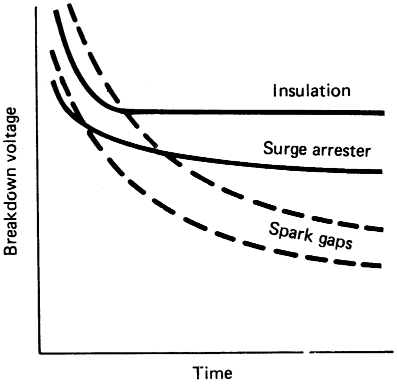


Figure 35.44 Typical voltage/time relationships for plant insulation and overvoltage protection devices

35.4.1 Insulation co-ordination

International and national standards for phase-to-earth insulation are referred to three system voltage ranges, *viz.* 1–36, 52–245 and 300–765 kV r.m.s. (Standards for other than phase-to-earth insulation are under consideration.) *Tables 35.1* to *35.3* give the standard voltage levels in terms of:

V_0 highest system voltage (in kilovolts r.m.s.) at which the plant and the protective equipment have to operate normally;
ν_{lp} rated lightning-impulse withstand voltage (in kilovolts peak);
υ_{sp} rated switching-impulse withstand voltage (in kilovolts peak); and
V_1 rated power-frequency short-duration withstand voltage (in kilovolts r.m.s.).

Lower range *Table 35.1* gives voltage values in two series based on current practice (I) in Europe and (II) in the USA and Canada, and applicable to other countries within the sphere of influence of (I) or (II). There are two lists in (I), the choice between them being made on consideration of the degree of exposure to lightning and switching impulse voltages, the type of system neutral earthing and (where applicable) the type of overvoltage protective devices. In (II) the data are based on the apparent power rating of the plant protected.

Middle range In *Table 35.2* more than one level of υ_{lp} is given, the highest being for plant in systems where the earth fault factor exceeds 1.4.

Higher range *Table 35.3* specifies no power-frequency values, the switching overvoltage having priority. The rated values of ν_{lp} associated with a standard ν_{sp} have been chosen in accordance with: (a) for plant protected by

Table 35.1 Insulation co-ordination, 1–36 kV

V_0 (kV *r.m.s*)	ν_{lp}		V_1 (kV *r.m.s.*)
	(kV *peak*)	(kV *peak*)	
Series I	*List 1*	*List 2*	
3.6	20	40	10
7.2	40	60	20
12	60	75	28
17.5	75	95	38
24	95	125	50
36	145	170	70
Series II	*≤500* kV-A	*>500* kV-A	
4.4	60	75	19
13.2–14.5	95	110	34
26.4	150	150	50
36.5	200	200	70

Table 35.2 Insulation co-ordination, 52–245 kV

V_0 (kV *r.m.s.*)	ν_{lp} (kV *peak*)	V$_1$ (kV *r.m.s.*)
52	250	95
72.5	325	140
123	450, 550	185, 230
145	550, 650	230, 275
170	650, 750	275, 325
245	750, 850, 950, 1050	325, 360, 395, 460

Table 35.3 Insulation co-ordination, 300–765 kV

V_0 (kV *r.m.s.*)	ν_{sp} (kV *peak*)	V_1 (kV *peak*)
300	750	850, 950
300	850	950, 1050
362	850	950, 1050
362	950	1050, 1175
420	950	1050, 1175
420	1050	1175, 1300, 1425
525	1050	1175, 1300, 1425
525	1175	1300, 1425, 1550
765	1300	1425, 1550, 1800
765	1425	1550, 1800, 2100
765	1550	1800, 1950, 2400

surge arresters the two lowest values apply, and (b) for plant not—or not effectively—protected by arresters only the highest value applies.

Several insulation levels may exist in the one power network, appropriate to different situations or to the characteristics of different protective equipments.

35.4.2 Protective equipment

35.4.2.1 Lightning

Underground cables are virtually immune from direct lightning strokes, so that preventive measures apply in general only to overhead lines.

Earth wire Shielding the line conductors by an earth wire is reasonably effective in preventing a direct stroke to the conductors, provided that the conductors lie within a segment subtending an angle of about 45° (or preferably 25° for towers above 50 m high) from the earth wire to the ground. Owing to the complex (and not fully understood) nature of the leader stroke as it approaches the earth, such protection sometimes fails; in cases of particular importance, e.g. near a major substation or where lightning is particularly prevalent, two earth wires may be installed.

Tower footing resistance A lightning stroke to the earth wire can produce a back-flashover to the line conductors resulting in a line surge voltage unless the tower footing resistance is very low, e.g. not greater than 1 Ω per 100 kV of impulse level. In difficult situations a buried *counterpoise* earthing system, comprising wires radiating from the foot of the tower to a distance of 30–60 m, or continuous wires from tower to tower, help to lower the footing resistance.

35.4.2.2 Switching surges

While switching surges cannot be entirely avoided, it is desirable on lines of 300 kV upward to limit them by shunting across circuit-breaker contacts, during closure, resistors of value approximating to the surge impedance of the line (300–400 Ω).

35.4.2.3 Damage to plant

Overhead-line damage by overvoltage can occasionally occur, but is generally repairable fairly quickly, but damage to substation plant (particularly to transformers) is costly and results in a long outage. Overvoltage protection is thus aimed primarily at the defence of transformers.

Surge attenuation A travelling wave of voltage attenuates (mainly by corona loss) as it is propagated along a line, and its magnitude and wavefront steepness are mitigated. Inductive coupling with the earth and with earth or counterpoise wires assists in this process. A reduction by one-half may thus occur in a line length of 5–8 km.

A length of underground cable between the terminal of an overhead line and the substation plant also reduces the magnitude of an incoming voltage surge, owing to the lower surge impedance of the cable. Additional reflections from the junctions may, however, partially offset this advantage and the cable is itself vulnerable. Although such a section of cable is often desirable for amenity or other reasons, it is rarely installed solely for protective purposes.

Surge arrester The aim is to direct a surge to earth before it reaches a vulnerable plant. The arrester must be located as near as is practicable to the terminals of the plant. An ideal arrester (a) takes no current at normal system voltage, (b) establishes an instantaneous path to earth for any voltage that is abnormally high, (c) is capable of carrying the full discharge current and (d) inhibits subsequent power-frequency current. The modern device comprises an assembly of small gaps and non-linear resistors in series, the whole being contained in a cylindrical porcelain housing. The use of multiple rather than single gaps gives the most rapid breakdown. The resistor elements offer a low resistance to surge currents (limiting the voltage across the arrester) and a much higher resistance to the power-frequency follow current. Careful gap design ensures that the follow current does not restrike.

Rod gap A plain air-gap is cheap, and it satisfies requirements (a), (b) and (c) above. However, it does not fulfil condition (d), and although gap breakdown protects the plant from overvoltage, there may be a power-frequency follow current which must be cleared by a circuit-breaker operation, involving an outage. A plain rod gap, connected between line and earth, has the following typical gap lengths giving breakdown at about 80% of the plant impulse level:

System voltage (kV)	36.2	72.5	145	300	420
Gap length (m):	0.23	0.35	0.66	1.25	1.70

35.4.2.4 Prevention of outages

About 80% of the earth faults on overhead lines are of a transient nature resulting from lightning, birds or other causes, and no damage to the equipment is caused; however, the ionised path caused by the transient flashover permits a power-frequency follow current to flow which must be cleared by a circuit-breaker operation. Such outages can, however, be avoided in certain cases by the use of arc suppression coils (Petersen coils) or auto-reclose circuit-breakers.

Arc suppression coil For reasons of insulation economy and/or safety it is customary to earth the neutral point of a system either directly or through a resistor. Any earth fault results in fault current and necessitates a circuit outage to clear it. Were the neutral point isolated, it might be expected that an earth fault would result in no fault current and that

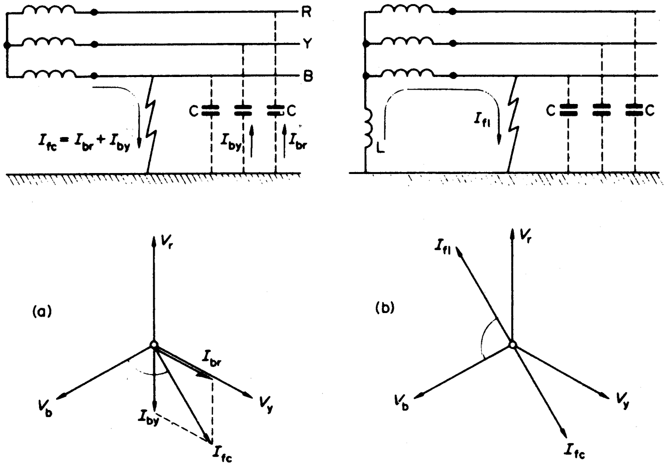


Figure 35.45 Action of arc-suppression coil: (a) isolated neutral; (b) neutral earthed through arc-suppression coil

the system could continue to operate, provided that the raising of the two sound phases to line voltage above earth potential were acceptable. However, owing to the system capacitance a leading current will flow through the fault as shown in *Figure 35.45*(*a*). If the fault involves an arc, any such current, if it exceeds a few amperes, is likely to cause damaging voltage surges and to be very persistent. Operation with a completely isolated neutral point is therefore impracticable except on very small isolated systems.

If, however, the neutral is earthed through a high inductive reactance X (several hundred ohms), the resulting lagging current can, neglecting losses, precisely neutralise the capacitance current, as shown in *Figure 35.45*(*b*). There will thus be no fault current and the system can be operated with the fault until such time as it can conveniently be repaired. To secure precise balance the value of the arc suppression coil reactance must be $X = 1/3\omega C$.

Arc suppression coils are effective on 12 kV, 36 kV and occasionally up to 245 kV systems, but at the higher voltages and with long lines the resistive and other losses prevent precise phase opposition of I_{fl} and I_{fc} so that there is a resultant current in the fault which, if it involves an arc, is inextinguishable and damaging.

Auto-reclose circuit-breaker A transient flashover rarely causes damage if the fault is cleared by the normal protective equipment. After an interval of 0.4–0.8 s, sufficient for the natural de-ionisation of the arc path, the circuit-breaker can be reclosed with safety.

Circuit-breakers with automatic reclosure was widely used on 12 kV and 36 kV radial distribution networks. They are required to provide time delay switching in the event of a permanent fault; permit of normal fuse discrimination to limit the area of interruption; operate rapidly on the initial passage of a fault current. *Figure 35.46*(*a*) shows three applications. In (i) all transient faults are cleared by the recloser, and in the case of permanent faults the recloser sequence provides a time delay trip to blow the fuse only on the faulty sub-circuit. In (ii) sectionalisers provide the disconnecting means for faulty sub-circuits. Method (iii) operates as (i), except that on permanent fault the recloser holds closed and back-up protection operates.

A typical current operated recloser comprises a normally closed oil circuit-breaker for pole mounting. The breaker is held closed by springs. When the current exceeds, for example, twice full-load value, movement of the plunger of a series solenoid causes the recloser to open. The plunger is then spring-reset and the reclosing is automatic. Relay features control the reclosing and opening times, and a mechanism is provided to enable the recloser to lock open or hold closed at the end of its operating sequence.

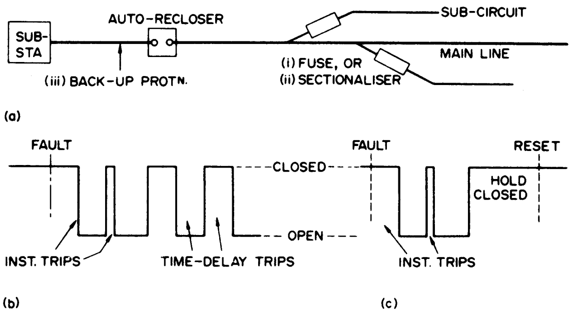


Figure 35.46 Auto-reclosing: (a) system; (b) lock-open sequence; (c) hold-closed sequence

Figure 35.46(*b*) shows a lock-open sequence, the open-circuit times being adjustable between 0.25 and 1 s. If the fault clears and operation is stopped during the cycle, the mechanism returns to the starting position. The ‘instantaneous’ trips are rapid (2–7 cycles) to avoid fuse deterioration: they give two chances for fault clearance. The time delay switching occurs twice to blow the fuses on a permanently faulted section. If the sequence continues, the operation terminates with the recloser locked open. In *Figure 35.46*(*c*) is shown a sequence ending in a hold-closed condition. A permanent fault is cleared by other means, after which the mechanism automatically resets. Transient faults are cleared as with type (*b*).

Where adequate fuse co-ordination is not obtainable, the alternative is the use of sectionalisers, as at (ii) in *Figure 35.46*(*a*), associated with automatic reclosers. A sectionaliser comprises a normally closed oil switch latched in against spring loading. When a fault occurs, the current passes through a series coil which actuates the mechanism and counts the fault current impulses during the recloser sequence, opening the sectionaliser automatically during a pre-set open-circuit period. The mechanism resets if the fault clears prior to the end of the sequence, but if the sectionaliser reaches the open condition, it must be reset by hand.

Auto-reclosing is also sometimes used on transmission systems to assist in maintaining stability by reclosing before synchronous machines have had time to lose synchronism. In some cases, particularly on single-circuit lines, the added complication of selecting, tripping and reclosing only the faulty phase is justified.

When the loss of supply may have serious consequences, duplicate supplies or a ring main may be installed.

Bibliography

ALSTOM, T & D, *Protective Relays Application Guide*, Alstom T & D, Protection and Control Ltd (1987)

BASAK, P. K. *et al.*, 'Survey of TRV conditions on the CEGB 400 kV system', *IEE Proceedings*, **128**(6), 342–350

BLOWER, R. W. *et al.*, *Experience with Medium Voltage Vacuum Circuit-breaker Equipment* (*IEE Conference Publication No. 197*), 51–55 (1981)

BOGGS, S. A. *et al.*, *Coupling Devices for the Detection of Partial Discharges in Gas-Insulated Switchgear* (*IEEE PES Paper 81 WM*)

CHAMIA, M., 'Transient behaviour of instrument transformers and associated high-speed distance and directional comparison protection', *Electra*, **72**, 115–139 (October 1980)

CHISHOLM, W. A. *et al.*, 'Lightning performance of overhead lines', *Ontario Hydro Review*, **3**, 19–24 (January 1981)

CHOWDHURI, P., *Electromagnetic Transients in Power Systems*, Wiley (1996)

DAVY, R. A. *et al.*, 'State of the art in protection and control for e.h.v. transmission', *Electrical Review*, **207**(19) 27–29 (21 November 1980)

DONON, J. and VOISIN, G., *Factors Influencing the Ageing of Insulating Structures in SF*$_6$, (*CIGRE Paper 15-04*), 11 (1980)

ELECTRICITY TRAINING ASSOCIATION, Power System Protection, a four volume text, *IEE.* (1997)

FLURSHEIM, C. H., Power Curcuit Breakers—Theory and Design, Revised Edition, *IEE.* (1982)

FORREST, J. S., 'Electricity supply—present and future', *IEE Electronics and Power*, **29**(11/12)

GARDNER G. E. *et al.*, 'Development testing techniques for air-blast and SF$_6$ switchgear', *IEE Proceedings*, *Part C*, **127**(5), 285–293 (September 1980)

IEE, *Developments in Design and Performance of EHV Switching Equipment* (*IEE Conference Publication No. 182*), Peter Peregrinus

IEE, *Developments in Power System Protection* (*IEE Conference Publication No. 185*), Peter Peregrinus

IEE, *Electricity Distribution* (*CIRED 1981*), (*IEE Conference Publication No. 197*), Peter Peregrinus (1981)

JOHN, M. N., 'Electricity supply—problems and possibilities', *IEE Electronics and Power*, **29**(10), 702–704

JOHNS, A. T. and SALMAN, S. K., Digital Protection for Power Systems, *IEE*, Power Engineering Books Series (1995)

JOHNS, A. T. and WALKER, E. P., 'Co-operative research in the engineering and design of a new digital directional comparison scheme', *Proceedings IEE*, *Part C*, **135**(4), (July 1988)

KEINERT, L. *et al.*, *Service Experience with, and Development of SF*$_6$ *Gas Circuit-Breakers Employing the Self-Extinguishing Principle* (*IEE Conference Publication No. 197*), 36–40 (1981)

KRIECHBAUM, K., 'Progress in h.v circuit-breaker engineering', *SRBE Bulletin*, **96**(3–4), 147–158 (1980)

LAT, M. V. *et al.*, 'Distribution arrester research', *IEEE PES Paper H WM 199–9*, 9 (1981)

LATHAM, R. V., *High-voltage Vacuum Insulation—the Physical Basis*, Academic Press, London (1981)

MADDOCK, B. J. *et al.*, 'Optical fibres in overhead power transmission systems for communication and control', *Proc. 29th International Wire and Cable Symposium, USA*, 402–409 (1980)

MALLER, V. N. *et al.*, *Advances in H.V Insulation and Arc Interruption in SF*$_6$ *and vacuum*, Pergamon, Oxford (1981)

MILLER, T. J. E., *Reactive Power Control in Electric Systems*, Wiley Interscience, New York (1982)

PHADKE, A. G. and THORP, J. S., *Computer Relating for Power Systems*, Research Studies Press, Wiley, New York (1988)

RAGALLER, K. (Ed.), 'Surges in h.v. networks', *Proc. of Brown Boveri Symposium*, 423 (1979)

ROSEN, P. *et al.*, 'Recent advances in HRC fuse technology', *IEE Electronics and Power*, **29**(6), 495–498

SONG, Y. H., JOHNS, A. T. and AGGARWAL, R. K., *Computational Intelligence Applications in Power Systems*, Science Press, Kluwer Academic Publishers (1996)

STAMACKA, E., 'Interruption of small inductive currents', *Electra*, **72,** 73–103 (October 1980)

TEDFORD, D. J. *et al.*, 'Modern research and development in SF$_6$ switchgear', *IEE Electronics and Power*, **29**(10), 719–723

'The use of ZnO surge diverters in the Austrian e.h.v. network', *Oesterrich Z. Electr.*, **33**(12), 433–437 (December 1980) (in German)

WRIGHT, A. and NEWBERY, P. G., *Electric Fuses*, 2nd Edition, *IEE*, Power Engineering Books Series (1995)

36

Electromagnetic Transients

R Aggarwal

Contents

36.1 Introduction 36/3

36.2 Basic concepts of transient analysis 36/3
  36.2.1 Lightning surges 36/3
  36.2.2 Switching surges 36/5
  36.2.3 Overvoltage induced by faults 36/5
  36.2.4 Resonance 36/7

36.3 Protection of system and equipment against transient overvoltage 36/7
  36.3.1 Protection of transmission lines against lightning 36/7
  36.3.2 Surge protection 36/7
  36.3.3 Transient voltages and grounding practices 36/8
  36.3.4 Insulation co-ordination 36/8

36.4 Power system simulators 36/8
  36.4.1 The transient network analyser (TNA) 36/9
  36.4.2 Digital computer simulation 36/9
  36.4.3 Digital real-time electromagnetic transient simulator 36/9

36.5 Waveforms associated with the electromagnetic transient phenomena 36/10
  36.5.1 Introduction 36/10
  36.5.2 Transient waveforms 36/10

36.1 Introduction

Although a power system operates in the steady state for a large part of its time, it must be capable of withstanding the overvoltage and overcurrent stresses generated during transient conditions. The size of transmission line towers, the clearance for conductors, the insulation of windings, the rating of circuit-breakers, the loading capabilities of lines, cable, etc., and the operating performance of protection relays are all dictated by consideration of the amplitude and frequency of the power system transients. Consequently, to design and operate a power system to deliver a reliable and high quality power supply, it is crucial to have a thorough understanding of its transient behaviour.

In terms of modelling and analyses of power system transients, since they usually consist of multiple frequencies, phasor representations of voltages, currents and impedances are no longer valid. Furthermore, representation of a power system by a single-line diagram derived from say, symmetrical components, is no longer appropriate since the latter are derived from steady state phasor analysis.

The foregoing analysis problem is compounded by the fact that a single physical component in power system network may have a different model representation (depending on a specific transient phenomenon being investigated) in order to achieve the requisite accuracy. For example, a transmission line (or an underground cable) may be modelled by a short bus section, an inductance, a resistance, a series RL circuit, an RLC II or T section (or multiples of such sections cascaded together), or a fully distributed line/cable model. Likewise, a transformer or a shunt reactor may be modelled by simply an inductance, or a network of capacitances or by a combination of the two.

Transient phenomena in power systems is caused by such disturbances as switching operations, faults including those induced by lightning strikes, partial discharges due to defects in power equipment, etc. They involve a frequency range from d.c. to several MHz. In this respect a rudimentary distinction is usually made between electromechanical transients, traditionally covered by transient stability studies, and electromagnetic transients. The latter types of transients can occur on a time scale ranging from microseconds to several cycles. They are a combination of travelling waves on overhead lines, underground cables and buses, and oscillations in lumped element circuits of generators, transformers and other devices. Some slower electromechanical transients, such as subsynchronous resonance (associated for example with series capacitor compensated lines), for which detailed machine models are needed, are usually included in this class of transients.

Transients may result in overvoltages, overcurrents, distorted waveforms or electromechanical oscillations. In general, a single event will cause all of these effects, but depending upon the requirements, a particular one might be of greatest concern. Briefly, the aforementioned four categories within the transient phenomena can be described as:

Overvoltages The design and insulation requirements of power equipment is determined by the overvoltages it must withstand. The overvoltage could be quasi-steady state (principally 50/60 Hz) or due to the short duration high frequency phenomena. Examples of the former are voltage excursions due to load rejection or loss of shunt reactive compensation. Examples of the latter are the voltage surge due to a lightning strike or the overvoltages caused by the opening and closing of a circuit-breaker. In this respect, a voltage surge generated by lightning may have a wavefront as steep as a fraction of a microsecond and may last for a few microseconds. Alternatively, switching surges may have wavefronts of several microseconds and may persist for a few power frequency cycles. It is thus apparent that if the effect of lightning is being studied, the system model must be capable of accurately simulating the power system transients at frequencies up to hundreds of kHz (or even MHz); for studying switching operations, frequencies up to tens of kHz (or even 100 kHz) will be of interest.

Overcurrents Overcurrents result from system faults and their knowledge helps to determine the performance of protection relays the required interrupting duty of circuit-breakers, and mechanical and thermal stresses on power equipment. Since a power system is mainly inductive in nature, the transients with short circuit currents are largely restricted to the lower frequency. However, it must be mentioned that although higher frequency transients are also present, they are relatively of much smaller magnitudes in comparison to the lower frequency current transients and/or high frequency voltage transients.

Distorted waveforms An accurate knowledge of the shape of the voltage and current signals during either transient or abnormal steady state operating conditions is of considerable importance. For example, when testing a high speed protection relay, the distortion in the voltage and current waveforms immediately after a fault will affect the operating time or stability of the relay. In addition, some protection relays will be affected by the harmonics generated by d.c. converters or non-linear circuit elements.

Electromechanical oscillations One particular type of such oscillations, viz, subsynchronous resonance type, are caused by a combination of the series capacitors and natural line inductance; this resonance can exacerbate the torsional oscillations on the generator shaft to cause mechanical failure. Such oscillations thus require detailed simulation models.

This chapter deals with several aspects associated with the electromagnetic transients. It describes in some detail some of the fundamental concepts for transient analysis and an outline of the tools available for simulating such transients through setting up models of the various components within a power system network. An understanding of the transient phenomena is enhanced by an illustration of typical voltage and current transient waveforms.

36.2 Basic concepts of transient analysis

As mentioned before, the main causes of power system transients are:

- lightning surges;
- switching surges;
- faults; and
- resonance.

36.2.1 Lightning surges

Lightning is the greatest single cause of line outages; for example, it accounts for about 70% of outages on the high voltage transmission system (275 kV, 400 kV, 500 kV, etc.). The physical phenomenon of lightning shows that clouds acquire charge, or at least become polarised. Electric fields become excessive to the extent that the dielectric of the intervening space can no longer support the electrical stress and breakdown or lightning flashover occurs; this is usually a high-current discharge.

The lightning strikes that create problems for the power engineers are those that terminate on or near to power lines. These can be looked upon as equivalent to closing a switch between the cloud and the power line or adjacent earth, thus constituting a circuit change condition. This is either a direct connection to the line or it completes a circuit with close mutual coupling to the line. Direct consequences of this phenomenon are:

- very often the line will be raised to such a potential that further flashes will occur to grounded structures;
- the grounded structures may be raised to such a potential that they flash over to the line.

The disturbance created on a power line due to the lightning phenomenon involves *travelling waves*. These are essentially voltage surges which, although of short duration, can nonetheless cause overvoltage well in excess of the insulating capabilities of power lines, thus posing a serious threat of damage to expensive equipment and causing disruption.

The basic wave shape of a lightning strike is the 1.2/50 μs wave and typifies the lighting surge. This is shown in *Figure 36.1*.

Figure 36.1 represents a current waveform which rises in 1.2 μs and falls to half the peak value in 50 μs. Normally only heavy current flowing over the first 50 μs is of importance and the magnitude of the peak current ranges from 20 kA to 200 kA. The current/time relationship for the aforementioned wave shape is given by:

$$i = I_{\text{peak}}\left(e^{-\alpha t} - e^{-\beta t}\right) \qquad (36.1)$$

where t is in μs. The values of the constants α and β depend on the nature of the surge and are typically $\alpha = 0.002$ and $\beta = 3.0$.

When a lightning strike arrives on an overhead conductor, equal current surges of the waveform shown in *Figure 36.1* are propagated in both directions, away from the point of impact. The magnitude of each voltage surge set up is therefore given by:

$$V = \frac{Z_0 I_{\text{peak}}\left[e^{-\alpha t} - e^{-\beta t}\right]}{2} \qquad (36.2)$$

where Z_0 is the conductor surge impedance.

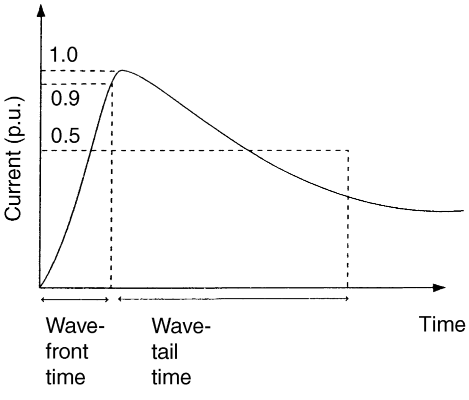


Figure 36.1 Typical waveform due to lightning

The voltage surge set up depends upon the effective surge impedance of the conductors into which the current discharges (and, of course, on the peak current) but this value is rarely less than about 3 MV peak. Such a voltage is well above the insulating capability of transmission line insulators. It would, however, be economically nonviable to design them to withstand such high voltages (it can be shown that cost $\propto V^2$). In this respect, overhead shield (or earth) wires are used, which largely prevent direct strikes to the phase conductors.

When an earth wire exists over the overhead line, a strike arriving on tower or on the wire itself sets up the surges in both directions along the wire. On reaching neighbouring towers, they are partially reflected and transmitted further and this process continues over the length of the line as and when towers are encountered. Under this condition, the voltage surges set up on the main phase conductors are significantly smaller than would be the case if the earth wire were absent. This is because, in the presence of an earth wire, the voltage surges set up on the phase conductors are due to the mutual coupling effect between the earth wire and the phase conductors. The coupling factor is typically in the range of $0.15 < k < 0.3$.

The initial voltage surge generated on the earth wire depends upon both the tower impedance (Z_T) and the earth wire surge impedance (Z_{EW}). In this respect, different tower designs present different surge impedances. For example, if the tower tops are connected by a single earth wire of surge impedance Z_{EW} then the effective surge impedance is given by:

$$Z_{TE} = \frac{Z_T \times 1/2 Z_{EW}}{Z_T + 1/2 Z_{EW}} \qquad (36.3)$$

The half is included with the Z_{EW} since a wave propagates in both directions from the stricken point.

Current and voltage waves radiate from the point of contact in both directions along the earth wire and down the tower if the tower is involved. These waves rapidly encounter discontinuities such as adjacent towers in the case of earth wires or footing resistance in the case of the tower.

A direct consequence of this phenomenon is that reflected waves are initiated, which cause yet other waves when they return to the stricken point. The effect of these waves will depend on the change of surge impedance at the point of discontinuity. For example, if the initial wave going down the tower encounters a low footing resistance R, then the reflected wave will be of opposite sign and will act to reduce the tower potential. The reverse will occur if the footing resistance is high.

Transmission line theory[1] shows that reflection coefficient ρ_G at ground is given by:

$$\rho_G = \frac{R - Z_T}{R + Z_T} \qquad (36.4)$$

Also of interest are the refracted (β_E) and reflected (ρ_E) coefficients experienced by the reflected waves returning to the tower top from the ground. Again, using transmission line theory, these can be shown to be:

$$\beta_E = \frac{Z_{EW}}{1/2 Z_{EW} + Z_T} \qquad (36.5)$$

$$\rho_E = \frac{1/2 Z_{EW} - Z_T}{1/2 Z_{EW} + Z_T} \qquad (36.6)$$

36.2.1.1 Back flashover

The potential difference across the suspension insulators is of particular concern since a flashover can occur and a fault is placed on the phase if this voltage becomes excessive. Waves travelling along the earth wire induce waves on the phase conductors, the conductor closest to the earth wire experiencing the highest induced voltage. The effect of this coupling between the earth wire and the phase conductors (typically between 0.15 and 0.3) is to lessen the stress on the line insulators. The induced voltages add to, or subtract from, the power frequency voltages. At any instant, at least one phase will have the same polarity as the lightning surge; such a phase is more likely to flashover. This phenomenon is more commonly known as *back flashover*.

It is desirable, therefore, to have a lower footing resistance. There are two considerations:

- the local resistivity of the earth itself; and
- the connection that is made between the tower and ground.

A typical value for the tower footing resistance R is 25 Ω. After a lightning strike, the time taken for the voltage surge to travel from the tower top to the tower footing and back at the speed of light is typically 0.25 µs. The resulting magnitude of the voltage surge is thus reduced to typically about 500 kV which is far less daunting than the above 3 MV surge generated at the instant of a lightning strike.

36.2.1.2 Summary

- Lightning-induced flashovers are the greatest cause of line outages.
- Overhead earthwires reduce the magnitude of phase conductor voltage surges.
- Low tower footing resistance will reduce tower surge potential and the likelihood of back flashover.

36.2.2 Switching surges

36.2.2.1 Interruptions of short-circuits and switching operations

When a fault occurs on a transmission line, due for example to a lightning strike, then in order to clear the fault (i.e. if the fault is transient in nature) and at the same time minimise damage, it is normal practice to isolate the faulted conductor(s). This is done by tripping the nearest circuit-breakers, once the fault has been detected by protective relays. However, no circuit which contains an appreciable inductance can have the current in it broken instantly. As power circuits are always inductive, a high induced voltage will occur across any break in the circuit at the instant of current interruption. This voltage is sufficient to ionise the gas molecules in the gap between circuit-breaker contact and conductor, resulting in an arc which strikes between circuit-breaker contact and conductor at the instant of a separation.

It is important to note that the extinction of arc current on opening of a particular circuit-breaker will only remain permanent if the restrike voltage is less than the recovery voltage (see *Figure 36.2*).

Note:

- Restrike voltage is the voltage across the contacts of a circuit-breaker at current extinctions.
- Recovery voltage is the voltage that can be withstood across the contacts of a circuit-breaker after current extinction. It increases with time. Re-ignition voltage is another term for recovery voltage.

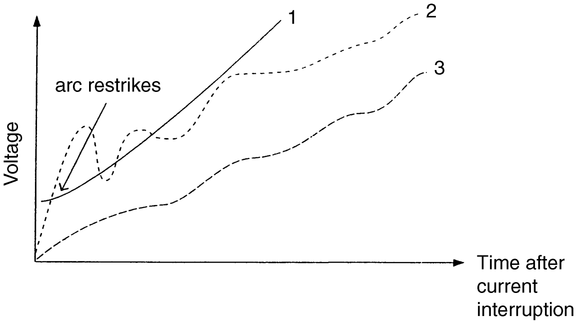


Figure 36.2 Breaker restrike voltage characteristics: 1—breaker recovery voltage, 2—restriking voltage transient with initial rapid rate of rise, 3—restriking voltage transient that does not cause a restrike

The restriking voltage transient is very largely independent of the breaker characteristics and very much depends upon the transient response of the system with respect to a particular breaker location.

36.2.2.2 Current chopping

Current chopping is the extinction of arc current before natural current zero. It is very important that circuit-breakers are designed so that they do not prematurely *chop* the fault current. Unfortunately, current chopping does arise with certain types of circuit breakers, namely the air-blast circuit breakers which operate on the same air pressure and velocity for all values of interruption current.

On low-current interruption, the breaker tends to open the circuit before the current natural-zero and the electromagnetic energy present is rapidly converted to the electrostatic energy.

36.2.3 Overvoltage induced by faults

Overvoltage may be produced by certain types of asymmetrical fault such as a ground fault on one of the phase conductors of a three-phase transmission line. The situation is somewhat analogous to the switching transients except that here instead of injecting a current, a voltage equal and opposite to the prefault voltage at the fault point is applied (see *Figures 36.3*(*a*)–(*d*)).

- *Figure 36.3*(*a*) shows the faulted system in which a ground fault has occurred on the 'a'-phase at point F.
- *Figure 36.3*(*b*) shows the prefault steady-state voltage V_{fSa} at the fault point.
- If it is assumed that a fault occurs at the peak of the prefault voltage and $t = 0$ at this instant, then the voltage injected at the fault point is as shown in *Figure 36.3*(*c*).
- *Figure 36.3*(*d*) shows the de-energised network to which is applied the voltage V_{ffa}. This is also known as the *superimposed* voltage.

We can study the problem of overvoltage by asking how the de-energised network behaves in response to the application of this voltage. Complex analysis based on travelling-wave theory (such as Bewley lattice diagrams[2]) can be used to depict the level of overvoltage likely to be induced, particularly

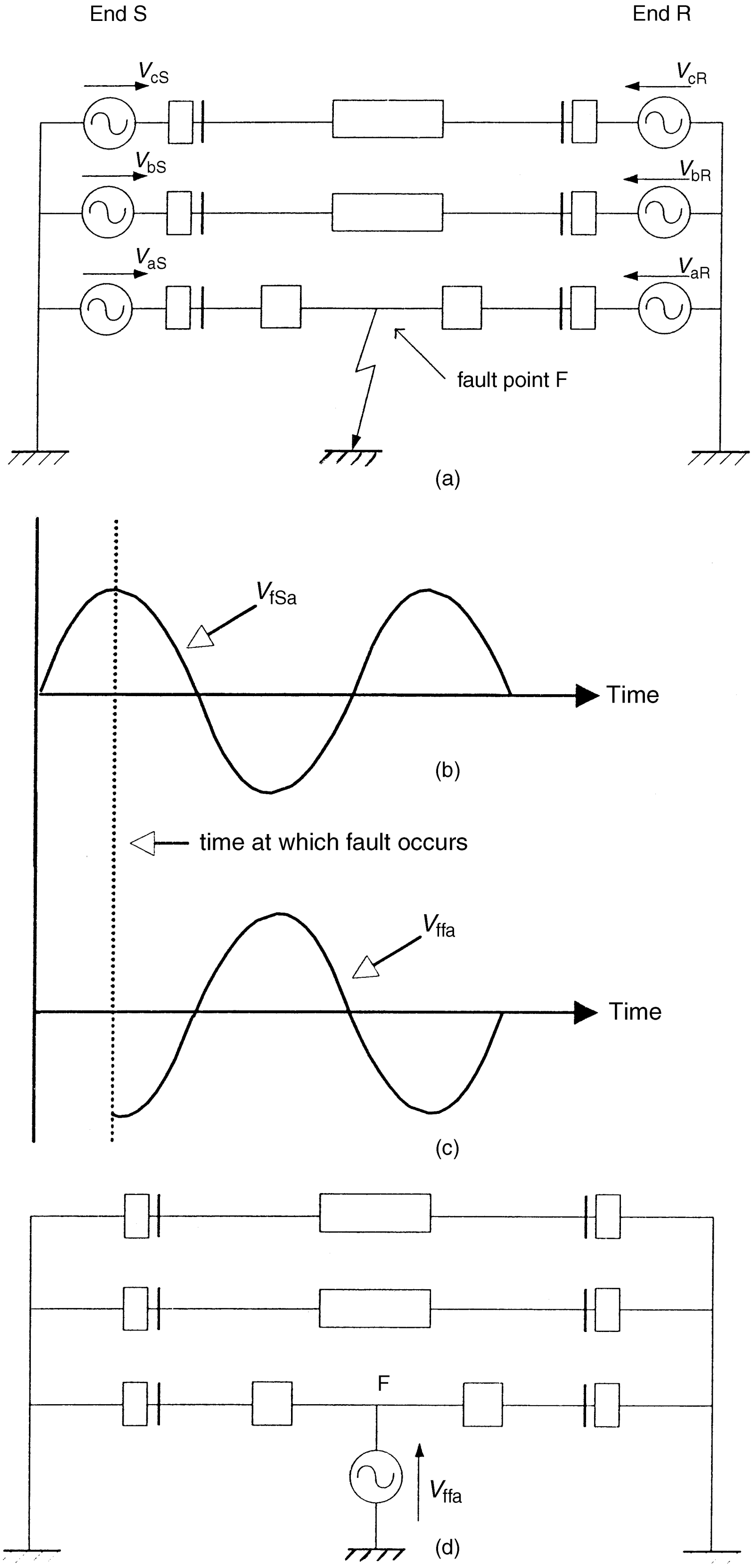

End S
End R
V_{cS}
V_{cR}
V_{bS}
V_{bR}
V_{aS}
V_{aR}
fault point F
(a)
V_{fSa}
Time
(b)
time at which fault occurs
V_{ffa}
Time
(c)
F
V_{ffa}
(d)

Figure 36.3 A faulted three-phase system: (a) faulted network; (b),(c) faulted network voltage; (d) superimposed network

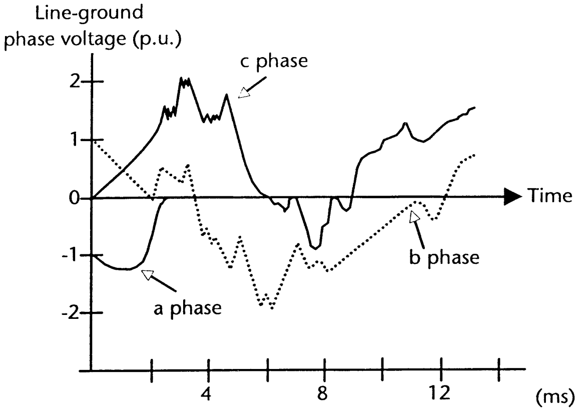


Figure 36.4 Overvoltage phenomena on a faulted three-phase system

on the unfaulted phases due to the injection of the suddenly applied voltage at the fault point. A line–ground fault of the type considered here can produce an overvoltage on an unfaulted phase as high as approximately twice the normal line-to-neutral voltage.

Figure 36.4 typifies an example of such an overvoltage, in particular when the fault occurs at the midpoint of the line; then maximum voltage is at the midpoint of the unfaulted conductor.

A more comprehensive set of transient voltage and current waveforms associated with different forms of electromagnetic disturbances will be illustrated later on in this chapter.

In practice, since insulator chains are designed to withstand approximately 1.4 times the normal steady-state voltage, there is a likelihood of a fault being induced on a healthy phase (this is also commonly known as an evolving fault) and this can have a detrimental effect on systems employing single-pole tripping.

36.2.4 Resonance

It is well-known that, in resonant circuits, severe overvoltage occurs, dependent on the resistance present. For example, voltage at resonance across the capacitance can be high. Although it is unlikely that resonance in a supply network can be obtained at normal supply frequencies, it is possible to have this condition at harmonic frequencies. Resonance is normally associated with the capacitance to earth of items of plant and often occurs on an opened phase due to a broken conductor or a fuse operating.

In circuits containing windings with iron cores, e.g. transformers, a condition due to the shape of the magnetisation curve known as *ferroresonance* is possible. This can produce resonance with overvoltage and also sudden changes from one condition to another with the possibility of an overvoltage surge.

The problems associated with subsynchronous resonance have already been briefly discussed before.

36.3 Protection of system and equipment against transient overvoltage

Surges in power systems can be very destructive. It is important to protect power equipment against them wherever possible, consistent with sound economics.

- Firstly, we must strive to prevent lightning surges gaining entry to the systems.
- Secondly, we must minimise the effects of those that do enter.

We must accept switching surges—they are the inevitable consequence of normal system operation. However, we can mitigate their effect by careful design.

The protection of terminal equipment in substations is of particular concern because damage there can be very costly in replacement and in outage time. A family of surge protection devices has been developed over the years which assist greatly in this task.

Surges can intrude into control circuits causing destruction and/or maloperation of relays, measuring equipment, monitors, etc.

36.3.1 Protection of transmission lines against lightning

As mentioned before, it is common practice to protect overhead lines against lightning by the employment of one (or more) earth wires that are strung from tower to tower above the phase conductors. These wires are bonded to the towers so that they are essentially at ground potential under normal conditions.

The intent in the disposition of these wires is to have them intercept lightning strikes that would have otherwise terminated on the phase conductors themselves, the effects of induced surges from indirect strikes is significantly lessened.

An aspect of vital importance, quite apart from the prevention of damage, is the maintenance of supply. Most flashovers do not cause permanent damage and therefore a complete and lasting removal of the circuit from operation is not necessary. Nowadays this is achieved by the use of *autoreclosing* circuit-breakers.

36.3.2 Surge protection

Inevitably, even the best shielding schemes will either not be fully effective or fail from time to time. Some switching operations and faults may generate overvoltage which can be potentially hazardous to the insulation of the plant. To protect the equipment against these contingencies, various devices have been developed. These devices are placed in parallel with the plant to be protected. Their purpose is to ensure that under no circumstances will the voltage across them exceed what the protected object can support.

36.3.2.1 Rod gaps and resistance suppressors

In the case of a transmission line insulator, a surge diverter in the form of *rod gaps* is commonly used as a crude protection device by placing it in parallel with the conductor insulator. However, it does have the disadvantage that often, after the surge responsible for breaking the gap between the rods has discharged, a power arc is maintained, i.e. a fault is thrown on the circuit. It is true that this can be removed by opening circuit-breakers at the two ends of the line and the power then restored by reclosing the breakers after the rod gap has de-ionised (the commonly employed circuit-breaker autoreclosing technique). Certainly, this is preferable to damaging a costly piece of equipment and perhaps sustaining a lengthy outage; but a device that can effect the voltage limiting without creating a fault is obviously more attractive. The *non-linear* resistor is such a device. These resistors have the property that their resistance diminishes sharply as the voltage across them increases.

36.3.2.2 Lightning arrester

Circumstances will arise in which it is impossible to obtain the desired protection by using non-linear resistors. For example, if the ceiling voltage to which surges must be limited are ≤3-times the normal system voltage, the values of the resistance at system voltage may be so low that they make the steady-state losses prohibitive.

An improved but more expensive surge diverter is the lightning arrester. In such devices, the non-linear resistance elements are as before but they have a gap(s) in series with them. In this way, the resistor is isolated from the circuit in normal conditions and is introduced when a surge appears by the sparkover of the gap(s).

36.3.2.3 Surge capacitors and surge reactors

If a capacitor is placed in parallel with a piece of equipment it will provide some measure of protection against voltage surges—the surge can only impress a high voltage on the equipment as quickly as it can charge the capacitor. The effectiveness of the capacitor in holding down the voltage depends on the energy conveyed by the surge and the capacitance of the capacitor. Such capacitors are particularly effective against voltage *spikes*, i.e. short duration voltage surges. In this respect, the combination of a capacitor and some other protective device such as an arrester, can be very useful because they complement each other.

Surge reactors are sometimes used to protect voltage surges, particularly within pieces of a.c./d.c. conversion equipment, in a manner that complements the use of surge capacitors. The capacitor parallels the equipment to be protected; the reactor is placed in series with the equipment. When a surge reaches such a series combination, the initial tendency is for the surge voltage to appear primarily across the reactor because this offers the highest impedance to the rapidly changing surge current.

Like capacitors, the ability of reactors to store energy is limited, but again like capacitors, they can take care of low energy surges, or act temporarily to protect a device until such time as a gap can flashover or an arrester operates.

36.3.3 Transient voltages and grounding practices

Good grounding practices can contribute quite significantly to more reliable and safer utility and industrial power systems. We are concerned with minimising the damage to equipment and false operations of equipment, under transient conditions and also with the avoidance of shock hazard to personnel.

Under steady-state, quiescent operating conditions in a power system:

- the integrity of equipment is preserved by adequate insulation; and
- the safety of personnel is secured by keeping sufficient separation between people and high-voltage conductors.

Preferably, a *grounded barrier* in the form of a metal enclosure, wire fence etc. is placed between the circuit and any personnel. In this respect, grounded barrier refers to a physical equipotential surface, at the same potential as local ground and any other grounded objects around. However, under abnormal conditions such as when a fault occurs, this situation may no longer prevail and *ground* becomes a relative thing.

In industrial plant, ground fault currents are usually confined to the specially laid ground mat of the building and to cable sheaths. In utility systems, *true* ground is often involved and it is vitally important to connect the buried mat to ground via metal rods (typically 2–3 m long) into the ground. The effectiveness of these rods is critically dependent on the resistivity of the soil; it varies widely with location and weather conditions, from a few thousand to hundreds of thousands of ohms cm.

36.3.4 Insulation co-ordination

The equipment used in a power system comprises items having different breakdown withstand voltages and different volt/time characteristics. In order to protect adequately all items of the system it is necessary to consider the situation as a whole and not items of plant in isolation, i.e. the insulation protection must be co-ordinated. To assist this process, standard insulation levels are recommended and are summarised in *Table 36.1* below.

36.3.4.1 Definitions

- Impulse refers to a transient voltage developed in a laboratory test and is an approximate representation of a surge that develops in nature.
- The term impulse is defined as the impulse peak voltage/crest value of the power frequency voltage to cause flashover.

36.4 Power system simulators

Simulation is one of the most powerful tools available to the power systems engineer when confronted with the need to study complex power systems phenomena. In this respect several tools have been used over the years to model and analyse electromagnetic transients. At early stages, miniature power system models, known as transient network analysers (TNA), were used. At present, the digital computer is the most popular tool, although TNAs are still used. In addition, the new generation of real-time digital systems are probably the most adequate tool in certain applications

Table 36.1 British substation practice

Nominal voltage (kV)	*Impulse* (peak kV)	*Power frequency withstand* (peak kV)	*Minimum clearance to ground* (m)	*Minimum clearance between phases* (m)	*Co-ord gap setting* (m)
11	100	29	0.2	0.25	2 × 0.03
132	550	300	1.1	1.25	0.66
400	1425	675	3.05	3.55	1.5–1.8

for which either a very high-speed or a truly real-time simulation is required.

36.4.1 The transient network analyser (TNA)

In a TNA, the power system components such as transmission lines, transformers, shunt reactors, circuit-breakers, etc. are scaled down to low current, low voltage, lumped inductance, capacitance and resistance circuits. A power network modelled using these circuits is then energised using generators modelled by ideal voltage sources behind *Thevenin* equivalent reactances.

The main advantages of a TNA are that the simulator operates in real time (providing analogue signals), there are no computationally unstable solutions and many users believe that a physical miniature model is closer to the actual system than a mathematical model.

The main disadvantages of a TNA are inflexibility, large capital cost (including regular maintenance cost), the time required to set up the system, limitations on the number/type of equipment models, problems associated with designing miniature inductors with the same '*R*' to '*L*' ratio as the actual power components, a drift in the component values due to ageing/environment and the finite length of lines that can be modelled using serially connected Π (or T) circuits.

36.4.2 Digital computer simulation

The studies to solve travelling wave problems by means of a digital computer were started in the early 1960s using two different techniques, the Bewley lattice diagram and the Bergeron method.[2,3] These techniques were applied to solve small networks with linear and non-linear lumped-parameter, and distributed-parameter elements. The extension to multi-node networks was made by Dommel[4]; this scheme essentially combined the Bergeron method and the trapezoidal rule into an algorithm capable of solving transients in single and multi-phase networks with lumped and distributed parameters. This solution was the birth of the now very widely employed and universally accepted Electromagnetic Transients Program (EMTP).

Essentially, the trapezoidal rule is used to convert the differential equations of the network components into algebraic equations involving voltages, currents and past values. The algebraic equations are assembled using a nodal approach given as:

$$[\mathrm{G}] \cdot [v(t)] = [i(t)] - [I] \qquad (36.7)\Leftarrow$$

where $[G]$ = nodal conductance matrix
$[v(t)]$ = vector of node voltages
$[i(t)]$ = vector of current sources
$[I]$ = vector of 'history' terms

Very often, the network contains voltage sources to ground; the equation is then split up into part E with unknown voltages and part F with known voltages given as:

$$[V_{\mathrm{E}}(t)] = [i_{\mathrm{E}}(t)] - [I_{\mathrm{E}}] \Leftarrow [G_{\mathrm{EF}}] \cdot [v_{\mathrm{F}}(t)] \Leftarrow \qquad (36.8)\Leftarrow$$

The resulting conductance matrix is symmetrical and remains unchanged as the integration is performed with a time-step size. The solution of the transient process is then obtained using triangular factorisation. One of the main advantages of this approach is that it can be applied to networks of arbitrary size in a very simple manner.

Bergeron's method can be efficiently used with lossless and distortionless lines. However, in practice, parameters of actual transmission systems are frequency dependent. Much effort has been expanded since then into developing alternative frequency-dependent line models[5,4] and these have been subsequently incorporated within the EMTP software. In this respect, it should also be mentioned that some researchers have overcome this problem by developing simulation techniques for transmission systems which are totally frequency-domain based such as the one developed by Johns and Aggarwal.[7] Although the latter approach gives a very accurate solution to the problem, the principal disadvantage is that for any time-domain circuit change (most circuit changes are time-domain based), the variable in question has to be transformed into its frequency spectrum first before a solution can be effected.

The drawback of the original Dommel's scheme was that it could be used to solve only linear networks. However, many power components such as transformers, reactors, surge arresters, circuit-breakers, etc., exhibit a non-linear behaviour. Several modifications to the basic method were thus proposed to cater for non-linear and non-stationary elements.[8]

The development of the very widely used EMTP software has been evolutionary in nature. Apart from the foregoing changes, a very important development was that of a section for representation of control systems, initially motivated by studies of HVDC links. The Transient Analysis of Control Systems (TACS) option was implemented in the EMTP in 1976.[9] Although the main goal was the simulation of HVDC converters, it soon became apparent that TACS had many other applications, a major attribute was in the fact that many externally developed non-linear models representing, for example, fault arcs (primary and secondary), dynamic arcs in circuit-breakers, protection relay, etc. can be interfaced to the main EMTP-based system model via TACS. Control systems are represented in TACS by block diagrams with interconnection between system elements. Control elements can be transfer functions, FORTRAN algebraic functions, logical expressions and emulation of special devices. Essentially, when employing TACS, the network solution is first advanced, network variables are next passed to the control section and then the control equations are solved. Finally, the network receives control commands. The whole procedure introduces a time-step delay as illustrated in *Figure 36.5*.

36.4.3 Digital real-time electromagnetic transient simulator

One of the limitations with digital computer based simulator programmes, such as the EMTP, is that a single run can

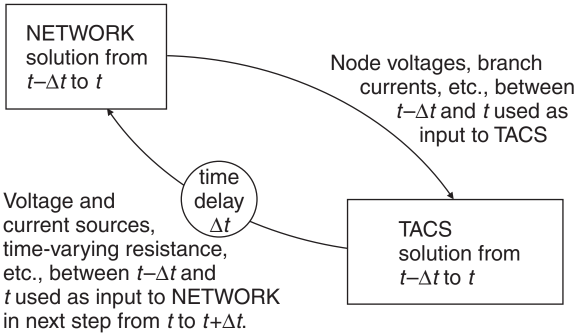


Figure 36.5 Interface between a network and a control system

take many minutes (or even hours) to complete the modelled system's response over a period of, say one second. Such non real-time operation precludes interfacing the simulator to physical control or protection devices. Generally, testing of such devices is done using an analogue simulator (such as a TNA) or a special device which plays back, in real time, the stored results of an off-line computer simulation to the device under test.

As mentioned before, one of the drawbacks of the analogue simulator is its high capital and operating costs. Furthermore, general accessibility is low since a single study may occupy the analogue simulator for many weeks. With the playback device, major drawbacks include the long time required to obtain the simulated results coupled with the fact that only open loop testing is possible i.e. it is not possible to study the effect of circuit changes (such as those due to circuit breaker operations) within a network on the devices under test.

The principal advantages of a fully digital simulator capable of real-time operation are that it incorporates the best features of both the computer-based digital simulation tool and the analogue simulator.

36.5 Waveforms associated with the electromagnetic transient phenomena

36.5.1 Introduction

As mentioned before, the electromagnetic transient phenomena in power system networks can have a wide range of frequencies varying from d.c. to several MHz. Modelling of power components taking into account the frequency-dependence of parameters can be practically made by developing mathematical models which are accurate enough for a specific range of frequencies. Each range of frequencies usually corresponds to some particular transient phenomena. A widely accepted classification of frequency ranges is in the following four groups (with some overlapping):

- low-frequency oscillations, 0.1 Hz–3 kHz;
- slow-fronted surges, 50/60 Hz–20 kHz;
- fast-fronted surges, 10 kHz–3 MHz; and
- very fast-fronted surges, 100 kHz–50 MHz.

This section illustrates some typical transient waveforms associated with the aforementioned categories of frequencies.

36.5.2 Transient waveforms

As mentioned before, a common cause of low-frequency electromagnetic transients is resonance:

(i) Sub-synchronous resonance (typically about 20 Hz) due to the interaction of an externally inserted series capacitor into a transmission line and the natural inductance of the line; such phenomenon can have a detrimental effect on the rotating shaft of a generator, bringing about its premature failure;
(ii) Ferro-resonance (typically about 16 Hz) due to the interaction of the saturable magnetising inductance of a transformer and a capacitive distributive cable or transmission line connected to the transformer. An abnormally large exciting current is drawn resulting in a significantly distorted transformer voltage output.

Slow fronted surges commonly emanate from switching of circuit breakers and shunt faults. The results involving such studies are useful for:

(i) evaluating the performance (or new design) of protection/control devices;
(ii) insulation co-ordination to determine overvoltage stresses on equipment;
(iii) determining the surge arrester characteristics;
(iv) determining the rupturing capacity of circuit breakers and/or the transient recovery voltage across circuit breakers; and
(v) assessing the effectiveness of transient mitigating devices such as pre-insertion resistors, inductors, etc.

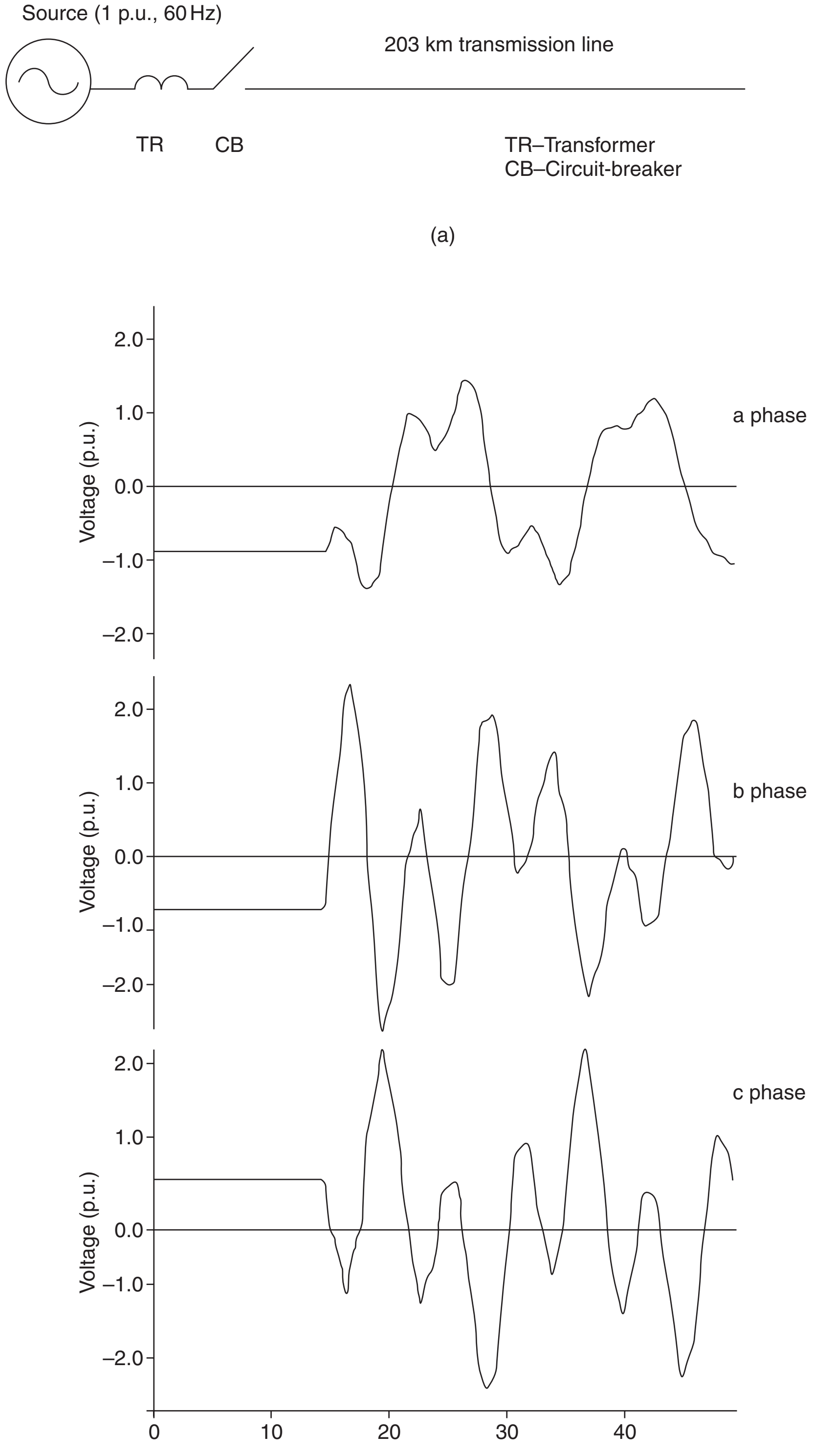
Figure 36.6(*b*) typifies the three voltage waveforms, simulated at the open end of a high voltage transmission system shown in *Figure 36.6*(*a*), when the line is energised by closing the circuit breaker. It is assumed that prior to circuit breaker reclosure, the line carries a trapped charge of -0.9, -0.8 and 0.8 per unit on the a, b and c phases, respectively. It is apparent that the transient overvoltages produced are significantly higher than the nominal source voltage.

The electromagnetic transient phenomena associated with faults on overhead transmission lines is largely dictated by a number of factors, the principal of which are: type of fault, fault location, fault inception angle, bus-bar terminations in terms of source capacities.[7] *Figure 36.8* typifies the voltage and current waveforms when an 'a'-phase-ground fault occurs near voltage maximum of the 'a'-phase, and at the mid-point of a 400 kV transmission system shown in *Figure 36.7* (the waveforms are simulated at end S). Comparing *Figure 36.8*(*a*) and *Figure 36.8*(*b*), it is apparent that the magnitude of the high-frequency distortion is very significantly mitigated by the relatively large source capacity termination at end S.

Figure 36.9 interestingly shows that when an 'a'-phase-ground fault occurs near minimum of the 'a'-phase, there is little high frequency distortion in the voltage waveforms; importantly, there is a large d.c. offset in the current waveforms.

The electromagnetic transient phenomenon of much interest is that associated with the sequence of events that take place on high voltage transmission systems employing three-phase tripping. The voltage and current waveforms simulated at end S of the transmission system as shown in *Figure 36.10* very vividly depicts the electromagnetic transients generated on waveforms (voltage waveforms in particular) following a fault (time T_1) and their characteristic features on circuit breaker opening (T_2)/fault break off (T_3). Of particular importance is the large d.c. levels on the previously healthy phases 'b' and 'c', and these are as a direct consequence of the line trapped charge; in practice, care needs to be exercised when reclosing the breakers as such d.c. levels can significantly accentuate the overvoltage phenomenon, posing serious problems for such devices as the capacitor voltage transformers (CVTs) employed for protection devices.

A study of the electromagnetic transient phenonemon associated with arcing faults on long distance high-voltage transmission lines, in particular those employing shunt compensation for suppressing secondary arcs in single pole auto-reclosure, is of particular interest to design engineers worldwide. *Figure 36.11* depicts the voltage and current waveforms attained for an 'a'-phase-earth fault on a shunt compensated 500 kV high voltage transmission line, 540 km in length. In the Figures, 't_1', signifies breaker opening time



(a)

(b)

Figure 36.6 Switching transients on a high voltage transmission line: (a) the transmission system studied; (b) the voltage waveforms simulated at the line open end

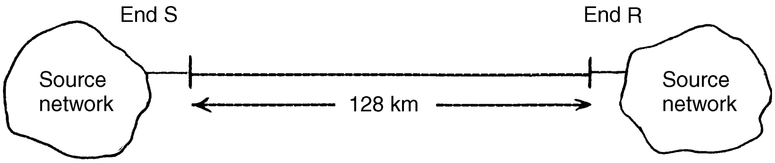
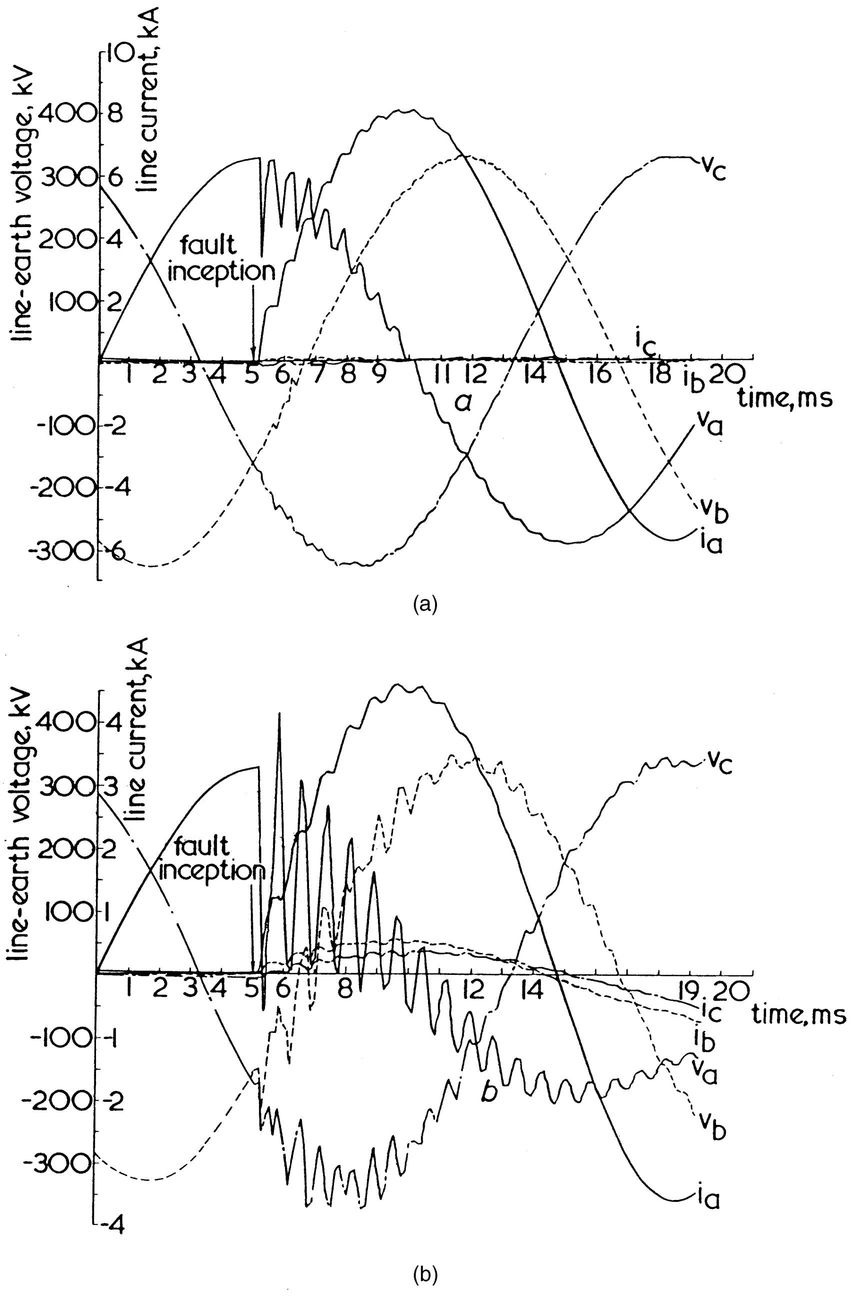

End S
End R
Source network
128 km
Source network

Figure 36.7 A typical transmission system simulated



(a)

(b)

Figure 36.8 Effect of source capacities on voltages and currents: 'a'—earth solid fault at mid-point of line; scl = short circuit level; (a) end S scl = 35 000 MVA, end R scl = 35 000 MVA; (b) end S scl = 5000 MVA, end R scl = 35 000 MVA

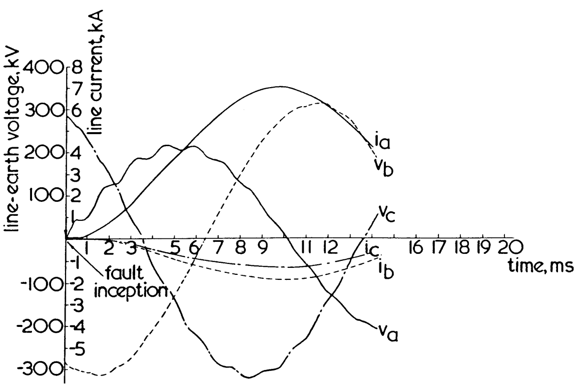


Figure 36.9 Waveforms for fault applied at voltage zero; 'a'—earth solid fault at receiving end; end S scl = 5000 MVA, end R scl = 35 000 MVA

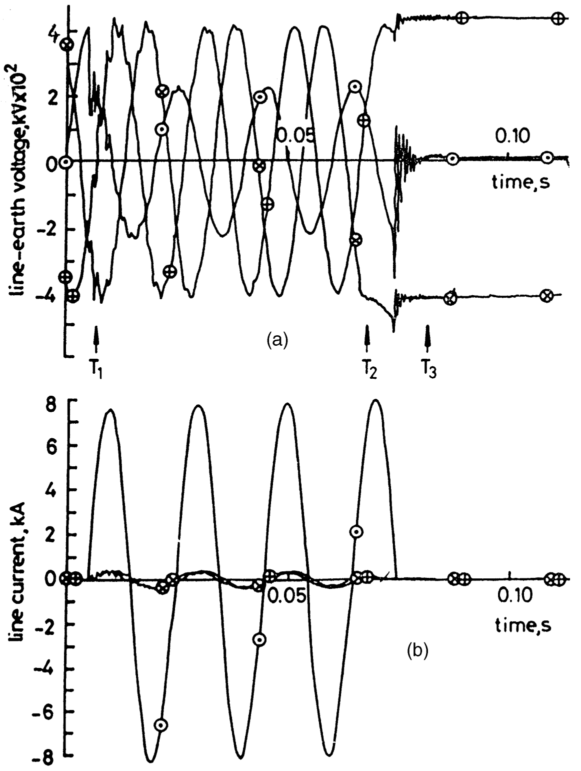


Figure 36.10 Three-phase tripping transient fault simulation; 'a'—earth mid-point fault; T_1 = fault inception (5 ms); T_2 = breaker contacts at ends S and R separate (65 ms); T_3 = fault breaker off (~80 ms) ⊙ a—phase; ⊕ b-phase; ⊗ c-phase; (a) voltages on line side of breaker at end *S*; (b) line currents at end *S*

on the faulted 'a'-phase conductor and 't_2' the secondary arc extinction time. Apart from the high frequency distortion, interestingly, the voltage waveforms also show the 'slow' beating frequency (typically 5 Hz) due to the interaction of the external shunt reactors with the natural line parameters.

The waveforms presented here are only a small sample of those encountered in practice. A more comprehensive knowledge of the electromagnetic transient phenomena on high voltage transmission lines can be attained from references.[7,10–12]

As mentioned before, the fast-fronted transients in power systems cover a frequency range from 10 kHz up to about 1 MHz. One of the principal causes of such transients is the lightning strikes to the transmission lines and associated backflashovers. It is important to note that although the foregoing waveforms (*Figures 36.8–36.11*) are also associated with faults on transmission lines caused by lightning strikes, the transients considered are at the lower-end of the frequency spectrum (<5 kHz). The transient waveforms presented henceforth typify those encountered inside a substation and are at the higher end of the frequency spectrum. One of the important objectives of performing power system studies involving lightning surge phenomena is to design substations, principally to:

(i) establish/verify surge arrester ratings;
(ii) find optimum location for surge arresters for lightning surge protection;
(iii) determine minimum phase-to-ground and phase-to-phase clearances;
(iv) calculate mean time between failure for the substation; and
(v) determine optimum location of capacitances to reduce steepness of surges.

Figure 36.12 typifies the steep-fronted voltage waveforms at varous equipment locations inside a 230 kV substation with 500 kV insulation breakdown level (BIL) when there is a $3/75\,\mu s$ lightning strike of 110 kA striking a tower approximately 150 m from the substation. It is apparent from the transient waveforms that in spite of the very large overvoltages encountered, the surges are well below the equipment BIL of the substation and this can be directly attributed to the presence of surge arresters placed at strategic points within the substation. It should be mentioned that the transients presented herein are for one case study only. However, there are situations whereby the BIL rating of the substation could be exceeded and hence it is crucial to perform a comprehensive set of studies in order to ascertain the type/rating of arresters and their optimal locations.

Very fast transient (VFT) overvoltages are commonly encountered in a gas insulated substation (GIS), and they are primarily generated during disconnector operations, but other events such as the operation of a circuit-breaker, the closing of a grounding switch or the occurrence of a fault can also cause VFTs. During a disconnector operation, a number of prestrikes or restrikes occur due to the relatively slow speed of the moving contact.

The generation and propagation of VFT from their original location throughout a GIS can produce internal and external overvoltages. The main concerns are internal overvoltages between the centre conductor and the enclosure. However, external VFT can be dangerous for secondary and adjacent equipment. These external transients include transient voltages between the enclosure and ground at GIS-air interfaces, voltages across insulating spacers in the vicinity of GIS current transformers, particularly when they do not have a metallic screen on the outside surface, voltages on the secondary terminals of GIS instrument transformers, radiated electromagnetic fields which can be dangerous to adjacent control or relay equipment.

An example of the foregoing transient phenomenon in an actual GIS is given in *Figure 36.13*, where one pre-strike of

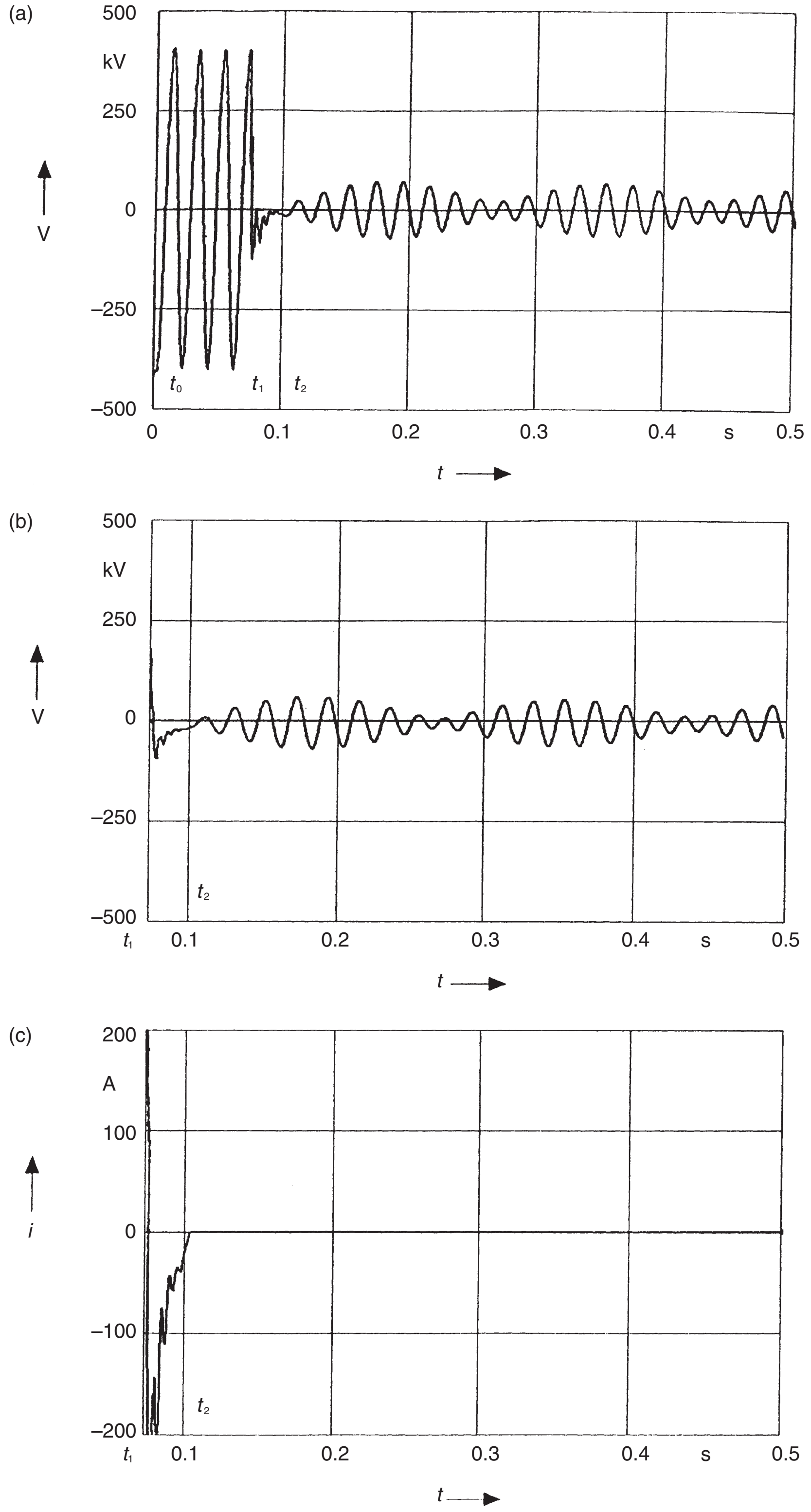


Figure 36.11 Voltage and current characteristics for a long distance practical system, phase-to-ground fault 400 km from the local end: (a) sending end voltage ($t_0 = 0$, $t_1 = 0.078$ s, $t_2 = 0.101$ s); (b) secondary arc voltage; (c) secondary arc current

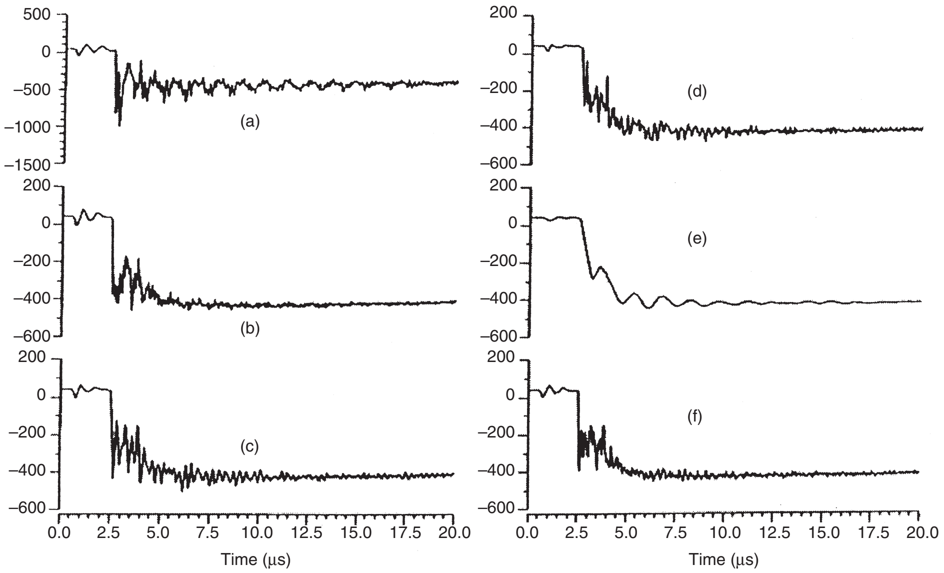


Figure 36.12 Phase-ground voltage waveforms inside the substation: (a) dead-end tower; (b) station entrance; (c) circuit-breaker; (d) remote switch; (e) operating bus; (f) inspection bus. All voltages in kV

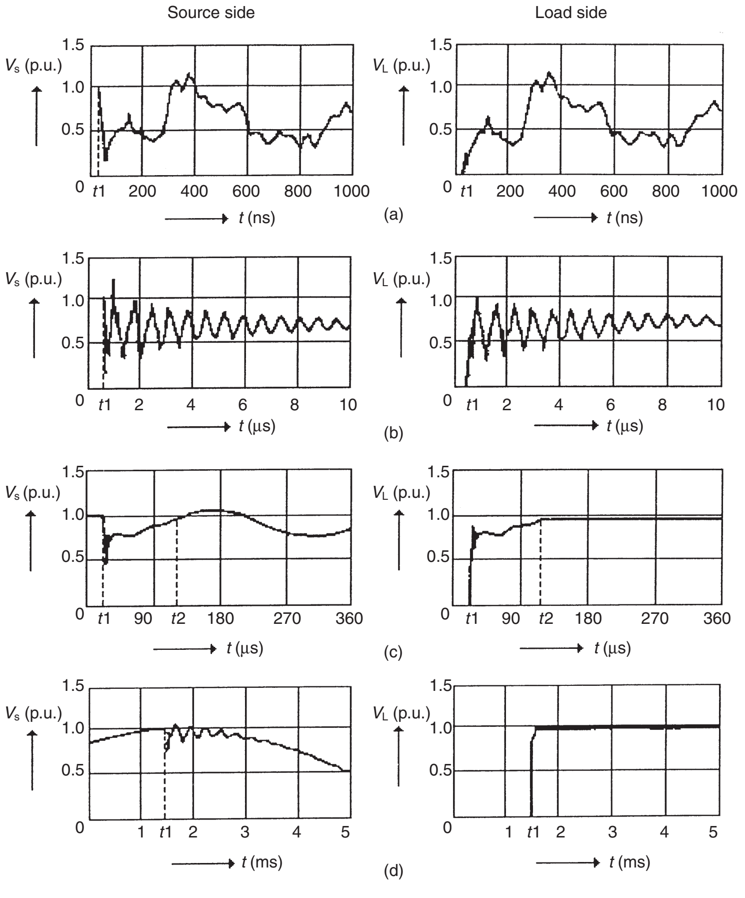


Figure 36.13 Transients on the source and load sides of a GIS due to disconnector switching: (a) steep voltage transients; (b) basic frequency components of the VFT in the MHz range; (c) overall transients in the KHz range; (d) low frequency transient and steady state condition

a disconnector switching is depicted showing the steep fronted voltage transients at the supply and load sides. The basic frequency component is in the MHz range; the overall transient and the steady state waveforms are also shown.

References

1 GRAINGER, J. J. and STEVENSON, W. D., 'Power System Analysis', text book published by McGraw-Hill International, ISBN 0-07-113338-0 (1994)

2 BARTHOLD, L. O. and CARTER, G. K., 'Digital travelling-wave solutions. 1-Single-phase equivalents', *AIEE Trans*. Vol. 80, Pt 3, pp. 812–820 (December 1967)

3 FREG, W. and ALTHAMMER, P., 'The calculation of transients on lines by means of a digital computer', *Brown Boveri Rev*., Vol. 48, 334–355 (May/June 1961)

4 DOMMEL, H. W., 'Digital computer solution of electromagnetic transients in single and multi-phase networks', *IEEE Trans. on Power Apparatus and Systems*, Vol. 88, No. 2, pp. 734–741 (April 1969)

5 MEYER, W. C. and DOMMEL, H. W., 'Numerical modelling of frequency dependent transmission line parameters in an electromagnetic transients program', *IEEE Trans. on Power Apparatus and Systems*, Vol. 93, No. 5, pp. 1401–1409 (September 1974)

6 MARTIN, J. R., 'Accurate modelling of frequency dependent transmission lines in electromagnetic transient simulations', *IEEE Trans. on Power Apparatus and Systems*, Vol. 101, No. 1, pp. 147–155 (January 1982)

7 JOHNS, A. T. and AGGARWAL R. K. 'Digital simulation of faulted ehv transmission lines with particular reference to very high speed protection', *Proc. IEE*, Vol. 123, No. 4, pp. 353–359 (April 1976)

8 DOMMEL, H. W., 'Non-linear and time-varying elements in digital simulation of electromagnetic transients', *IEEE Trans. on Power Apparatus and Systems* Vol. 90, No. 6, pp. 2561–2667 (November 1971)

9 DUBE, L. and DOMMEL, H. W., 'Simulation of control systems in an electromagnetic transients program TACS', *Proc. of IEEE PICA*, pp. 266–271 (1977)

10 JOHNS, A. T. and AGGARWAL, R. K., 'Digital simulation of faulted autoreclosure sequences with particular reference to the performance evaluation of protection for ehv transmission lines', *Proc. IEE*, Vol. 128, Pt C, No. 4, pp. 183–195 (July 1981)

11 AGGARWAL, R. K., JOHNS, A. T. and KALAM, A., 'Computer modelling of series compensated ehv transmission systems', *Proc. IEE*, Vol. 131, Pt C5 (September 1984)

12 SONG, Y. H., AGGARWAL, R. K. and JOHNS, A. T., 'Digital simulation of fault arcs on long-distance compensated transmission systems with particular reference to adaptive autoreclosure', *European Trans. on Electrical Power Engineering*, Vol. 5, No. 5, pp. 315–324 (September/October 1995)

37

Optical Fibres in Power Systems

R Tricker MSc, IEng, FIEE(elec), FInstM, FIQA, MIRSE

Contents

37.1 Introduction 37/3

37.2 Optical fibre fundamentals 37/3
  37.2.1 Optical propagation in fibres: ray theory 37/3
  37.2.2 Acceptance angle and numerical aperture 37/4
  37.2.3 Basic fibre types, modes, mode conversion and bandwidth 37/4
  37.2.4 Fibre protection 37/7
  37.2.5 Fibre strength 37/7

37.3 Optical fibre cables 37/8
  37.3.1 General 37/8
  37.3.2 Low fibre count cables 37/8
  37.3.3 High fibre count cables 37/8
  37.3.4 Cable protection 37/9
  37.3.5 Cable usage 37/10
  37.3.6 Splices and connectors 37/10

37.4 British and International Standards 37/12

37.5 Optical fibre telemetry on overhead power lines 37/13
  37.5.1 Introduction 37/13
  37.5.2 Fibre cable configurations 37/14
  37.5.3 Factors governing system design 37/15

37.6 Power equipment monitoring with optical fibre sensors 37/15
  37.6.1 Introduction 37/15
  37.6.2 Technology implementation difficulties 37/15

37.1 Introduction

In the 1980s when optical fibres were in their infancy, they were primarily used by the telecommunications industry to replace the copper lines normally used in telephone networks and the long-haul coaxial trunk links between telephone exchanges. Nowadays, this cost-effective medium with its high transmission bandwidth, low attenuation and relative immunity to ElectroMagnetic Interference (EMI) has become increasingly advantageous in the fields of data communications, Local Area Networks (LANs), Urban Broadband Service Networks (UBSNs), Community Antenna TeleVision (CATV), and control applications. Fibre (especially in digital communications) overcomes the problems of different earth potentials between locations, is unaffected by the electric fields found in high-voltage environments, lightning strikes and other electromagnetic interference (e.g. the high magnetic fields caused by transformers and motors etc.). As the data transfer rates of local area networks increase, fibre becomes increasingly attractive and more economically viable. In terms of pence per metre per unit bandwidth, in an increasing number of applications, fibre as a transmission medium is the obvious choice.

In the electrical power industry, these advantages have meant that fibre optics and the technology of fibre optics has replaced normal cable systems and is now widely used for a variety of purposes. The adoption of fibre optics by the telecommunications technology is already well advanced and the realisation of reliable optical data transmission for protection and control has been convincingly demonstrated. Furthermore, research into optical fibre based parameter sensing has reached the stage that properly engineered systems are now available and their integration into optical fibre networks has produced a purely optical monitoring and transmission system.

This chapter of the *Electrical Engineers' Reference Book* contains a brief introduction to optical propagation in fibres followed by a description of the technology of cabling and interconnecting fibres. The present status of communication and telemetry along high voltage transmission lines is discussed and progress towards the optical fibre monitoring of power equipment in substations and elsewhere on a power system is described.

37.2 Optical fibre fundamentals

37.2.1 Optical propagation in fibres: ray theory

Strictly speaking, one should not use the term 'light' when referring to most optical fibre transmission systems. Normally, near infra-red radiation (with a wavelength in the range 780–1550 nm), derived from a light-emitting diode (LED) or semiconductor laser, is employed and as such, the radiation is not usually visible to the naked eye. However, 'light' is the term commonly employed, and is used here.

Light propagation through an optical fibre depends upon total internal reflection at the interface between two transparent materials with high and low refractive indices. *Figure 37.1* shows that as a ray approaches a boundary within a transparent medium, it can be totally internally reflected at the high–low refractive index interface, and will be guided along the high refractive index medium. As the angle at which the approaching ray increases, a critical value (θ_2) may be reached beyond which light 'leaks' out of both media.

The most basic type of optical fibre can be developed from the above principle by using a cylindrical geometry. Rays beyond the critical angle are trapped within the fibre core and travel down the fibre.

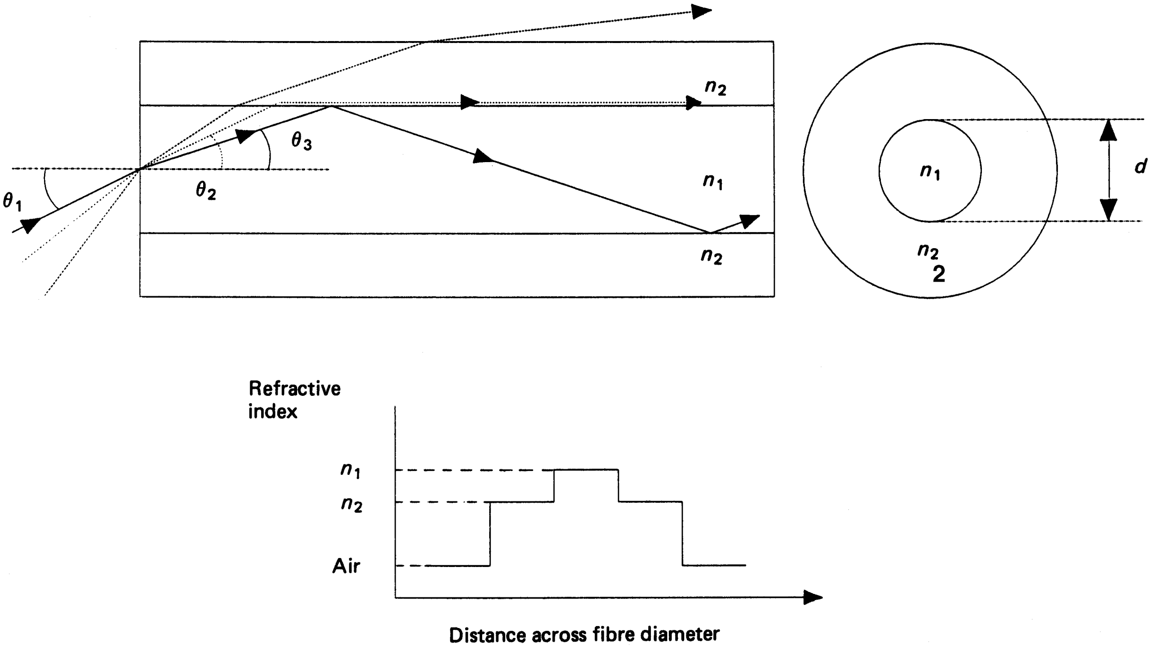


Figure 37.1 Ray diagram for multimode step index fibre

37.2.2 Acceptance angle and numerical aperture

The acceptance angle is a function of the refractive indices of the core and the cladding materials. The sine of the acceptance angle is called the numerical aperture.

Consider the fibre illustrated in *Figure 37.1*, having a circular core of diameter *d*, and a uniform refractive index n_1, surrounded by a cladding layer of uniform refractive index n_2. Light launched into the core at angles θ_1 will be propagated within the core at angles θ_3 up to a maximum value θ_2 to the axis. Light at angles greater than θ_2 will not be internally reflected and will be refracted into the cladding. The maximum launch or acceptance angle θ_1 can be expressed as a function of the Numerical Aperture (NA), where:

$$\mathrm{NA} = (n_1^2 - n_2^2)^{1/2} = \sin\theta_1 = n_1 \sin\theta_2 \qquad (37.1)\Leftarrow$$

Note that reciprocity dictates that what is true for light entering the core of a fibre is also true for light exiting. The fibre core diameter (d), the NA, and the operating wavelength (λ) are often used together in a single parameter, known as the normalised frequency, waveguide parameter, or fibre parameter (V), which is of importance in characterising all fibres:

$$V = \Leftarrow \frac{\pi d}{\lambda \varsigma} .\varsigma \mathrm{NA} \qquad (37.2)\Leftarrow$$

37.2.3 Basic fibre types, modes, mode conversion and bandwidth

Two types of fibre are used for optical transmission (step-index and graded-index) and two modes (single-mode and multimode).

37.2.3.1 Single-mode fibre

Single-mode fibre (also referred to as fundamental or mono-mode fibre) will permit only one mode to propagate and, as such, cannot suffer mode delay differences. Single-mode fibres are capable of wide bandwidths (e.g. >40 GHz) and are, therefore, ideally suited for long-haul and high capacity circuits. Single-mode fibre are used almost universally in telecommunications over 1 km or so and are generally used at the 1300 nm and 1550 nm wavelengths where attenuation is low and sources and detectors are available.

37.2.3.2 Multimode fibre

In comparison to single-mode fibre, multimode fibre has a relatively large core (typically 50–60 µm) and a high numerical aperture. As the name implies, multimode fibres are capable of propagating more than one mode at a time and they are ideally sited for high bandwidth (i.e. a few GHz) and medium haul applications. Multimode fibre has become increasingly popular in the last decade and is now widely used in high-speed (e.g. 100 Mbit/s) local area networks, and 62.5/125 (core/cladding diameters in µm) fibre was specified as the first choice for the ANSI X3T9.5 committee Fibre Distributed Data Interface (FDDI) standard.

37.2.3.3 Step index fibre

A small glass fibre in air has a large critical angle of approximately 42°.⇐This enables propagation along paths that are much longer than the axial route but attenuation and dispersion losses are substantially increased. By making the fibre core extremely narrow (e.g. 150 µm) and enclosing it in a cladding material whose refractive index is only slightly lower than the fibre, will not only reduce the critical angle but also reduce the losses.

Waveguide theory shows that the total number of modes which can be sustained in a step index fibre is given by

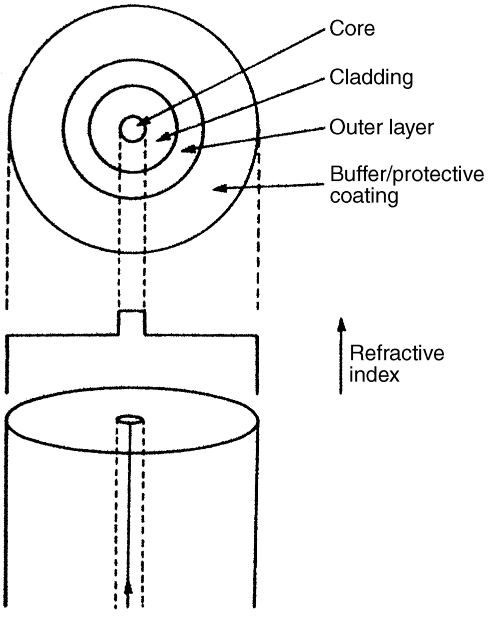


Figure 37.2 Single-mode optical fibre, together with its refractive index profile and cross-section

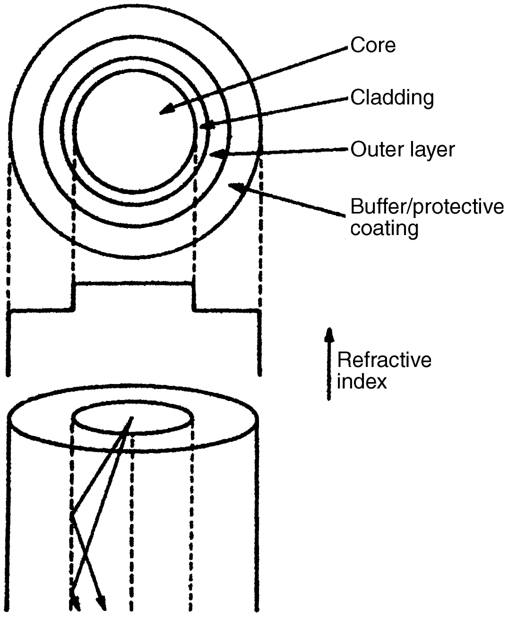
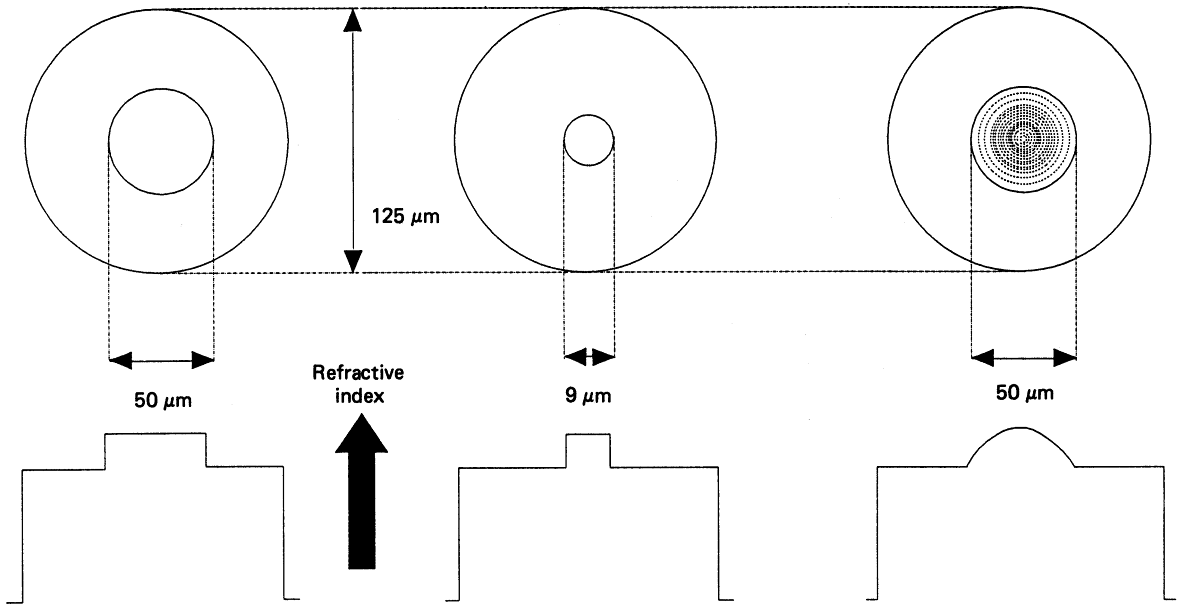


Figure 37.3 Step-index multimode fibre, together with its refractive index profile and cross-section



(a) Step index multimode **(b) Step index single mode** **(c) Graded index multimode**

Figure 37.4 Refractive index profiles (typical dimensions shown)

$N = V^2/2$. Modes can be visualised as rays propagating at differing angles to the fibre core. Discrete and definable modes only can propagate because of the geometrical constraints of the fibre, and are analogous to modes in hollow metallic waveguides used at microwave frequencies (ca. 1–100 GHz). Typical multimode fibres, with core diameters of 50–200 μm propagate 100–1000 modes. It is not necessary for the general user of fibres, however, to understand the mathematics of propagation.

In a multimode step index fibre (*Figure 37.4*(*a*)), no matter however careful one is to launch a single mode, conversion between modes or ray angles is inevitable because of bending and fibre imperfections. This is a great drawback for such fibres, because different modes travel at different speeds down the fibre, causing different arrival times at the receive end. The difference in transit time between the extreme ray paths for a multimode step index fibre of length L is given by:

$$\Delta T_{\text{intermodal}} = \frac{L}{c}(n_1 - n_2) \qquad (37.3)\Leftarrow$$

where c is the velocity of light. The difference in transit times causes a sharp pulse of launched light to become spread at the distant end, limiting the bandwidth of a system. Typically, for an all-silica based fibre of $\text{NA} \approx 0.2$, pulse spreading is of the order of 50 ns/km, and is inversely proportional to the length of the system; the longer the system, the lower the bandwidth.

As V is reduced, less guided rays or modes can be supported, and when $V < 2.405$, only a single waveguide mode can propagate. Such single mode fibres have a core diameter which is comparable with the wavelength of light (d is commonly 8–10 μm for telecommunications fibres see *Figure 37.4*(*b*)), making fibre-fibre and fibre-device interconnection more difficult and generally less efficient than for multimode fibres. Intermodal dispersion does not occur, and bandwidths can be very high.

It should be noted, however, that a fibre which is single mode at, say, 1300 nm will not necessarily be single mode at 850 nm or below. V increases as the wavelength decreases, and will generally be greater than 2.405 at 850 nm. A wavelength known as the cut-off wavelength is an important manufacturing parameter defining the onset of multimode behaviour.

Simple ray optics cannot describe the propagation of energy through a single mode fibre very well as it is difficult to depict a single ray being guided. Mathematical modelling of single and multimode fibre may be achieved by solving Maxwell's equations with boundary conditions defined by fibre geometry and wavelength. This involves Bessel functions and is beyond the scope of this introduction. One of the important results of mathematical modelling, however, is that optical power is not confined to the core alone, but extends appreciably into the cladding region—the power distribution is approximately Gaussian. The extent of cladding penetration is dependent primarily on refractive index difference and wavelength.

37.2.3.4 Graded index fibre

In an effort to overcome bandwidth and connection difficulties, a fibre type (known as graded index fibre) was developed as an intermediate step between index multimode and step index single mode Although previously only used for long-haul (i.e. trunk) networks, graded index fibres are nowadays frequently being employed in local networks at the 850 and 1300 nm wavelengths.

The principle behind graded index fibre is shown in *Figure 37.4*(*c*). Here, the refractive index profile is graded, and ray paths are curved as the rays are continually refracted (*Figure 37.5*). Rays which travel closer to the core-cladding boundary are in a region of lower refractive index, and will, therefore, travel faster than those in the denser central core area. The overall effect, given the appropriate refractive index profile, is that rays travelling

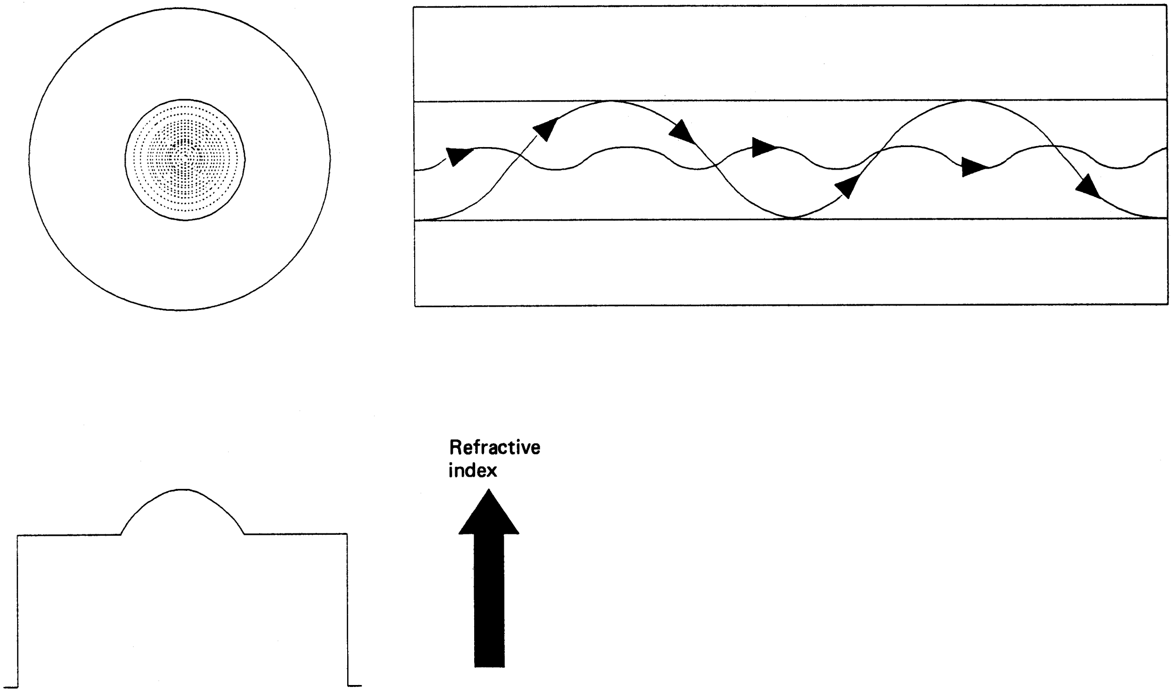


Figure 37.5 Typical ray paths for graded index fibre

different paths arrive at the far end at approximately the same time. The exact index profile to minimise dispersion effects is dependent on the composition of the fibre and the operating wavelength, but is approximately parabolic.

37.2.3.5 Material dispersion

As the refractive index of a glass prism varies with wavelength, different wavelengths (from an LED and even a narrower linewidth laser source) travelling at different velocities down the fibre will arrive at different times. This causes an energy loss which is called material dispersion, and which predominates in single-mode fibres. The speed of light propagating in the fibre is inversely proportional to the refractive index of the propagating medium, and the refractive index of silica drops from 1.46 to 1.44 between 600 and 200 nm, approximately. The variation of the material dispersion parameter $M(\lambda)$ with respect to wavelength is given by

$$M(\lambda) = -\frac{\lambda d^2 n_1}{c\, d\lambda^2}\ \text{(ps/nm/km)} \qquad (37.4)\Leftarrow$$

and is shown for silica in *Figure 37.6*. Pulse broadening in a particular case can be calculated by multiplying the value of $M(\lambda)$ by both the length of fibre in question and the linewidth of the source in nanometres. An LED source, for example, operating at 850 nm, and with a linewidth of 40 nm will give pulse spreading of some 4 ns/km, and this spreading may be significantly reduced by using a laser source with much reduced linewidth of typically 4 nm or less. The bandwidth length product is generally specified for multimode telecommunications and data communications grade fibres, and is generally of the order of a few hundred megaherz-kilometre product.

Note that $M(\lambda)$ goes through zero at approximately 1300 nm. First generation telecommunication systems operated with multimode fibre in the region of 850 nm, called the first window (where many sources and detectors were available, and fibre losses were acceptable at approximately 3 dB/km or greater), but second generation (or second window) systems operate at 1300 nm and can exploit the dispersion zero to give high bandwidths. At 1300 nm, losses can also be substantially below 1 dB/km, giving far greater transmission distances before regeneration is required. A third window at 1550 nm, at which attenuation can be less than 0.2 dB/km, is now widely used throughout the telecommunications industry. *Figure 37.7* shows nominal attenuation

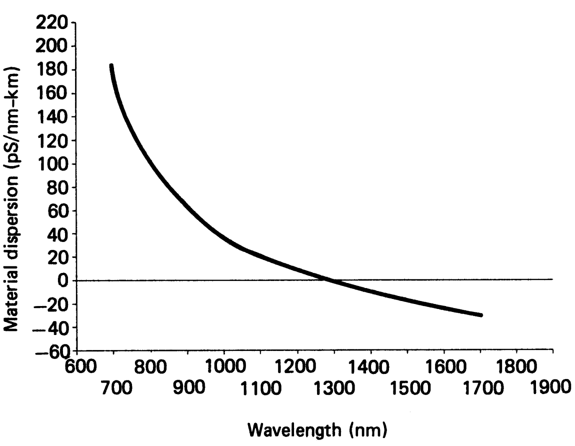


Figure 37.6 Material dispersion versus wavelength for silica

Table 37.1 Fibre types

Type	*Core/cladding diameter* (μm)	*Typical attenuation* (dB/km)	*Typical bandwidth* (MHz/km)	*Applications*
All silica				
Step index multimode	50/125–200/300	3–10 at 850 nm	20	Data links
Graded index multimode	50/125–100/140	3 at 850 nm <1 at 1300 nm	200–1000	Telecommunications, data links
Single mode	5/125–10/125	<0.5 at 1300 nm <0.25 at 1550 nm	>1000	Telecommunications, high speed data links
Other				
PCS	50/125–200/300	5–50 at 850 nm	20	Data links
All plastic	50/100–50/1000	>100	<20	Light pipes, electrical isolation, short data links

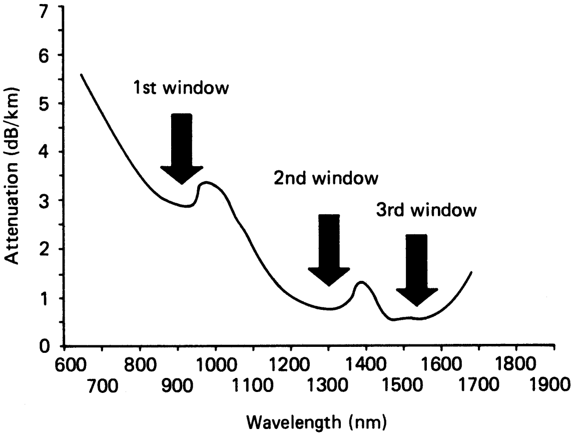


Figure 37.7 Nominal attenuation versus wavelength for silica fibre (showing operating windows)

versus wavelength for all-silica fibre, and indicates operating windows.

By modifying the chemical composition of single-mode fibre, and the geometry of the core and cladding, the zero of $M(\lambda)$ can be moved to 1550 nm (the third window), which gives further advantages in terms of attenuation, as is explained in the next section. $M(\lambda)$ can also be flattened to give near zero dispersion at both 1300 and 1550 nm, but this is achieved only at the expense of an attenuation penalty and is not common practice. (High-bandwidth systems are generally achieved by using very narrow spectral linewidth lasers.)

The above treatise generally refers to all-silica fibres which are widely used in the telecommunications industry. The two other common types are Plastic-Clad-Silica (PCS) and all-plastic fibres.

PCS fibres have an all-silica core and a polymer based cladding (commonly a silicone resin) which also serves as a protective layer. They are generally less expensive to manufacture, but are characterised by higher attenuation and lower bandwidth (as they have a step index) than all-silica fibres, but are used for relatively short data links.

All-plastic fibres, generally manufactured from Poly-Methyl-MethAcrylate (PMMA) are the least expensive type. They currently have the highest attenuation of commonly available fibres, and are generally step index. They have application as 'light pipes' over short distances (i.e. a few metres).

37.2.4 Fibre protection

37.2.4.1 Microbending losses

If a fibre is subjected to mechanical stress, local discontinuities can be introduced. Curvatures of the fibre involving axial displacements of a few millimetres may cause a light ray on the outside radius of a bend to approach or exceed the optical angle and light may be lost through the cladding and also result in additional attenuation losses. These are referred to as microbending losses, or in the case of macroscopic axial deviation of the fibre from a straight line, macrobending losses. To overcome these problems, high-density laser diodes are frequently used.

A guide to the susceptibility to microbending of a particular multimode fibre type can be made by using the following 'figure of merit' (this is based on step-index analysis but can be used as a general guide):

$$\gamma\varsigma = \lessdot K \frac{d^4_{\text{core}}}{\text{NA}^6 d^6_{\text{cladding}}} \qquad (37.5)\Leftarrow$$

An optical fibre must therefore be protected against radial forces. This is accomplished by mechanically decoupling the fibre from its immediate surroundings, and is commonly achieved by surrounding the fibre with a very low-elastic modulus material such as a silicone rubber followed by an extruded polymer layer (termed a 'tight' packaged fibre), or by encapsulating the fibre loosely within a polymer tube ('loose' packaging). The type of protection depends upon the particular application.

Note that the pristine surface of an all-silica fibre is always protected during manufacture by a thin UltraViolet (UV) radiation or heat cured polymer layer.

37.2.5 Fibre strength

37.2.5.1 Proof testing

Silica is, to all intents and purposes, purely elastic and has no plastic deformation prior to tensile failure. The strain at which a fibre breaks is dictated by its surface condition.

Minute flaws, 1 µm or less in size, and generally at the surface, act as stress raising points and the effective stress at the flaws may be much higher than that directly imposed. The flaws are generally created during fibre fabrication, and are caused by (inevitable) contamination. Although all manufacturers take immense care to produce fibre in scrupulously clean conditions, some particles (from the hot furnace element, for example) inevitably reach the fibre. A flaw 1 µm deep will cause failure at about 1% strain.

The stress intensity at a crack flaw tip can be explained by classic Griffith crack theory. Local stress intensification can be described by an intensification factor K_1, which reaches a critical value K_{IC} when fracture occurs:

$$\sigma\varsigma = \frac{K_{IC}}{Ya^{1/2}} \tag{37.6}$$

where σ, is the applied stress, a the crack depth, and Y is a geometric factor. $\sigma\varsigma$ is always less than the theoretical breaking stress for the material in question.

Flaws can be introduced easily, and that is one reason why fibres are coated, within seconds of being drawn from the melt, with a UV curing polyurethane acrylate or a thermally curing silicone material to a diameter of 250 µm.

Strength is guaranteed by subjecting all delivered fibre to a proof or screen test, commonly 0.5–1% strain. The strain is applied after production by passing the fibre continuously through two capstans, and loading the section in between.

Fibre manufacturers will generally specify the minimum radius to which a fibre should be bent (about 50 mm for telecommunications fibres). There are two reasons; a bend causes tensile strain on the outer edge, and severe bending will cause light leakage.

37.2.5.2 Static fatigue

Static fatigue is caused by the combined action of tensile stress and moisture on a fibre surface, and causes weakening over time. Because glass is a supercooled liquid, it has an amorphous or non-crystalline structure. The lattice is quite open and water (in the form of OH^- ions), if it reaches the fibre surface, destroys silica–silica surface bonds. The energy for this to happen is supplied by external (tensile) stress, and the rate increases with the applied stress.

Although there has been much research on the phenomenon, it is still not fully understood. The generally accepted model is for a growth law of the form:

$$V = AK_I^n \tag{37.7}$$

where V is the velocity of crack growth, K, is the stress intensification factor mentioned above, and A and n are constants. If the above is combined with the Griffith equation and integrated, the following result is achieved:

$$t = BS_I^{n-2}\sigma\varsigma^{n} \tag{37.8}$$

where t is the time to failure, S_I is the inert strength of the material (the stress required to produce instant failure or failure in the absence of crack growth), B is a constant related to both a and n, and σ, is the applied stress.

The time to failure is thus finite and inversely proportional to the nth power of the applied stress or strain. Values of n usually lie in the range 14–30, the higher the better. The lowest value corresponds to 100% relative humidity and higher values to dry laboratory conditions. The time to failure is thus extremely sensitive to changes in both $\sigma\varsigma$ and n, and accurate values are needed to enable cable lifetime to be reliably extrapolated. During cable design, manufacturers take account of static fatigue; to ensure that design lifetimes are achieved, maximum installation and service loads must be adhered to.

37.3 Optical fibre cables

37.3.1 General

Unlike copper cables, where the performance of conductors is largely immune to bending and tensile forces, optical cables are designed to protect the contained fibres. This does not suggest that properly designed fibre cables are fragile, but merely that care has been taken during both design and manufacture. Optical cables have and are being installed in environments that are as harsh as those experienced by copper cables—for example, sub-sea, between high-voltage electricity pylons, and in the military environment.

Suppliers of optical cables use many different manufacturing techniques and, although it is beyond the scope of this introduction to include them all, the most common types encountered in data communications and telephony are covered.

37.3.2 Low fibre count cables

The requirement for low fibre count cables is generally for a cost-effective, compact construction which provides a package with ample mechanical protection. Typical uses are for inbuilding telephony, data and control links, rack-to-rack links, and electrical isolation.

Both 'tight' and 'loose' designs of cable are used. 'Tight' designs utilise fibre which has been packaged in a closely fitting polymer jacket, commonly nylon, PolyVinyl Chloride (PVC), polypropylene, or polyester elastomer, the outer diameter of which is between 0.5 and 1.0 mm. One or more individually jacketed fibres are generally positioned at or near the neutral bend axis of the cable, with a peripheral tensile strength member. The simplest design is for a single jacketed central fibre provided with stranded polyaramid yarns (e.g. Kavlar and Twaron) for axial reinforcement, and an overall sheath. Similar designs are available for 'loose' packaged fibre with a tube outer diameter of 1.5–3.0 mm, but these may have an increased overall diameter.

'Tight' fibre designs have the advantage that, once broken out of the cable, the fibre still maintains some measure of protection, and may be routed over short distances without a splice. Utilisation of this approach can be cost-effective and will generally keep overall losses to a minimum. *Figure 37.8* shows a typical single-fibre cable design.

37.3.3 High fibre count cables

As the fibre count increases, suppliers generally encounter a cost and overall diameter trade-off between centrally and peripherally positioned fibres. The break point varies between manufacturers and manufacturing methods, but is typically between 8 and 20 fibres.

The approach adopted in most cases is to bundle the fibres together in some form of unit packaging or element. This may take the form of tubes containing one or more loose fibres, placing the fibres in a grooved or slotted former, using fibre ribbons akin to copper ribbon cabling

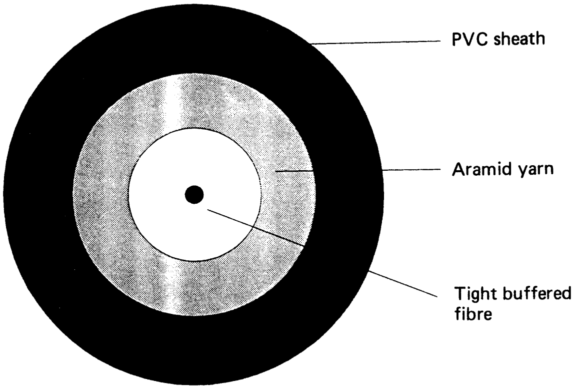

PVC sheath
Aramid yarn
Tight buffered fibre

Figure 37.8 Single-fibre cable

(but with much reduced dimensions), or a combination of the above.

An extruded thermoplastic sheath is then applied to elements stranded helically or in an oscillating fashion (commonly called 'S-Z' stranding) around a central axial strength member. In external cables, interstitial filling compounds and/or an aluminium-polymer laminate barrier are used to prevent moisture reaching the fibres and exacerbating static fatigue effects. *Figures 37.9* and *37.10* show typical high fibre count cable designs.

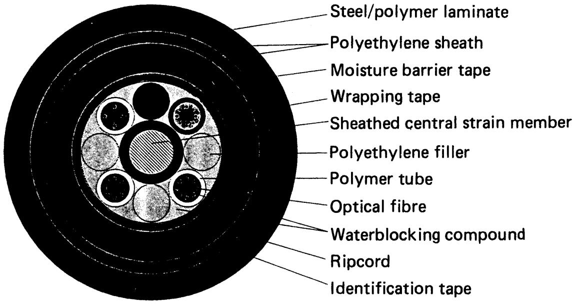

Steel/polymer laminate
Polyethylene sheath
Moisture barrier tape
Wrapping tape
Sheathed central strain member
Polyethylene filler
Polymer tube
Optical fibre
Waterblocking compound
Ripcord
Identification tape

Figure 37.9 Loose tube type cable

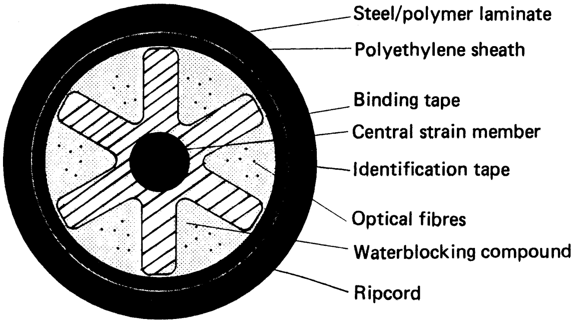

Steel/polymer laminate
Polyethylene sheath
Binding tape
Central strain member
Identification tape
Optical fibres
Waterblocking compound
Ripcord

Figure 37.10 Slotted core type cable

37.3.4 Cable protection

37.3.4.1 Ruggedisation

Ruggedisation of the cable can be achieved by utilising some form of strength member (e.g. Kavlar) which is either laid helically or braided around the fibre coating. This is then surrounded by a tough outer sheath to provide the required environmental and mechanical protection. *Figures 37.11* and *37.12* show two examples of ruggedised optical fibre cables.

37.3.4.2 Armouring

Sometimes even ruggedising an optical cable is insufficient for some environments, especially where the cable is liable to mechanical radial forces (caused by digging implements) and rodents. In these cases the optical cable will need to be armoured. The most usual forms of armour are stranded steel wires, polymer laminated tapes, Kavlar and corrugated steel tape, the latter being particularly effective against rodent bites. The steel must, of course, be protected against corrosion by moisture.

37.3.4.3 Fire safety (internal cables)

As the fire safety properties of a cable are dependent on both the materials and the construction, cables which are to be sited within buildings require constructions which make them safe to use in the event of fire. Polyethylene, used commonly as a sheathing compound for external cables, has poor fire safety properties, and flame retardant PVC sheaths have been and are used for internal (electric) power and data cables. For more demanding applications, for example at sites where the public has common access, it is prudent to use cables which in a fire will:

- resist the spread of flame (flame retardance);
- produce minimal smoke; and
- produce minimal toxic emissions.

37.3.4.4 Environmental considerations

The service environment of a cable will, again, affect both the materials and the construction used. Principal considerations will be temperature, aggressive gaseous and liquid chemicals, radiation, electric fields, and extreme mechanical and other forces such as hydrostatic pressure.

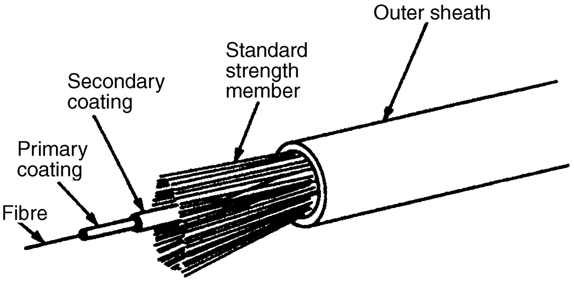

Outer sheath
Standard strength member
Secondary coating
Primary coating
Fibre

Figure 37.11 Construction of a single-fibre ruggedised optical fibre cables

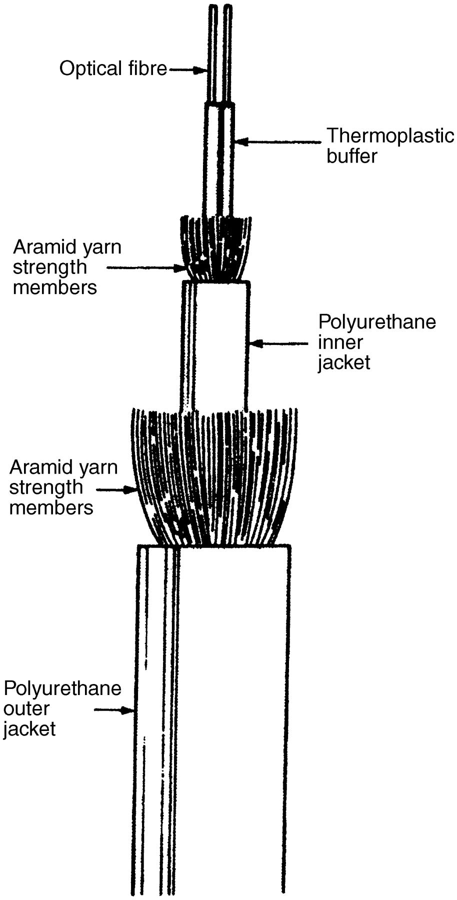

Optical fibre
Thermoplastic buffer
Aramid yarn strength members
Polyurethane inner jacket
Aramid yarn strength members
Polyurethane outer jacket

Figure 37.12 Construction of a multi-fibre ruggedised optical fibre cable, suitable for harsh, military environments

37.3.5 Cable usage

Optical fibre cables, because of their mechanical strength and low weight, can be pulled into ducts, ploughed in, cleated to walls, installed in vertical runs of over 1000 m in length (particularly important in multistorey office blocks and mines etc.) placed on cable trays and planer shelves (or PVC conduit) installed aerially, attached to supporting wires (e.g. high-voltage cables), or even immersed in water (oceanographic systems).

The technique used, of course, not only depends upon the environment but on the length of cable to be installed. Today, one of the most practised methods of installing short lengths into highly populated duct systems is to use compressed air to inject an auxiliary rope. This will then be connected to a winch rope which in turn will be connected to the cable.

37.3.5.1 In buildings

In buildings, cables tend to be of the 'tight' type, for two main reasons. Firstly, they allow for connectors to be directly mounted on to buffered fibre, minimising the number of splices and 'pig-tails' (i.e. factory assembled single-fibre cables with a connector at one end). This is generally a more cost-effective means of installation, and also keeps route losses to a minimum. Secondly, the tight construction relieves axial fibre stress in vertical runs due to its own weight (i.e. the fibre is continually supported by the cable strength member along its entire length).

37.3.5.2 External

In external applications, it is important to keep moisture away from the fibre surface, and to minimise fibre stress. 'Loose' cable types prevail, generally using a combination of filling or flooding compounds designed to displace water, plus an aluminium–plastic–laminate moisture barrier. Axial strength is provided by a stranded steel or glass reinforced member, and/or polyaramid yarns.

Because optical fibre cables are very light they can be incorporated into overhead power line routes without any additional stress being exerted onto the existing mast. In addition, as the optical fibre cable is completely non-metallic, any problems associated with inductive interference are eliminated.

The process is very similar to installing normal copper aerial cables and the optical fibre is merely lashed to the phase conductors or ground wires of the overhead power line. This is further explained in Section 37.5.

37.3.6 Splices and connectors

When an optical fibre cable is damaged it can be repaired either by splicing a new piece of cable into the existing line or, if sufficient cable is available, by cutting out the damaged portion and splicing the two ends together. This process is referred to as concatenation.

Most techniques for joining fibres rely on accurate geometry. The core/cladding concentricity is also very important in all-silica fibres where alignment of the cores usually depends on alignment of the outside surfaces of the cladding. In any optical system, there is likely to be a mix of splices and connectors for active device to fibre interconnection and for fibre to fibre interconnection (e.g. external cable to in-building cable, or for patching).

Both splices and connectors require the achievement of fibre end faces which are flat and at 90° to the fibre axis. Two methods (or a combination of both) are commonly used; cleaving, and cut-and-polish. Tools for either method are commonly available.

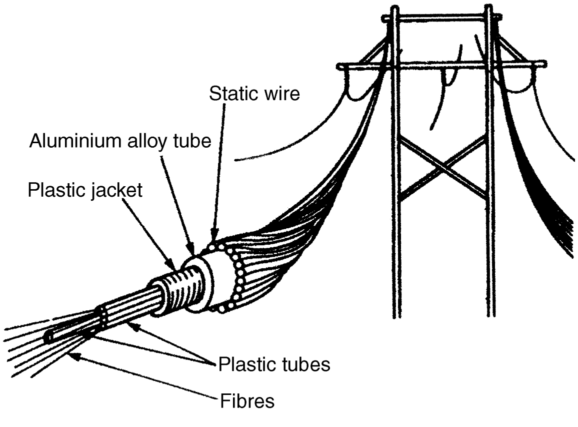


Figure 37.13 Ground wire, incorporating optical fibres, mounted on the top of a pylon power line structure

37.3.6.1 Splices

A number of techniques for permanently and temporarily splicing optical fibres are available, including fusion, V-groove (now largely obsolete), and sleeve splices. Sleeve splices have obvious mechanical advantages for single-fibre jointing in the field, but have the disadvantage of 'slop', due to the clearance required to insert the fibres in the sleeve. Fusion splices are attractive but they require precise external alignment of the fibres to be joined. The fusion type is permanent, but some sleeve splices can be demounted and re-used.

In a typical sleeve splice, an elastomeric tube is used to locate the fibre ends which are simply inserted into the sleeve, together with a small amount of index adhesive to fuse the two fibre ends together. Alternatively, the splice connector can be crimped to the optical fibre cable. Supports are available to house the sleeve and to clamp the fibre/cable. An insertion loss of 0.2 dB is typical for both single and multimode fibres, and the elastomeric tube can be re-used up to 50 times.

Fusion splicing for silica fibres has been developed by numerous companies world-wide. In this technique, the fibres to be joined are visually or automatically aligned using micromanipulator movements. The aligned fibres are then fused together using a suitable localised heat source. Electric arcs are most common, but for high-strength splices, an oxy-gas torch is sometimes used. Losses less than 0.1–0.2 dB are routinely and reliably achieved for both multi- and single-mode fibres. Alignment is commonly achieved by locally injecting light in one fibre and detecting in the other, using a severe bend to cause core/cladding mode coupling. Fusion is activated when the locally detected signal is maximised.

Fibres are usually protected from the environment after splicing by some type of reinforced heat shrink tube, or by coating reinstatement.

Splice losses Splice losses can be due to variations in the outer diameter of the fibre core, differences in index profile, differences in the ellipticity of the core, misalignment of the fibre ends, poor quality of the refractive index match at fibre ends, waveguide imperfections, etc.

In practice, splice losses of about 0.5 dB are typical for fusion-spliced multimode fibres while single-mode fibres (because of the narrowness of the core diameter) are not as much (e.g. 0.1–0.2 dB). Mechanical splicing, whilst still being more lossy than fusion-splicing, is nevertheless appreciably less than that of a connector.

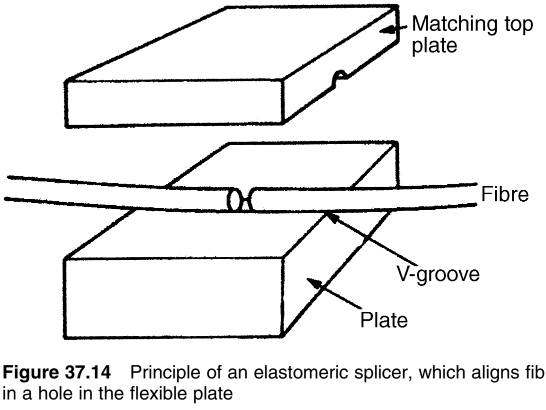


Figure 37.14 Principle of an elastomeric splicer, which aligns fibres in a hole in the flexible plate

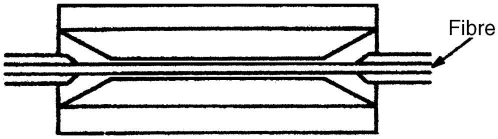


Figure 37.15 Principle of an elastomeric splicer, which aligns fibres in a hole in the flexible plate

37.3.6.2 Connectors

Demountable connectors for optical fibre (cables) need to perform several functions. The primary function is to couple light from one fibre to another efficiently and repeatably. If a connector is to operate satisfactorily it must also protect the fibre ends from damage which may occur due to handling, to protect against environmental factors such as moisture and dust, and to carry tensile loads on the cable, whilst allowing rapid connection and disconnection when required. For optimum performance in hostile environments the cable and connector must be considered as an integral unit.

An optical fibre connector can usefully be considered in three parts:

(1) Fibre terminations which protect and locate the fibre end;
(2) Alignment guides which position the pair of fibre terminations for optimum coupling; and
(3) Connector shells which protect the optical contacts from the environment, hold the alignment guides and fibre terminations in place, and terminate the cable sheath and strain member.

There are two main categories of demountable optical connector. The first is the butt joint in which the prepared ends are close to each other and are aligned so that their fibre axes coincide. The second major category uses the expanded beam technique. In this approach the diameter of the transmitted beam is increased by one half of a connector and this expanded beam is reduced again to a size compatible with the core of the receiving fibre by the second half of the connector.

This expansion can be achieved by tapering the fibre, but generally lenses are used.

Butt jointing In a butt joint, a ferrule usually protects the fibre. The main types are jewelled, ceramic, and tri-ball. A typical jewelled metal ferrule is manufactured with an accurate outside diameter and a 1 mm counterbore at one end. A standard watch jewel, with a hole size closely matching the diameter of the fibre to be used, is press-fitted into the counterbore giving a fibre size hole which is accurately concentric with the outside diameter of the ferrule. The close diameter and concentricity tolerances of the fibre hole are achieved more easily and cheaply by drilling a 1 mm hole and using a watch jewel than by drilling a small hole directly into the end of the metal ferrule.

To terminate a fibre in a jewelled ferrule, the protective coating is removed over a short length from the end, and the fibre is fed into the ferrule filled with a suitable adhesive and through the close fitting jewel hole at the front. The fibre is then polished back flush with the ferrule end. Typical concentricity errors between the fibre core and the outside diameter of the ferrule are 2–6 μm. The termination technique has been developed for silica and plastic-clad-silica fibres.

Another similar technique uses a precision ceramic rod with a central hole, and has fast become the standard for most fibre connectors as the rods can be manufactured very cost effectively.

Another technique uses three precision balls (as used in the bearing industry) to locate a fibre centrally within a conical hole.

Butt joint connectors are by far the most common; those regularly encountered are SMA, bi-conic (both types for multimode all-silica fibres), ST (for both multimode and single mode), and FC/PC (for single mode).

Expanded beams The expanded beam method uses a different mechanism to achieve alignment. When a prepared fibre end is fixed at the focus of a convex lens, a collimated beam, with a diameter greater than the fibre core diameter, emerges from the lens. When two of these terminations are aligned an optical connector is produced. The fibre must be positioned at the focus with the same accuracy as two fibres in a butt joint since the receiving fibre is in effect forming a butt joint with the image of the transmitting fibre. At first sight this may appear to forfeit the advantages of this technique; however, since the accurate positioning of the fibre in the lens termination unit is only required once, it would normally be done in a factory. Making and breaking of the connection occurs between the two lenses. The required connection tolerances are reduced since the increased beam diameter allows greater lateral misalignment than with a simple butt. Owing to the collimation of the beam, a small separation of the terminations, and/or angular misalignment can be tolerated without significantly increasing attenuation.

The increased beam diameter reduces the effect of dust on the connector loss. The separation of the terminations minimises the risk of permanent damage arising from grit scratching or chipping the optical surfaces when the connector is inadvertently coupled in a dirty condition. For these reasons, this type of connector is useful in harsh environments.

Alignment and connector shells For accurate alignment of two fibres within the connector shell, precision holes are used. Watch jewels are used for fibre guidance as well as precision bores, bioconical systems, and helical springs. In the FC/PC type (used for single mode), a factory set keying system is used to optimise connector angular orientation to optimise connector angular orientation to optimise connector loss. Springs overcome the need for a sliding fit tolerance by using the fact that as they are unwound, the inner diameter increases; at this stage the two ferrules are inserted, and then the spring relaxed.

Many connector bodies or shells are available, generally using bayonet or screw threads to mate. Hermaphroditic types are also available. The shell may be for a single connection, or for multiple fibres.

Connectors tend to be very specific to the type of cable which they will accept and the need to clamp onto the cable strength member rather than the fibre itself for strain relief is always advisable.

37.4 British and International Standards

Listed below are current British (BS), American (ANSI) and international (ISO) standards that concern optical fibres and Fibre Distributed Data Interface (FDDI).

ANSI X3.139 (1997)—INFORMATION SYSTEMS—FIBER DISTRIBUTED DATA INTERFACE (FDDI)—TOKEN RING MEDIA ACCESS CONTROL (MAC)

Describes Media Access Control, lower sublayer of Data Link Layer for FDDI, which provides high-bandwidth (100 Mbits/s), general-purpose interconnection among computers and peripheral equipment using fibre optics in ring configuration.

ANSI X3.148 (1999)—INFORMATION SYSTEMS—FIBER DISTRIBUTED DATA INTERFACE (FDDI)—TOKEN RING PHYSICAL LAYER PROTOCOL (PHY)

Describes Physical Layer Protocol standard, upper sublayer of Physical Layer, for FDDI which provides high-bandwidth (100 Mbit/s), general interconnection among computers and peripheral equipment using fibre optics. Coverage includes definitions, conventions, coding, symbol set, line states, coding overview, general organization, smoothing function, repeat filter and ring latency. Also gives detailed diagrams and tables.

ANSI X3.166 (1995)—FIBER DISTRIBUTED DATA INTERFACE (FDDI) PHYSICAL LAYER MEDIUM DEPENDENT (PMD)

Gives a specification for Physical Layer, Medium Dependent (PMD) requirements for the FDDI high-bandwidth (100 Mbit/s) general purpose interconnection among peripheral and computer equipment that use fibre optics as the transmission medium.

ANSI X3.263 (1995)—FIBRE DISTRIBUTED DATA INTERFACE (FDDI)—TOKEN RING TWISTED PAIR PHYSICAL LAYER MEDIUM DEPENDENT (TP-PMD)

Defines Twisted Pair Physical Layer Medium Dependent (TP-PMD) criteria for FDDI high-bandwidth, (100 Mbit/s) general purpose interconnection among computers and equipment using fibre optics and twisted pair as the transmission media. It can be shaped to support a sustained data transfer rate of 80 Mbit/s or more, and allows connection for nodes distributed over distances of several kilometres. Default values are calculated on the basis of 1000 physical links and total fibre path of 200 km in length. (Typically corresponding to 500 nodes and 100 km of dual fibre cable.)

BS 6004 (1995)—SPECIFICATION FOR PVC-INSULATED CABLES (NON-ARMOURED) FOR ELECTRIC POWER AND LIGHTING

Specifies requirements and dimensions for non-armoured PVC insulated cables for fixed installation and for operation at voltages up to and including 450 V to earth and 750 V a.c. between conductors. Coverage includes: voltage designation, conductors, insulation, core identification, sheath, marking, electrical requirements and fillers and extruded inner covering. Also gives definitions, tables, diagrams and annexes.

BS EN 186000 PT1 (1994)—HARMONIZED SYSTEM OF QUALITY ASSESSMENT FOR ELECTRONIC COMPONENTS. GENERIC SPECIFICATION: CONNECTOR SETS FOR OPTICAL FIBRES & CABLES—REQUIREMENTS, TEST METHODS AND QUALIFICATION APPROVAL PROCEDURES

Applies to fibre optic connector sets for optical fibres and cables. Coverage includes definitions, environmental category, gauges, corrosion resistance, component marking, package marking, spectral loss, cable torsion, cable pulling, static load, nuclear radiation, solar radiation, spectral loss, axial compression and industrial atmosphere.

BS EN 187000 (1997)—GENERIC SPECIFICATION FOR OPTICAL FIBRE CABLES

Applicable to optical fibre cables to be used with telecommunication equipment and devices having similar techniques and combining optical fibres and electrical conductors.

IEC 60793-1 (1992)—OPTICAL FIBRES—GENERIC SPECIFICATION

Applies to primary coated or primary buffered optical fibres for use in telecommunication equipment and in devices employing similar techniques. Establishes uniform requirements for the geometrical, optical, transmission, mechanical and environmental properties of optical fibres and includes measuring methods for dimensions, transmission and optical characteristics. A guide for fibres for short distance links is in Annex A.

IEC 60793-1.1 (1999)—OPTICAL FIBRES—GENERIC SPECIFICATION—GENERAL

Applies to primary coated or buffered optical fibres for use in telecommunication equipment and in devices employing similar techniques and defines categories of optical fibres as well as packaging.

IEC 60793-1.2 (1996)—OPTICAL FIBRES—GENERIC SPECIFICATION—MEASURING METHODS FOR DIMENSIONS

Gives the measuring methods applicable to environmental tests of optical fibres. The methods are to be used for inspection of optical fibres for commercial purposes. They establish uniform requirements for geometrical characteristics of optical fibres.

IEC 60793-1.3 (2000)—OPTICAL FIBRES—PART 1–3: GENERIC SPECIFICATION—MEASURING METHODS FOR MECHANICAL CHARACTERISTICS

Applicable to the tests of mechanical strength, ease of handling or the recognition of physical defects or primary coated or primary buffered optical glass fibres. The methods are to be used for inspecting fibres for commercial purposes. The aim of this part is to establish uniform requirements for mechanical characteristics of optical fibres.

IEC 60793-1.4 (1998)—OPTICAL FIBRES—GENERIC SPECIFICATION—MEASURING METHODS FOR TRANSMISSION AND OPTICAL CHARACTERISTICS

Applies to the practical measurements of transmission and optical parameters of fibre. The methods are to be used for inspection of fibres and cables for commercial purposes to establish uniform requirements.

IEC 60793–1.5 (1995)—OPTICAL FIBRES—GENERIC SPECIFICATION—MEASURING METHODS FOR ENVIRONMENTAL CHARACTERISTICS

Describes measuring methods which apply to environmental tests of optical fibres. The methods are to be used for inspection of optical fibres.

IEC 61218 IEC (1993)—FIBRE OPTICS—SAFETY GUIDE

An interpretation of IEC 825 with respect to fibre optic transmission systems. Some additions have been made in this guide wherever IEC 825 and its amendment do not cover a specific fibre optic subject. Applies to the wavelength range of 400 nm–10 (to the power 5) nm. IEC 825 addresses equipment based on lasers only.

IEC 61292-1 (1998)—FIBRE OPTICS—PARAMETERS OF AMPLIFIER COMPONENTS

Applicable to optical components of Optical Fibre Amplifiers (OFAs). Provides information on the most relevant parameters of OFA optical components, but does not define included definitions as these require more research.

IEC 61930 (1998)—FIBRE OPTIC GRAPHICAL SYMBOLOGY

Applicable to graphical symbols that are used in IEC publications dealing with fibre optics. The aim is to give uniform graphical symbols for the various fibre optic elements and devices.

IEC TR 61282-1 (2000)—FIBRE OPTIC COMMUNICATION SYSTEM DESIGN GUIDES—PART 1: SINGLE-MODE DIGITAL AND ANALOGUE SYSTEMS

Gives guidance for system design methodology for users and suppliers of fibre optic transmission systems for telecommunications and broadband video distribution applications. The function of the systems is to interconnect signals between defined digital and analogue interfaces via a fibre optic link. Generally, fibre optic systems are built-up from Basic Fibre Optic Systems (BFOS).

ISO 9314-1 (1989)—INFORMATION PROCESSING SYSTEMS—FIBRE DISTRIBUTED DATA INTERFACE (FDDI)—TOKEN RING PHYSICAL LAYER PROTOCOL (PHY)

Specifies the Physical layer protocol (PHY), for FDDI.

ISO 9314-2 (1989)—INFORMATION PROCESSING SYSTEMS—FIBRE DISTRIBUTED DATA INTERFACE (FDDI) TOKEN RING MEDIA ACCESS CONTROL (MAC)

Specifies the Media Access Control (MAC), the lower sublayer of the Data Link Layer (DLL), for an FDDI high-bandwidth (100 Mbit/s), general-purpose interconnection among computers and peripheral equipment using a transmission medium of fibre optics in a ring configuration.

ISO 9314-3 (1990)—INFORMATION PROCESSING SYSTEMS—FIBRE DISTRIBUTED DATA INTERFACE (FDDI)—PHYSICAL LAYER MEDIUM DEPENDENT (PMD)

Describes Physical Layer, Medium Dependent requirements for FDDI high-bandwidth (100 Mbit/s) general-purpose interconnection among computers and peripheral equipment using fibre optics as the transmission medium. May be configured to support a sustained transfer rate of approximately 80 Mbit/s. May not meet the requirements of all unbuffered high-speed devices. It establishes the connection among many nodes distributed over distances of several kilometres. Default values were calculated on the basis of 1000 physical connections and a total fibre path length of 200 km.

ISO 9314-13 (1998)—INFORMATION TECHNOLOGY—FIBRE DISTRIBUTED DATA INTERFACE (FDDI)—CONFORMANCE TEST PROTOCOL IMPLEMENTATION—CONFORMANCE STATEMENT (CT-PICS) PROFORMA

Gives the PICS proforma for the FDDI specified in the base standards as denoted in clause 5.

37.5 Optical fibre telemetry on overhead power lines

37.5.1 Introduction

The advent of optical fibre links has provided significant improvements in the communication facilities needed for the effective operation of power systems. Such links are well suited to the electrically noisy environments of power systems due to their immunity from electromagnetic interference. They are also free from the difficulties encountered with purely electronic systems due to local earth potential variations which may arise from the heavy earth currents that can occur during a power system fault. The wide information-carrying bandwidths of such systems leads to low costs per channel on trunk routes, can conveniently accommodate extra information transmission and provide the means for more rapid signal transmission which is important for protection and control under power system fault conditions. In addition, they are electromagnetically compatible in not generating stray or spurious interference as can be the case with microwave and power-line carrier systems.

The main uses on power systems for long-haul optical fibre links include telephony, telemetry, telecontrol, protection (the control of circuit-breakers), data transmission between computers, and possibly video signalling.

37.5.2 Fibre cable configurations

Optical fibre cables are generally available for use on overhead power lines. Such cables fall into three main generic types; (1) All-Dielectric Self-Supporting (ADSS); (2) spiral wrap-on or lashed; and (3) Optical Ground Wire (OPGW) where the optical fibres are included in the earth wire.

All-dielectric self-supporting cables require tensile members which offer a high strength-to-weight ratio such as glass reinforced plastic or aramid yarns. The presence of these cables imposes an additional load on the support structures which must be considered with regard to statutory support safety factors. An ADSS cable is inherently light and small relative to the phase conductor and earth wire which limits the effects of environmental factors such as wind and ice loads minimising any additional steel work or structural reinforcement needed. Clamping arrangements can also be chosen which allow slippage of the cable under conditions of differential loading (i.e. when one span adjacent to a support structure is loaded more or less than the other span adjacent to that support structure) further safeguarding the integrity of the supports. The optical fibre cables must also meet any ground clearance requirements imposed by the power utility. The principal benefits of ADSS cables are that they offer post-fit solutions for power utilities and can allow installation with one or both power circuits live avoiding costly supply interruptions.

Wrap-on or lashed cables can be fitted to either the earth wire or a phase conductor, although the earth wire is most common. The wrap-on cables are spirally applied to minimise the susceptibility of the resulting composite line to wind-induced vibration known as 'galloping'. Galloping is caused by light winds flowing over the cable giving rise to high and low pressure regions (similar to the action of an aircraft wing) producing up-lift or down-force depending on the orientation of the cables. By spirally applying the wrap-on cable the effective profile presented to the wind varies along the span such that consecutive spirals oppose each other cancelling any net effect.

Wrap-on or lashed cables also offer the advantage of being post-fit systems if the cable is installed on the earth wire. Live line installation of such systems on the phase conductor are being discussed.

Optical ground wire (OPGW) cables encase the optical fibres within the earth wire which is used to carry fault currents and to provide shielding from lightning. One structure of such a composite cable includes aluminium or steel strands helically wound around a hollow aluminium core carrying a polyethylene sheath which holds typically between two and four pairs of optical fibres, along with the option of a polymer tension member.

The OPGW cables offer the advantage of not imposing additional loads on the support structure as they replace the existing earth conductor. However, these cables are not post-fit solutions to the needs of power utilities for telecommunications purposes as the replacement of the existing earth wire is costly and time consuming. Therefore, such cables are generally used on new lines or when the existing earth wire is scheduled for replacement.

Any optical fibre cable used for aerial applications must be capable of operating under a range of environmental conditions without eroding the optical and power transmission characteristics or the service life of the cables and support structures.

Optical fibre cables incorporate strain relief between the strength member and the fibres to protect the optical fibres during service. Strain relief can be provided by overfeeding fibre into small polymer tubes which are then stranded around a central member or by overfeeding ribbon arrays of fibres within a slot. The amount of strain relief required is determined by relating the maximum cable strain expected during service to the maximum allowable fibre strain calculated for the required service lifetime (a function of the initial fibre proof test level) and the optical performance of the fibres when they are strained. Multiple fibres constrained in loose tubes tend to interact with each other when strained, giving increased attenuation.

The variation in cable tension and strain produced by climatic changes such as temperature, wind strength, and ice loads can be solved mathematically. The optical cable, as with the phase conductors and earth wires, adopts the shape of a catenary when suspended between two points. An initial set of known conditions (usually corresponding to those during installation) are used to determine the basic catenary and then all changes in tension due to climatic variations are related to this. The principal effect of a wind acting on the cable, apart from that on the tension in the cable, is to cause the cable to 'blow out' from the vertical.

Cable or conductor ageing must also be considered. ADSS cables generally have good self-damping characteristics which vibration in the cable and that transmitted to the support structure. Wrap-on and lashed cables tend to vibrate in harmony with the conductor or earth wire on which they are installed and hence do not produce a significant effect on the vibration behaviour of the host. OPGW cables are used instead of standard earth wires and behave in a similar fashion to those wires. Therefore, OPGW cables tend to use the same precautions (typically dampers) as earth wires.

In addition to vibration, a further consideration is a phenomenon known as 'creep' where the presence of a permanent tensile load produces a reduction in the strength of the cable or conductor leading to failure below the initial breaking load. Optical cables use strength members manufactured from materials which have been shown to be resistant to 'creep' within the operational strain window experienced during service.

The electric field in which these optical fibre cables reside can combine with onerous climatic conditions and have an effect on the sheath materials used. ADSS and wrap-on cables can be degraded by the local electric field. Careful selection of sheath materials which are resistant to dry band arcing, the phenomenon which causes sheath degradation, ensures that such cables fulfil service lifetime requirements. OPGW cables have metallic outer layers and are used for earthing purposes and so are not prone to damage from dry band arcing but can be subject to galvanic action between the dissimilar metals used in the cable construction.

37.5.3 Factors governing system design

In the UK, telephony traffic over single-mode fibre is common with bit rates up to 565 Mbit/s, and systems are installed operating at over 2 Gbit/s. The major telephony carriers use single-mode fibre only. Operation is generally at an optical wavelength of 1300 nm, with longer haul systems (e.g. intercontinental submarine links) using 1.55 µm where intrinsic fibre losses are lower (ca. 0.2 dB/km, and ca. 0.35 km at 1.3 µm). For a laser diode source with output power of typically—3 dB/m, and photodiode receiver sensitivity—50 dB/m, link lengths greater than 50 km and bit error rates of 10^{-9} are readily achieved, with a power margin to allow for component ageing and possible fibre repair splices. Such lengths generally allow for signal repeaters/regenerators to be situated within buildings.

Multimode fibre is used for data transport, analogue video transmission, etc., over short distances to approximately 2 km. Both 62.5/125 µm and 50/125 µm (core/cladding diameters) are commonly used, the former being the preferred FDDI standard. Multimode fibre, despite its higher cost, is used to allow the use of more cost-effective connectors, and launch and receive devices.

37.6 Power equipment monitoring with optical fibre sensors

37.6.1 Introduction

Although early forms of optical fibre sensing systems have illustrated the potential of such methods for power system monitoring there have been practical deficiencies which have detracted from their widespread use. The original systems were cumbersome, unreliable, costly and involved much unfamiliar optical processing and interfacing. It may be argued that insufficient attention was given to the real needs of power system monitoring. The problem has been exacerbated because the power industry has been unsure of the modes in which such novel technology might best be used whilst simultaneously a major thrust of optical fibre sensors research has been for methods which are over-sophisticated and costly for power system applications.

However, the many fundamental advantages of optical fibre sensing systems remain attractive. For instance, a major advantage is for the condition monitoring of a circuit-breaker during fault current interruption when signatures of impending faults may be more distinguishable. A further implication is that such an approach minimises supply disruption since such monitoring would be undertaken live, a capability which emerges because of such a system's electromagnetic immunity and its inherent electrical insulation properties.

This contribution examines the advantages of optical fibre monitoring for power equipment applications, reasons for limited progress in implementing such technology and possible strategies which are evolving for the site testing and evaluation of these systems.

37.6.2 Technology implementation difficulties

One objective of the realisation of optical fibre sensing in the power industry is for the implementation of an optically controlled substation system. Such an objective involves producing optical fibre sensors for monitoring a range of parameters governing the condition and operation of power equipment such as circuit-breakers and transformers. It entails interlinking the various monitoring systems with an optical fibre system and connecting these to various data stations and control units (*Figure 37.16*).

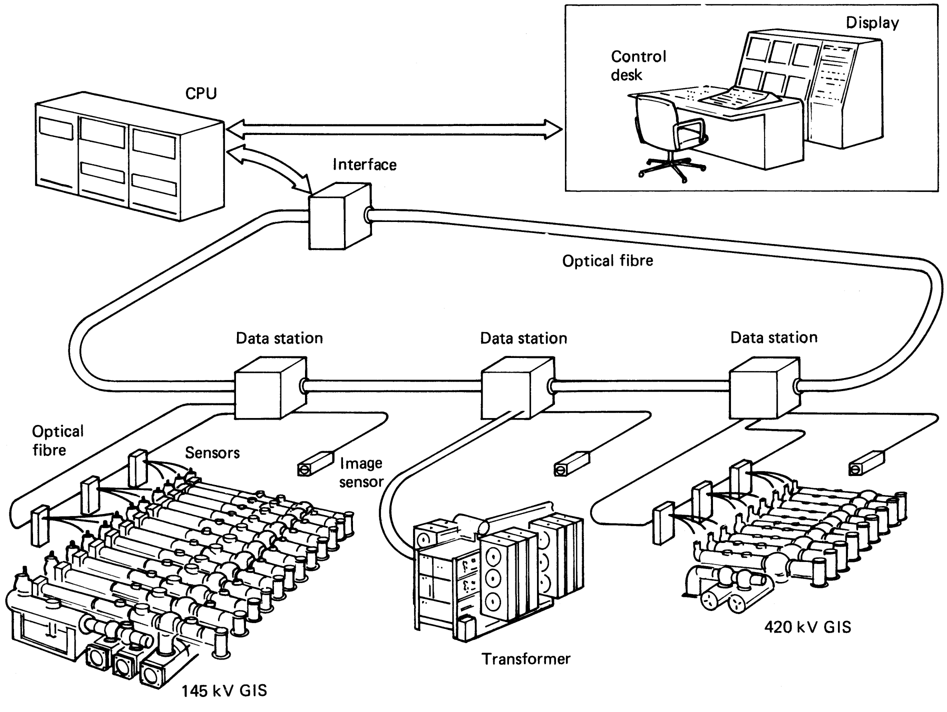


Figure 37.16 Optically monitored substation concept. (Source: Courtesy of Mitsubishi)

The attraction of optical fibre sensors for such monitoring is that their intrinsic properties offer many advantages in hostile and challenging environments. For high-voltage power apparatus applications these properties include:

- inherent isolation of electronic instruments;
- immunity from electromagnetic interference;
- geometric flexibility;
- inherent electrical insulation;
- no electrical shock hazard;
- corrosion resistance; and
- compact and lightweight.

Despite such attractive advantages, the uptake of the fibre sensing technology has been retarded because of general market penetration problems. These include:

- uncompetitive prices;
- end-user unfamiliarity;
- conservative attitude of large industries;
- range of different optical systems for monitoring various parameters;
- failure of early, immature systems; and
- absence of established markets.

Recent developments in optical fibre sensing have addressed these commercial problems. In order to appreciate the significance of the developments it is necessary to review briefly the methods available for modulating an optical signal.

The basic modulation methods involve modifying either the amplitude, phase or polarisation of the optical signal (*Figure 37.17*). Absolute amplitude monitoring is unreliable for sensing purposes because a number of external (e.g. fibre bending) and intrinsic factors (e.g. source variability) in addition to the sensor may modify the amplitude. Phase monitoring is based upon optical interferometry which involves complex instrumentation and because of its digital nature is not so attractive for situations in which the instrument power supply may be interrupted. Methods based upon monitoring changes in the plane of polarisation of an optical signal have hitherto relied upon even more complex optical systems.

Attempts to overcome the deficiencies of the amplitude modulation method by referencing a modulated signal at one wavelength (λ_1) with respect to an unmodulated signal at a second wavelength (λ_2) has met only with limited success because the referencing is not completely reliable and because of power-budget limitations.

The problem with these basic methods is that the intention with any measurement system is to seek an output (V) which is proportional to the measurand (M) via a constant which is the sensitivity S:

$$V = S \cdot M \qquad (37.9)\Leftarrow$$

With an optical fibre system the relationship (Equation (37.10)) between the output V and modulation (by measurand) $M_1(\lambda)$ is complicated and depends upon parameters such as source power ($P(\lambda)$), fibre transmission ($T(\lambda)$), and fibre perturbation ($M_2(\lambda)$) which can be induced by external influences or age.

$$V = T \sum_{1.m} \int_{\lambda_s} \left([P(\lambda)T(\lambda)M_2(\lambda)\mathrm{d}l] R(\lambda) M_1(\lambda)\mathrm{d}\lambda \right)^{P} \qquad (37.10)\Leftarrow$$

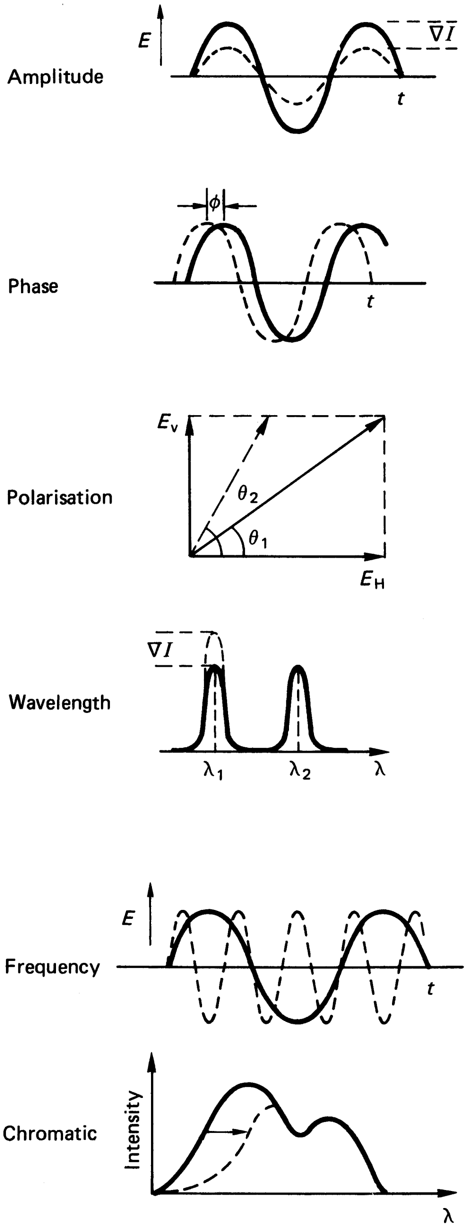


Figure 37.17 Basic optical modulation methods

It was, therefore, necessary to seek modulation methods which are less susceptible to these effects. Two main methods emerged. The first relied upon frequency modulation which was solution adopted by the telecommunication industry. The second method relied upon chromatic modulation whereby the spectral content of the optical signal is varied and the optical signal is monitored over its entire spectral range by several detectors each having a different but overlapping spectral response.

Both these methods are intensity independent and lead to practical, cost effective fibre sensing systems. The chromatic approach has the added bonus of using common instrumentation for monitoring a range of different parameters.

Bibliography

1 KAO, C. K., *Optical Fibres*, Peter Peregrinus (1988)
2 MURATA, H., *Handbook of Optical Fibres and Cables*, Marcel Dekker, New York (1988)
3 SENIOR, J. M., *Optical Fibre Communications, Principles and Practice*, Prentice Hall, New York (1985)
4 SNYDER, A. W. and LOVE, J. D., *Optical Waveguide Theory*, Chaptman and Hall, London (1983)
5 TRICKER, R. L., *Optoelectronic Line Transmission*, Heinemann Newnes, London (1989)
6 TRICKER, R. L., *Optoelectronic and Fiber Optic Technology*, Newnes, Oxford (2002)

38

Installation

P Whitewood
Parsons Brinckerhoff Ltd

Contents

38.1 Layout 38/3
    38.1.1 System supply 38/3

38.2 Regulations and specifications 38/3
    38.2.1 Purpose 38/4
    38.2.2 Scope 38/4

38.3 High-voltage supplies 38/4

38.4 Fault currents 38/5

38.5 Substations 38/5
    38.5.1 Low-voltage equipment 38/7
    38.5.2 Packaged substations 38/7
    38.5.3 Low-voltage distribution 38/7

38.6 Wiring systems 38/9
    38.6.1 Steel conduit 38/9
    38.6.2 Plastic conduit 38/9
    38.6.3 Trunking 38/10
    38.6.4 Cabling 38/10

38.7 Lighting and small power 38/11
    38.7.1 Lighting circuits 38/11
    38.7.2 Small power circuits 38/11

38.8 Floor trunking 38/12
    38.8.1 Underfloor trunking 38/12
    38.8.2 Open-top trunking 38/12
    38.8.3 Cavity-floor systems/dado trunking 38/12

38.9 Stand-by and emergency supplies 38/13

38.10 Special buildings 38/13
    38.10.1 High-rise buildings 38/13
    38.10.2 Public buildings 38/13
    38.10.3 Domestic premises 38/13

38.11 Low-voltage switchgear and protection 38/13
    38.11.1 Air circuit-breakers 38/14
    38.11.2 Moulded-case circuit-breakers 38/14
    38.11.3 Miniature circuit-breakers 38/15
    38.11.4 Fuses 38/15
    38.11.5 Prospective fault current 38/15
    38.11.6 Discrimination 38/16
    38.11.7 Motor control gear 38/16

38.12 Transformers 38/17
    38.12.1 Installation 38/18
    38.12.2 Transformer protective devices 38/19

38.13 Power-factor correction 38/19
    38.13.1 Capacitor rating 38/19

38.14 Earthing 38/19
    38.14.1 Earth electrode 38/20
    38.14.2 Soil resistivity 38/21
    38.14.3 Electrode installation 38/21
    38.14.4 Resistivity and earth resistance measurement 38/22
    38.14.5 Protective conductors 38/22
    38.14.6 System earthing 38/23

38.15 Inspection and testing 38/23
    38.15.1 Inspection 38/23
    38.15.2 Testing 38/23
    38.15.3 Test gear 38/25

38.1 Layout

This chapter deals with the installation of a power supply to a consumer. One example might be a large industrial concern with a factory complex spreading over several hectares; another the single tower-block with one supply entry point in the basement but with additional substations at intermediate floors. At the other extreme is the small dwelling-house.

The general rules of safety of personnel and equipment, continuity of supply and ease of operation and maintenance apply to all consumers, but the layout naturally varies with the individual establishment.

38.1.1 System supply

When an electricity supply is to be provided to a factory or building complex, there are several features that must be taken into account. Safety is one that cannot be measured in terms of capital cost, but it should not be difficult to achieve an acceptable standard, provided that good-quality equipment is purchased, properly installed and well maintained, and operated by experienced staff.

Loss of supply may result in loss of output from factory or danger to life in a hospital. The system designer must take these matters into account. To eliminate all risk of outage is likely to be very costly: the alternative of accepting some measure of risk is often preferable, bearing in mind the present-day reliability of the public electricity supply and the equipment available for installation.

Losses occur in cables, overhead lines and transformers, and the system should be designed to minimise these, by locating substations as near as possible to the centres of the load. The system power factor is also relevant, for it can affect equipment sizing, I^2R losses and the cost of electrical energy.

In few installations is maintenance afforded the importance it deserves. The design and installation engineer cannot control this, but he can ensure that the equipment supplied allows for proper isolation and regular testing with the minimum interruption of supply.

Few installations remain unchanged throughout their life. Whether it be the provision of a spare way on a distribution board or space for an additional high-voltage ring-main circuit, thought should be given to the problem at the time of design and installation.

In determining the cost of installation it must be borne in mind that, in addition to the capital cost of the equipment, attention must be given to the running cost, and also to the cost of providing accommodation for the equipment. For example, it may be better to provide outdoor equipment rather than less costly indoor plant when provision of a building to house it, is taken into consideration.

38.2 Regulations and specifications

In the UK, the supply of electricity to premises is governed by the Electricity Supply Regulations, which require that all electrical equipment in the premises shall be maintained in an efficient state and that, as regards high-voltage equipment, authorised persons are available to cut-off supply in emergency. Users must also be mindful of the Electricity at Work Regulations 1989 (EAW). These have extended the powers of the Health and Safety Executive to all places of work. They impose legal requirements on a wide range of personnel, with the object of insuring safe working practices in the use, operation and maintenance of electrical installations. The reader's attention is particularly drawn to an associated document 'Memorandum of Guidance', which as its title suggests, provides extremely useful help in interpreting the EAW Regulations. It should be pointed out that the Health and Safety Executive regards compliance with the Institution of Electrical Engineers (IEE) Wiring Regulations [BS7671:1992 with amendments] (referred to in detail in the following paragraphs) as being likely to achieve compliance with the relevant aspects of the EAW Regulations, but it is also necessary to have safe operational and maintenance procedures for personnel working on, or near, electrical installations to ensure full compliance with the EAW Regulations.

The actual erection and installation should comply with the IEE Wiring Regulations—Requirements for Electrical Installations: compliance with these Regulations will produce a high standard of work and allow a good factor of safety against possible breakdown.

In 1981 the new edition (15th) of these Regulations (referred to as WR15) marked a radical change in both style and approach. It is based on internationally agreed installation rules, the two bodies concerned being the International Electro-technical Commission (IEC) and the European Committee of Electro-technical Standardisation (CENELEC). This approach has now been carried a stage further with the introduction in May 1991 of the 16th edition of the Wiring Regulations (WR16). The concept has been carried further with the 1992 updated edition of the 16th edition of the Wiring Regulations (WR16) and the 2000 amendment no 3 available on the internet.

Every IEC Standard issued expresses, as nearly as possible, an international consensus on its subject, and the intention is that every member country should adopt the text of the standard, as far as the individual country's conditions will permit. Additionally, the members of the EEC are committed to removing trade barriers, and CENELEC aims to help in this by attempting to 'harmonise' the corresponding national standards. In 1968 IEC started work to formulate standard rules. It was soon realised that combining the existing rules of member countries was not feasible, and so IEC decided to go back to fundamentals. Whilst the work continues, some sections have been published, and the IEE made use of the IEC plan and the technical content already agreed by IEC in revising the Wiring Regulations. This resulted in the 15th and 16th editions of the Wiring Regulations being very different from any previously published. Basically, they were aimed directly at the designer of electrical installations rather than at the site installer. They demanded an analytical, mathematical approach to the design of the installation. Additional design time is needed, but the designer is given a greater degree of freedom to produce an economic design.

In the latest edition (WR16) there is no fundamental change in this approach, but harmonisation is taken a stage further, as can be seen from the list of CENELEC harmonisation documents in the Preface. The number of appendices has been greatly reduced, but in their place the IEE has published a series of guidance notes on specific subjects. Of particular relevance is the 'On Site Guide', which should prove invaluable for the installer who may not be totally familiar with every detail of the Wiring Regulations. By following the 'On Site Guide' in his work the installer can be assured that the installation will be acceptable, and safe in operation, although it may not be the most economical design possible.

Mention must also be made of two books sponsored by the Electrical Contractor's Association, the Handbook on the 16th Edition of the IEE Regulations (which is also sponsored by the National Inspection Council for Electrical Installation Contracting) and Electrical Installation Calculations, the latter by B. D. Jenkins.

It must be pointed out that casual and occasional reference to the Wiring Regulations is not enough to inform readers of the requirements. This can only be achieved by study of the Regulations in detail.

Mention can be made of other regulations and specifications used in other parts of the world. In the former UK colonies earlier versions of the IEE Wiring Regulations have been adopted and modified to suit the engineering environment of the country, i.e. The Central African Standards (CAS Wiring Rules) are based on the earlier 13th and 14th IEE Wiring Regulations and are used in central Africa including Zimbabwe and Malawi.

In parts of the world where there is an American influence then the NFPA70 (National Electrical Code) is adopted.

The National Fire Protection Association has acted as sponsor of the National Electrical Code since 1911. The original Code document was developed in 1897 as a result of the united efforts of various insurance, electrical, architectural, and allied interests.

NFPA has an Electrical Section that provides particular opportunity for NFPA members interested in electrical safety to become better informed and to contribute to the development of the National Electrical Code and other NFPA electrical standards. Each of the Code-Making Panels and the Chairman of the Correlating Committee reported their recommendations to meetings of the Electrical Section at the 1995 NFPA Annual Meeting. The Electrical Section thus had opportunity to discuss and review the report of the National Electrical Code Committee prior to the adoption of this edition of the Code by the Association.

38.2.1 Purpose

Practical Safeguarding The purpose of this Code is the practical safeguarding of persons and property from hazards arising from the use of electricity.

Adequacy This Code contains provisions considered necessary for safety. Compliance therewith and proper maintenance will result in an installation essentially free from hazard but not necessarily efficient, convenient, or adequate for good service or future expansion of electrical use.

Hazards often occur because of overloading of wiring systems by methods or usage not in conformity with this Code. This occurs because initial wiring did not provide for increases in the use of electricity. An initial adequate installation and reasonable provisions for system changes will provide for future increases in the use of electricity.

Intention This Code is not intended as a design specification nor an instruction manual for untrained persons.

38.2.2 Scope

The Code covers:

1. Installations of electric conductors and equipment within or on public and private buildings or other structures, including mobile homes, recreational vehicles, and floating buildings; and other premises such as yards, carnival, parking, and other lots, and industrial substations.

   For additional information concerning such installations in an industrial or multibuilding complex, see the *National Electrical Safety Code*, ANSI C2-1993.
2. Installations of conductors and equipment that connect to the supply of electricity.
3. Installations of other outside conductors and equipment on the premises.
4. Installations of optical fibre cable.
5. Installations in buildings used by the electric utility, such as office buildings, warehouses, garages, machine shops, and recreational buildings that are not an integral part of a generating plant, substation, or control centre.

38.3 High-voltage supplies

The general method of supplying bulk power to factories or other complexes is by means of an HV supply, usually at 11 kV, occasionally at 6.6 kV. The installation comprises a main substation at the point of entry, with HV cables to supply subsidiary substations located near load centres. Where desirable, it may be possible for the consumer to obtain two supplies from the power authority, and in special cases where loss of supply could give rise to a particular hazard (e.g. the danger to life in a hospital), the two supplies may be provided from separate sources. In buildings containing IT equipment requiring $n+1$ reliability in power supplies then dependability calculations would be undertaken to determine the Reliability, Availability, Maintainability and Safety (RAMS) of the bulk power supply systems.

The distribution system generally comprises either radial feeders or a ring main using underground cables to supply the subsidiary substations. The emphasis today is towards ring-main supplies, and this is considerably helped by the ready availability of competitively priced ring-main switchgear. The basic unit comprises two switches, earlier installations used to be oil but SF_6 has now taken over, capable of making on to a fault and of breaking load current, and

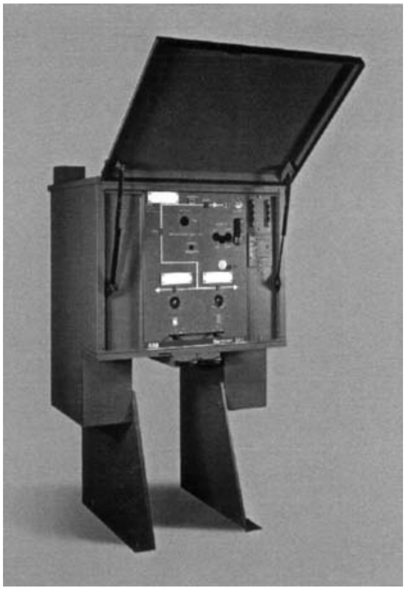


Figure 38.1 SF_6 insulated ring main unit. (Courtesy of ABB)

a fuse switch. Other units are available, assembled into various configurations. They can include circuit-breakers. The SF_6 insulated ring-main unit switching takes place in a SF_6 pressurised compartment. This equipment has obvious advantages for indoor installation because of the reduced fire risk.

A SF_6 ring-main unit is illustrated in *Figure 38.1*.

Four distribution arrangements are illustrated in *Figure 38.2*(*a–d*), although variants are also available.

Figure 38.2(*a*) is a *radial-feeder* arrangement with each feeder controlled by its own automatic circuit breaker. A cable or transformer fault will result in loss of supply only to the one sub-station, but supply cannot be restored until the repair is effected.

Figure 38.2(*b*) is a simple ring-main arrangement which allows for any individual transformer substation to be isolated for inspection or test without interrupting the supply to the other substations. Any fault occurring on the system will result in a total shut down, but the fault can be isolated and the supply restored to all substations but one, unless the fault is in the cable between substations C and D, in which case all substations can be re-energised.

In *Figure 38.2*(*c*) the distribution substations have their outgoing cables controlled by switches and the transformer by a fuse switch. Hence, every transformer can be independently switched, and should a fault occur and one of the fuses blow, the fuse switch is arranged to open all three phases automatically, thus disconnecting the transformer from the system.

The system would normally be operated with the ring open, but any cable fault will result in part of the system being shut down until the fault is located, when power can be restored to all the substations.

By using automatic circuit-breakers throughout, as shown in *Figure 38.2*(*d*), a greater degree of security of supply is obtained. The system can be operated as a closed ring, and even greater security is obtained by duplicating the transformers as shown in the figure.

Modern transformers and switchgear are very reliable, and, provided that cables are installed carefully, the number of faults that will occur will be extremely small. Since cost will always be of importance, it is common to use the simplest ring-main switchgear configuration that is acceptable for the particular application, and switches with a fuse-switch tee-off represent very good value for money and are widely used both by public electricity authorities and by industry. The situation is helped by the availability of fault monitoring and locating equipment, usually of the earth fault passage indication variety. This comprises a core balance current transformer on the out-going feeder cable with an associated hand or self-reset relay. Remote tripping of the switch fuse is also available. With the advent of low-cost micro-processor equipment the system of fault location and monitoring has been extended to provide more exact information to identify the location of a fault.

38.4 Fault currents

A relatively simple calculation determines the required interrupting capacity of switchgear. This is particularly useful to the installation engineer when it is necessary to provide extensions to an existing system. He must ensure that any changes in fault currents are safely interrupted via the breaking capacity of 13.1 kA at 11 kV switchgear. Where old gear is already in use, problems could arise.

The technique involves using the reactances of the equipment. In the case of underground cables, the resistance is high compared with the reactance and so for calculation purposes the impedance of the cable is used. For generators the sub-transient reactance is used, as it is effective for the first few cycles following a fault.

The ratio between the phase voltage drop and the normal line-to-neutral voltage for a stated current (or kilovolt-amperes (kVA)) is the per-unit (or per cent) impedance at that current. For example, if a 1000 kVA transformer has a 0.05 p.u. (or 5%) reactance, then its voltage drop on rated load is 0.05 of the line-to-neutral voltage. It follows that if the transformer output terminals are short-circuited, the volt drop is 1.0 p.u. (100%) and the short circuit kVA loading is $1000 \times (1/0.05) = 20\,000$ kVA, and this is the maximum that the transformer can pass. Again, consider a length of 11 kV cable of impedance 1 Ω/phase and carrying 1000 kVA. The current is $I = 1000/(11\sqrt{3}) = 52.5$ A, the phase voltage is $E_{\mathrm{ph}} = 11/\sqrt{3} = 6.35$ kV, the volt drop is $v = 52.5 \times 1.0 = 52.5$ V, and the impedance of the cable is $(52.5/6350) = 0.0083$ p.u. (0.83%).

Calculation of the interrupting capacity S required at a point in a system is by assuming an arbitrary kVA/S_0 in the system, and evaluating the voltage required to pass this load as a fraction x of the line-to-neutral voltage. Then $S = S_0/x$. The effective reactance of the system configuration involves summing branches in series and parallel and, in some cases, the application of the star/delta transformation technique. The availability of computer software that enables engineers to undertake fault analyses, protection and regulation calculations now speeds up the calculation process. *Table 38.1* gives typical values of reactance for generators and transformers.

38.5 Substations

The space requirements of a substation depend on the equipment to be housed, and on whether a new building can be erected for it or it has to be fitted into an existing building. In the latter case it may be difficult to achieve an ideal solution, but where no severe limitations are imposed the layout in *Figure 38.3* would prove satisfactory. This is suitable for a main 11 kV substation, also supplying local l.v. distribution, and it will be seen that it meets most of the following requirements:

(1) There is adequate clearance around the equipment and space to withdraw circuit-breakers for maintenance.
(2) Equipment operating at different voltages is segregated (advisable but not mandatory).
(3) There is sufficient space for drawing in and connecting cables, and for the delivery and erection of additional switchgear: doorways are high and wide enough.
(4) The walls can act as fire barriers. If a h.v. bus-section switch is included, a decision must be made as to the need for fire barrier walls between the sections. Even with oil circuit-breakers, this risk is small and many engineers disregard it.
(5) Transformers are usually housed in open or semi-open compounds; but if an indoor location is essential, particular attention must be paid to its ventilation. The risk of a transformer fire is extremely small: nevertheless, an oil sump (usually filled with pebbles) should be provided to trap burning oil which could escape following a fire.

The substation design described above can be said to be traditional, but in recent years, with the almost universal use of vacuum and gas-insulated switchgear from 11 to 15 kV, and with the growth in availability of cast-resin

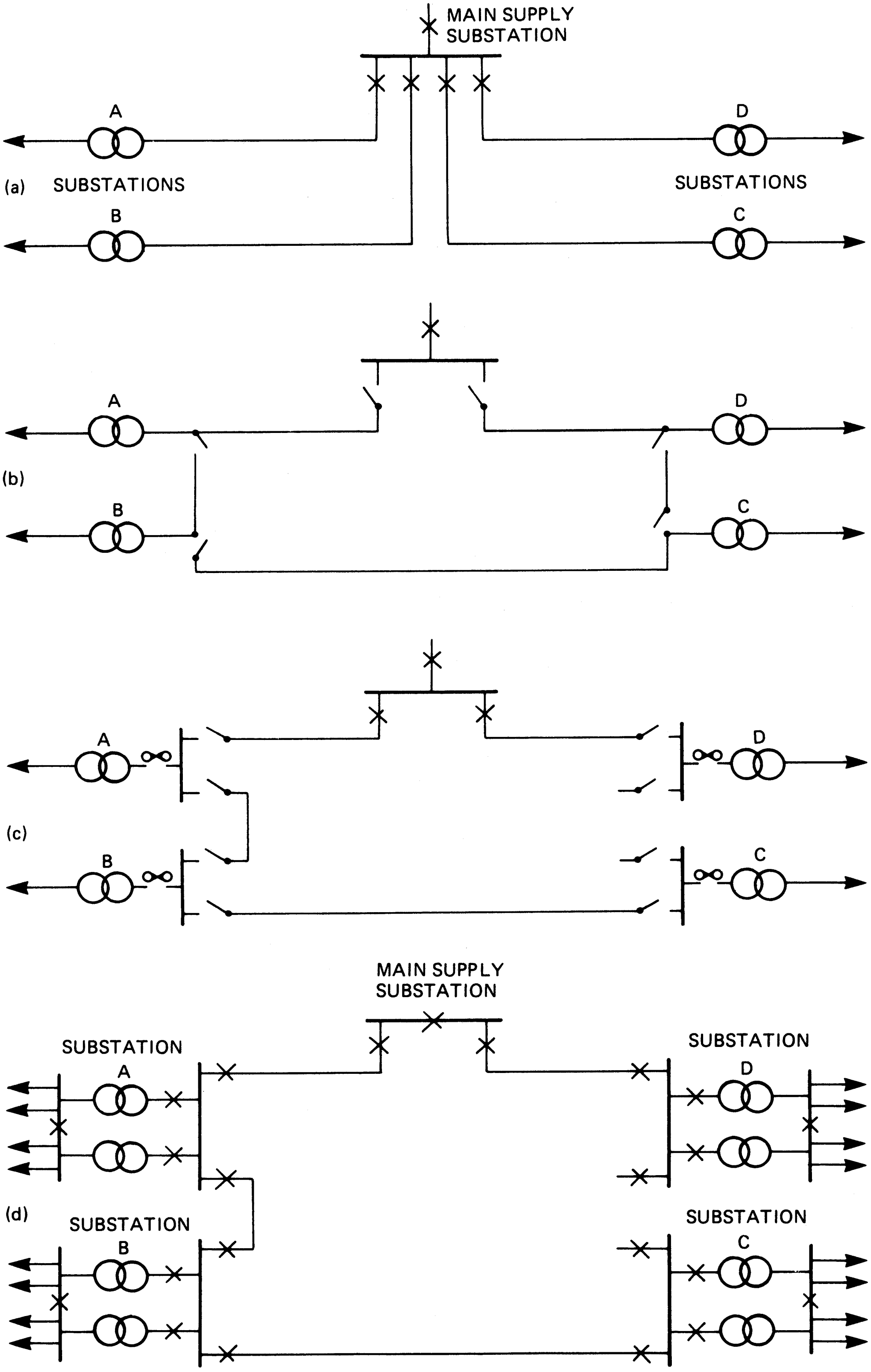

MAIN SUPPLY SUBSTATION
A
D
(a)
SUBSTATIONS
SUBSTATIONS
B
C
A
D
(b)
B
C
A
D
(c)
B
C
MAIN SUPPLY SUBSTATION
SUBSTATION
A
SUBSTATION
D
(d)
SUBSTATION
B
SUBSTATION
C

Figure 38.2 Alternative supply system designs: (a) radial feeder; (b) simple ring main arrangement; (c) distribution substation with outgoing cables controlled by switch and fuses; (d) HV installation with automatic circuit-breakers

Table 38.1 Typical per-unit reactances

(a) Synchronous generators: subtransient reactance (p.u.)

Rating (MVA)	*Voltage* (kV)	*Reactance* (p.u.)
<20	11	0.12
20–30	11	0.135
>30	22, 33	0.17, 0.20

(b) Transformers: leakage reactance (p.u.)

Rating (MVA)	*6.6/0.4 or 11/0.4* kV	*33/11* kV	*132/11* kV
0.05	0.034	—	—
0.1	0.036	—	—
0.2	0.040	—	—
0.5	0.043	—	—
1	0.045	—	—
2	0.056	0.06	0.075
5	0.07	0.075	0.09
10	—	0.09	0.10
20	—	0.10	0.10
45	—	0.10	0.11

distribution transformers, there has been a re-think on substation design for many applications. Over the past 20 years the concept of package substations was introduced whereby cast-resin transformers, low voltage switchgear, automatic power factor correction equipment and 11 kV switchgear is incorporated into a composite arrangement. It should be realised that there are savings in installation costs with possible improvements in safety as mentioned in 38.5.2. Today it is not unusual to find commercial buildings with electrical loads totalling tens of megawatts. The recent explosion of the IT industry and internet server installations has seen electrical loads in these bespoke buildings of up to 400 megawatts. The traditional approach of locating the transformers in the basement or in an outdoor compound with 415 V cabling to the l.v. switchboards may result in a very expensive distribution system. A new approach is to run 11 kV cables throughout the building to substations which incorporate the 11/0.415 kV transformer(s) as well as the 415 V switchgear. Typical of this is the combined board shown in *Figure 38.4*. In this illustration the transformer is of the dry insulated type, but increasingly cast-resin units are used. Both exhibit strong fire-resistance characteristics. This same approach can be as readily applied to industrial buildings.

To simplify the stock of spares and to ensure ready interchangeability between gear in different substations, as much standard equipment as possible should be used, even at the expense of varying from the ideal installation. In general, substations should be limited to a capacity of about 2000 or 3000 kVA, with individual transformers no larger than 1500 kVA, to allow for the use of commercial l.v. switchgear of about 43 kA rupturing capacity.

38.5.1 Low-voltage equipment

At substations the l.v. distribution gear consist usually of a circuit-breaker for each transformer with circuit-breakers, or switches and fuses for the outgoing feeders. L.v. feeders are usually supplied with ammeters and often with meters to measure the energy consumption of the feeder. In many cases maximum-demand meters are included so that this feature can be periodically monitored—a useful facility in large industrial complexes. Intelligent l.v. switchgear is now available where strategically placed voltage sensors and circuit telemetry is monitored by a PLC. Thermal monitoring of bus-bars and connections using continuous infra-red detectors are also available.

Feeders radiate to the various sectional distribution centres, where they terminate at switchboards to which smaller sub-main cables are connected, supplying power to the various departments or shops by means of other small distribution boards. The most important feeders will again be duplicated or interconnected to some extent, to safeguard the supply, so that a l.v. breakdown will result only in a temporary shutdown of one section. With duplicate cables, one may be entirely spare or both cables may share the load, provided that they are both of sufficient capacity to carry the total current independently in emergency.

38.5.2 Packaged substations

In addition to the substation designs referred to in Section 38.5 the so-called 'packaged substation' has become increasingly popular. A typical design incorporates an h.v. SF_6 switch, a cast resin transformer and fused l.v. outgoing ways. They are popular with supply authorities partly because of their compact construction, which makes them attractive for installation where space is at a premium. They also require the minimum of site erection work. They can be supplied complete with a prefabricated enclosure, often of moulded reinforced fibreglass construction, or unclad and suitable for direct installation in a building.

38.5.3 Low-voltage distribution

From the substations described earlier, l.v. feeders run to subsidiary substations or load centres. These can include multi-motor starter boards, air-conditioning control boards, lighting and heating power distribution boards, vertical or horizontal bus-bars, street lighting supplies, etc. The range of alternatives is wide and the following paragraphs outline only some of the solutions available.

Multi-motor starter boards Large motors are usually supplied with independently separate feeders. Motor starter boards will probably have the facility for intelligent motor protection. This is advisable because of the heavy fluctuating load, which might otherwise cause disturbance to other plant on the same supply. However, where a number of small or medium sized motors are in reasonable proximity, it is convenient to group the starters in a multi-motor starter panel, which often includes one or more distribution boards for lighting.

Air conditioning and ventilation control boards These are variants of the multi-motor starter board. As well as starters for the fans and pumps of the system, they also include the specialist control equipment needed for the automatic control of the air conditioning and ventilation. Similar developments are found in many industries where plant process control equipment is incorporated into combined motor starter and small-power boards. These starter boards may also incorporate equipment for Building Management systems and communications capability to connect onto

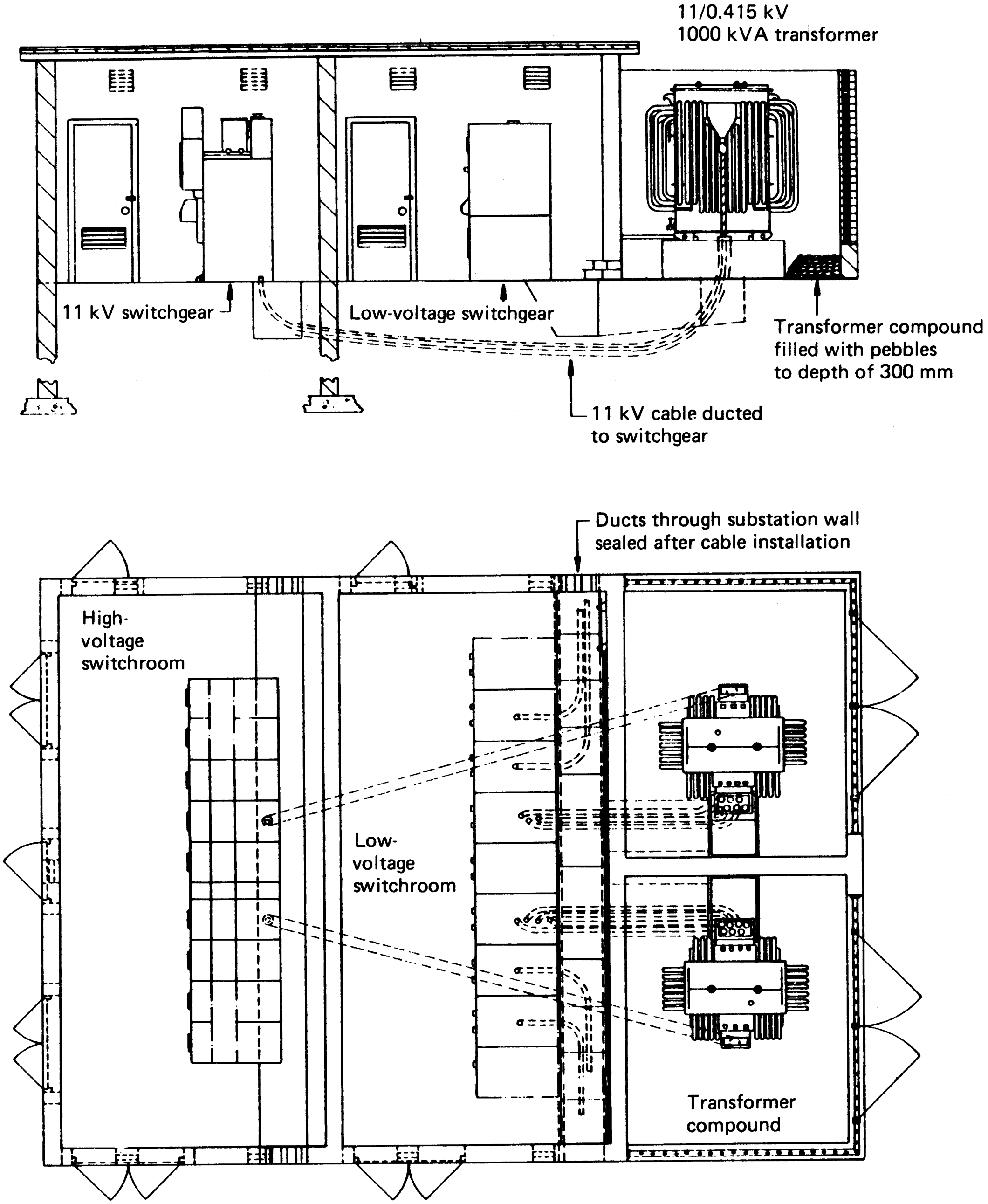

11/0.415 kV
1000 kVA transformer
11 kV switchgear
Low-voltage switchgear
Transformer compound filled with pebbles to depth of 300 mm
11 kV cable ducted to switchgear
Ducts through substation wall sealed after cable installation
High-voltage switchroom
Low-voltage switchroom
Transformer compound

Figure 38.3 Typical 11/0.415 kV substation layout

circuit breakers and MCCBs which allows diagnostic data, indication signals and energy management to take place.

Intelligent motor control can be incorporated to monitor overload, overheating and earth faults.

Bus-bar systems In general, wherever there are continuous rows of machine tools to be fed, an overhead bus-bar system provides the required degree of flexibility. Used vertically, a bus-bar system offers simple and flexible provision of power in a high-rise building. Several manufacturers market copper/aluminium bus-bars, embedded in epoxy cast resin, enclosed in steel trunking and provided with tapping points at intervals. Standard tees, bends and other accessories are available. At the tapping points it is possible to insert a tapping box, which can be safely applied or removed with the internal bus-bars live. When positions of machines have to be changed or new machines installed, it is easy to insert a tapping box at the appropriate point in the run of the bus-bar trunking. The maintenance cost is low, depreciation is minimal and there is a high recovery value if the trunking

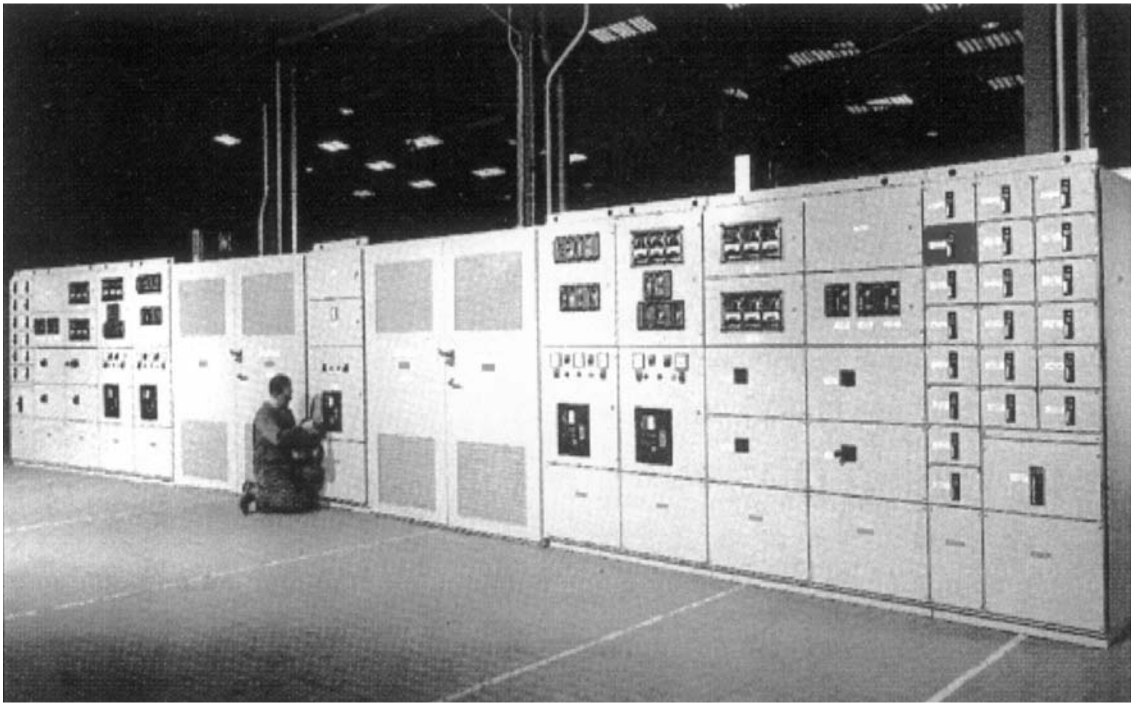


Figure 38.4 ABB 415 V switchboard, incorporating dry type 11/0.415 kV transformer, main incoming l.v. air circuit-breaker and outgoing l.v. circuits

has to be demounted and rerun. Earth continuity can be provided by an external link bolted between lengths of trunking, but for large ratings a separate earth conductor is necessary. No cables other than the bus-bars themselves may be included within the trunking. Fire barriers are required for vertical runs, and for long runs it is advisable to provide bus-bar expansion joints.

From the individual fuse switch (or circuit-breaker) units on the overhead bus-bar, cables in conduit or flexible metallic tubing are run to the machine tools. In the case of tools with more than one motor, distribution boards can be mounted at a convenient point on the equipment.

Vertical bus-bars are usually run in a vertical duct within the building. A lock-proof cupboard is formed at each floor level and the distribution boards and switches are accommodated therein, with the outgoing supplies fed to meet the lighting and power needs of the floor occupants. Where possible, it is preferable to use only one phase on each floor of the building, as this eliminates risk of voltages in excess of 240 V being encountered by personnel.

38.6 Wiring systems

The choice of wiring system for industrial, commercial and domestic installations depends on the structure, application, supply voltage, load, appearance and cost. An essential feature is a proper earthing system giving an earth continuity conducting path for protection and safety.

38.6.1 Steel conduit

Screwed steel conduit, either galvanised for corrosion resistance or black enamel, provides a potentially good earth continuity, and makes it possible to enclose cables throughout their length with strong mechanical protection. The conduit is normally of a heavy-gauge welded construction.

Skill is required to achieve a proper system of good appearance. The conduit should be erected completely before cables are drawn in; tube ends should be reamered to prevent damage to cables; draw-in boxes should be amply proportioned and accessible; the conduit should be adequately supported by saddles or clips. There should be no more than two right-angle bends (or their equivalent) between successive draw-in boxes. The whole system should be mechanically continuous throughout, including connections to switches, fittings, distribution boards, motor control gear and the metal cases of other equipment; and it must be properly earthed.

Most conduit systems in factories are run on the surface, but in offices, or where conduit would be affected by corrosive fumes, it may be sunk in concrete in the floor or in walls behind plaster or cement. The layout must be carefully preplanned, and joints must be tight, to prevent the ingress of wet concrete, cement or plaster. A draw wire should be pulled through as soon as possible after pouring or plastering, to remove any intrusive material before it hardens. The actual wiring is left until all is dry, and all wiring in one conduit must be drawn in at the same time.

38.6.2 Plastic conduit

In recent years there has been an upsurge in the use of plastic conduit. Polyvinyl chloride (PVC) components are generally unaffected by water, acids, oxidising agents, oils, aggressive soils, fungi and bacteria, and they can be buried in concrete or plaster. They provide slight resistance to heat and flame, and in this respect are inferior to steel conduit. Their mechanical strength is also much less.

Flexible heavy duty nylon conduit is also available for cable protection in installation systems. This system is self-extinguishing, free of phosphor, halogen and cadmium and is shock resistant.

They are not electrically conductive and separate earth continuity conductors have to be run, which on occasion can require a larger diameter than for a steel conduit. Expansion couplings may need to be included in long straight runs, as linear expansion of PVC gives it an extension of about 1.5 mm/m for a temperature rise of 20°C. On the other hand, the conduit can be readily cut by hacksaw and bent when gently heated. PVC conduit can be screwed using suitable dies, but jointing of lengths is commonly by a solvent welding compound. This is easier, quicker and provides an entirely watertight joint.

PVC conduit and fittings are marginally cheaper than the equivalent steel conduit components, but PVC is a by-product of the oil industry and its price comparison with steel conduit is likely to vary.

38.6.3 Trunking

Trunking is used to bunch numbers of cables or wires which follow the same route, with branches to motors, switchgear and lighting circuits by spurs in trunking or conduit. Trunking is made in various sections from 50 mm × ~~50~~ mm upward and in standard lengths. Tees, reducers, crossings, elbows and other fittings are made as shown in *Figure 38.5*.

Both steel and plastic trunking are used. With the former, low-resistance joints between lengths of trunking and fittings are essential. If the trunking is itself used as a 'protective conductor' under the requirements of the Wiring Regulation, its bonding and jointing are specified. Where conduits connect to the trunking, a clearance hole is drilled in the trunking wall and the conduit connected with a socket and male bush. Connections between trunking and switchboards and similar panels are generally effected by flanges securely bonded to the trunking and bolted to the switchgear.

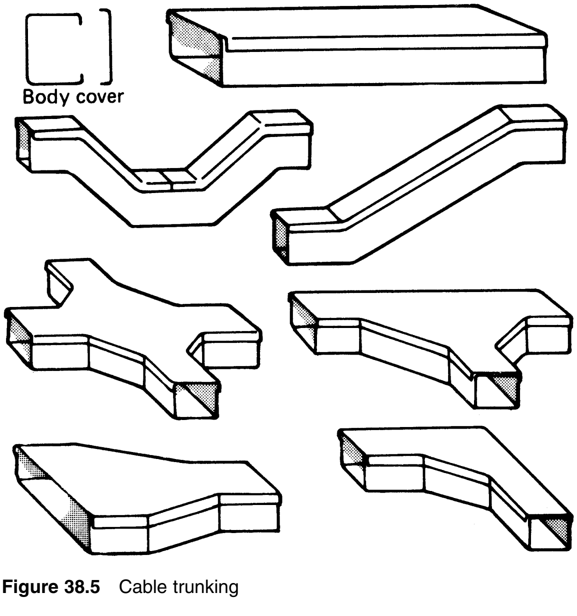


Figure 38.5 Cable trunking

On a switchboard or distribution panel in which lengths of trunking enclose the various interconnecting cables, the units should be bonded to the trunking by a copper earth tape. In the case of vertical runs of trunking, internal fire barriers must be fitted at each floor level. These barriers must also be fitted in horizontal runs where the trunking passes from one zone of fire protection to another.

Plastics trunking is made in sizes similar to those for steel trunking, small sizes being extruded and large ones formed from PVC laminated sheet. Common fittings are available, but as PVC is easily 'formed', special fittings can be 'welded' with the help of a hot-air welding gun. The characteristics of PVC trunking make it particularly suitable for fume-laden atmospheres, humid tropical conditions and salt-laden coastal air. The material has the physical characteristics mentioned in Section 38.6.2 and for the same reason requires a copper conductor to be included to provide an effective earth. Screwed or snap covers may be used.

38.6.4 Cabling

Details of insulated cables or conductors are given in Chapter 31. The following cables have applications special to the wiring of buildings:

38.6.4.1 Mineral insulated metal sheathed cables

A mineral insulated metal sheathed (MIMS) cable normally consists of high-conductivity single-strand copper conductors inside a seamless copper sheath packed with magnesium oxide powder as the insulant. As an alternative aluminium conductors and sheath are also available. The sheath acts as a robust and malleable conduit, capable of withstanding considerable mechanical abuse. It is strongly fire resistant, and when covered overall with an LSF sheathing can withstand chemical corrosion.

Because magnesium oxide is hygroscopic, the exposed cable ends must be sealed. The sealants are suitable for operating temperatures up to 70°C, but special oil-resistant and high-temperature seals are available for special applications.

Although dearer than the equivalent steel conduit installation, there are applications where MIMS cabling is superior and preferred. For example, some fire authorities insist on MIMS cables for all fire protection, detection and alarm circuits.

38.6.4.2 PVC/PVC sheathed cables

For domestic installations, PVC insulated and sheathed cables are often used, run on the surface, or more usually buried in plaster. Since there is at least a perceived danger of buried cables being penetrated by a nail or screw driven into the wall, the IEE Regulations have since 1987 been amended to lay down strict rules as to the positioning of buried cables. For example, cable connected to any point or accessory on the wall must be run horizontally or vertically to the point or accessory; in this regard, the consumer unit is treated as an accessory.

In small industrial installations where cost is a prime consideration, it may sometimes be acceptable to run PVC sheathed cabling fixed by cleats spaced about 1 m apart. Cables should preferably not be concealed and must be protected where they run through floors, and also on walls if fixed at a level below about 1.6 m from the floor.

38.7 Lighting and small power

Whatever type of electrical installation is involved—be it for a factory, a public or commercial building, or domestic premises—supplies for electric lighting and small power will be needed.

38.7.1 Lighting circuits

There are two main ways of running lighting circuits, 'loop-in' and 'junction box'. Both can be used in the same installation. In the loop-in system (*a*) in *Figure 38.6* the terminals for joining the cable ends form part of the ceiling rose. The junction-box system (*b*) is used when the lighting fittings have no loop-in terminals, or to save cable when the lamp and the switch are far apart. The connections are similar to those in the loop-in system. The new 3 and 4 pin plug in lighting system now offers the installer greater flexibility and ease of installation especially with dimmable fluorescent lighting and emergency lighting circuits.

38.7.1.1 Two-way switching

The two-way switch is a single-pole changeover switch. When interconnected in pairs, two-way switches provide control from two positions and are therefore installed on landings and staircases, in long halls and in any room with two doors. The arrangement is as shown diagrammatically in *Figure 38.7*(*a*).

38.7.1.2 Intermediate switches

Where there are long halls, corridors or passageways, with several doors, it may be convenient to introduce additional switching positions. The 'intermediate' switch enables any number of additional switching points to be introduced to the two-way switch circuit, as in *Figure 38.7*(*b*).

38.7.1.3 Master control switching

A master control switch is sometimes provided to give overall control to a number of lamps which are, in addition, independently switched, as in *Figure 38.7*(*c*). A typical installation would be an hotel bedroom or suite, where a double-pole switch at the entrance to the room will switch off all lights on the occupants leaving the room.

In new office installations, increasing use is being made of master control switching, usually by way of sophisticated relay systems which allow groups of lights (sometimes whole floors) to be controlled by timer, or by measuring the ambient light conditions and switching lights 'off' or 'on' to suit.

More modern lighting controls will consist of internal and external illumination sensors connected to micro-processor controls which monitor and control luminaires, from the window line, to achieve energy savings and sophisticated lighting control via high frequency dimming for T5 fluorescent lamps.

The new generation of micro-processor based control units incorporate LCD touch panels to enable activation of pre-set lighting schemes together with adjustment of blinds, air conditioning and sound systems.

38.7.2 Small power circuits

The socket outlet is a safe and convenient means by which free-standing or portable apparatus can be connected to an electric supply. It is false economy to limit the number of socket outlets installed. Sufficient should be provided to match the predicted needs of the consumer, and should be located adjacent to the most convenient point of use of the apparatus.

Both radial and ring final sub-circuits may be used to connect socket outlets, although in the UK the 30 A ring-main system using BS 1363:1984 fused plugs has held sway for many years. With this system, the current-carrying and earth-continuity conductors are in the form of loops, both ends of which are connected to a single way in the distribution board. The conductors pass unbroken through socket outlets or junction boxes (or must be joined in an approved manner). The ring may feed an unlimited number of socket outlets or fixed appliances but the floor area covered shall not exceed 100 m^2. In domestic premises particular consideration should be given to the loading in the kitchen, which may require an additional circuit.

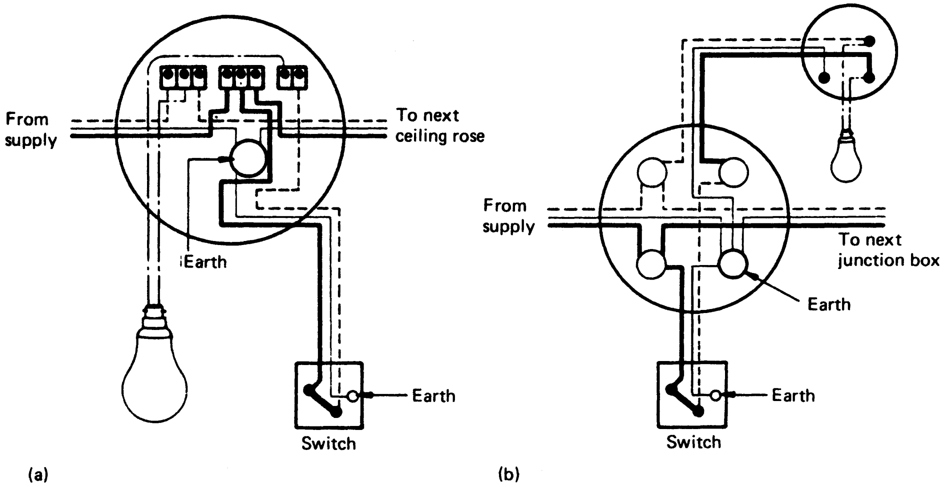


Figure 38.6 Ceiling rose loop-in and junction-box wiring: (a) loop-in system; (b) junction-box system. ▬▬, Red, always live; - - - - -, black, neutral except where returning from a switch; ——, earth; –·–·–, flex to lamp

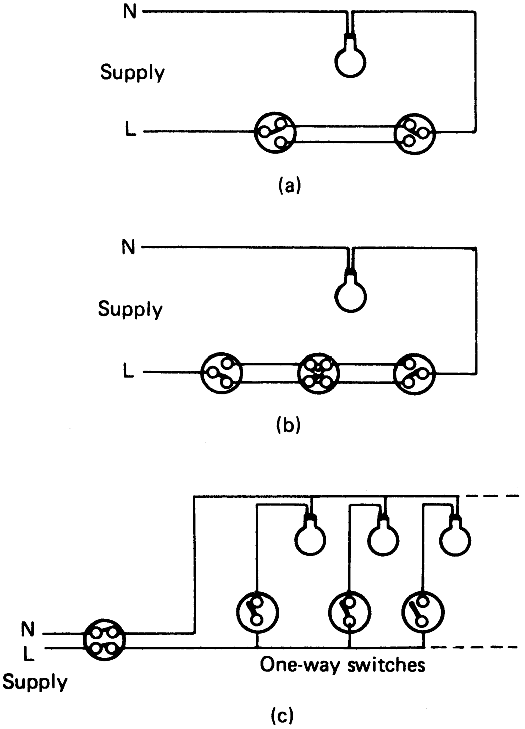


Figure 38.7 Lamp switching: (a) two-way switching; (b) two-way and intermediate switching; (c) master control switching

In the UK an unlimited number of fused connection units may be installed in a ring circuit, but the number of *non*-fused connection units must not exceed the number of socket outlets, or one permanently connected item of equipment. A typical ring circuit is shown in *Figure 38.8*.

Portable apparatus is now normally supplied with moulded plugs. At the same time more and more consumer units are being equipped with miniature circuit-breakers in place of fuses. It is also obvious that 13 A socket outlets to BS 1363:1984: Pt 1:1995 are bigger, more obtrusive and more expensive than the Continental or American counterparts. Is it perhaps not time that we give some thought in the UK to a reversion to radial circuits and to non-fused plugs?

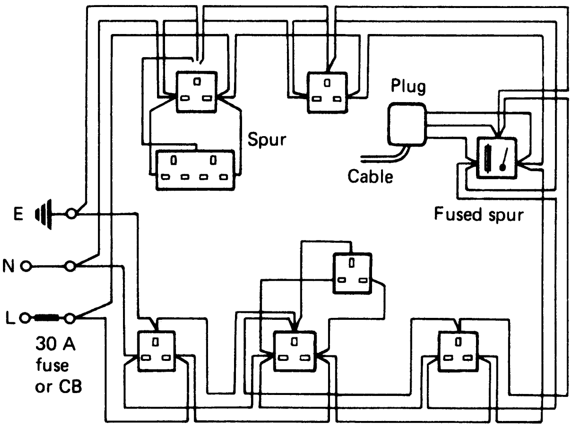


Figure 38.8 Ring circuit for 13 A outputs. CB, circuit-breaker

38.8 Floor trunking

Floor trunking is used mainly as a flexible means of providing small-power and telephone facilities throughout an open floor area. The basic methods are underfloor, open-top and cavity-floor systems. All must have continuity between metal sections and be adequately earthed.

38.8.1 Underfloor trunking

The several proprietary brands available consist basically of a closed trunking with one, two or three compartments, laid in a grid pattern on the floor slab. Specially designed intersection boxes take care of the crossovers in the lines of trunking, and fixed outlet boxes from the trunking are installed to meet all possible future needs. When the trunking is screeded over, the intersection and outlet box lids are flush with the final floor level.

38.8.2 Open-top trunking

Open-top trunking is a more recent concept of floor trunking. As with the underfloor trunking it is laid on the floor slab before screeding, but in this case the heavy-duty cover plates of the trunking are flush with the final floor finish, which makes it easy for outlet boxes to be added anywhere along the length of the trunking as changing needs demand, hence giving a greater degree of flexibility.

38.8.3 Cavity-floor systems/dado trunking

This flexible system comprises wood or metal square plates resting on support jacks. The services are run in conduit or trunking on the surface of the floor slab under the floor plates. Flush surface boxes can be installed into holes cut almost anywhere in the suspended floor, and wiring to power and telephone outlets is routed via conduits or

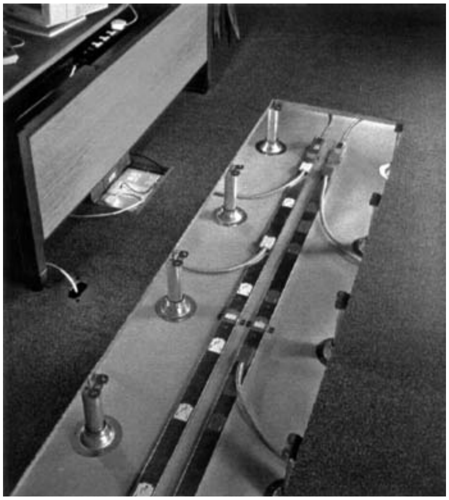


Figure 38.9 ‘ALPHATRAK’ cavity-floor trunking system. (Courtesy of Power Plan Systems Ltd)

trunking to central junction boxes. (A typical installation is shown in *Figure 38.9*.) There are now slim shallow products on the market with desk modules to connect to user power, voice and data appliances. Underfloor power distribution, available in 63 A, 40 A and 25 A is suitable for power to PC circuits and is available with clean and standard earthing systems.

Dado wall mounted cable management trunking consists of a series of single, twin and multi-compartment systems in various sizes produced in high impact UPVC. The trunking will accept wiring accessories and a bus-bar variant. The cable compartments are designed for power, communications and data outlets.

Data tray cable basket systems are manufactured from 50 mm × ~~50~~ mm wire mesh grid with the facility to provide take off or support positions throughout the length of the installation. The installed matrix system is ideal for most power, data, low voltage signal and some pneumatic systems.

38.9 Stand-by and emergency supplies

Provision for supply failure has sometimes to be considered, most commonly for emergency lighting in the form of (i) escape lighting (to ensure the provision of adequate illumination of emergency escape routes), and (ii) stand-by lighting (to enable normal activities to continue during the period of mains failure). The two most common systems of emergency lighting are: *maintained*, where normal lighting is fully maintained by automatic switching to an alternative source should mains failure occur; and *non-maintained*, where the emergency lamps are energised only when the normal lighting has failed.

Both systems are used extensively. The maintained system is essential in certain public buildings, and is the system preferred for 'stand-by' lighting.

Standards of emergency lighting are derived from the Fire Precautions Act 1971 and Fire Precautions Regulations Updates 1997. The CIBSE Lighting Code and BS 5266: Part 1 7:1999 also provides some useful guidance.

In the past, most emergency lighting systems were supplied from a central battery and charger unit, but self-contained emergency lighting luminaires have become readily available for both maintained and non-maintained systems. Each unit comprises a lamp, battery, charger and control equipment, and is wired conventionally into the normal lighting system. Usually a neon or diode indicator is incorporated to provide a visual indication that the battery is under charge. In general, where fewer than 20 fittings are involved, the self-contained luminaires are likely to give the more cost-effective solution. To be effective it is essential that emergency lighting is regularly tested and a maintenance procedure set in place. This can be done by semi-automatic and fully automatic testing.

38.10 Special buildings

38.10.1 High-rise buildings

With large high-rise buildings, the heavy electrical and mechanical services plant is usually located in the basement and/or basement and part-way up the building. However, there are occasions when air-conditioning chillers, stand-by generating plant and even boilers are located on the roof, and this can influence the positioning of the main electrical switchboards. Nevertheless, since the main power supply enters at basement level, it is usual to locate the main transformer and the primary substation in the basement. Basement-mounted equipment will normally provide supply for 10–12 floors, for taller buildings additional substations are required every 10–15 floors.

When equipment is installed in buildings to be used as offices or apartments, particular attention must be given to preventing unnecessary vibration. Rotating plant should be provided with anti-vibration mountings. In a new building the architect and the structural engineer should co-operate in providing a 'floating' floor to support equipment liable to vibration.

38.10.2 Public buildings

In installations for large buildings and special establishments open to the public (such as theatres, cinemas, etc.) special precautions must be taken in the layout of the electrical equipment. National standards and any regulations formed by the local authorities must be strictly observed in planning the scheme. These deal more especially with the requirements necessary to avoid danger to the public from fire, explosion, electric shock, etc. and are also concerned with the question of dual supply or emergency lighting in the event of a breakdown of the power mains.

38.10.3 Domestic premises

Domestic premises are commonly connected to the street supply by a one-phase l.v. PVC cable, often with aluminium conductors. The incoming supply is rated at between 60 and 100 A and is capable of supplying the lighting, heating and cooking demands of the house or flat.

The incoming cable terminates in a service cut-out and connects via the house meter to a consumer's unit. There is an increasing desire with new housing to locate the meter in a position where the meter reader can read it without access to the house. The consumer unit is equipped with an incoming supply switch and outgoing ways using high rupturing capacity (HRC) fuses or miniature circuit-breakers. One or more 30 A ring mains for socket outlets will be provided. Immersion heaters fitted to hot-water cylinders in excess of 15 litre capacity should be supplied from an independent circuit. Electric cookers are fed by a final sub-circuit through a cooker control unit which may also include a 13 A socket outlet.

When the ratings of the protective devices or the quality of the earthing is such that an earth fault would not be cleared within the prescribed time, a residual-current device (r.c.d) should be fitted. It could cover a whole domestic installation, but there can then be a risk of spurious tripping caused by the normal leakage currents of healthy equipment.

An alternative would be to use one r.c.d on a socket outlet circuit. Certainly, if a socket outlet is used to connect garden tools, this should be protected by an r.c.d. The r.c.d can be incorporated in the consumer unit; a unit comprising one socket outlet and an r.c.d unit can be contained in a twin-socket box or a plug in unit.

38.11 Low-voltage switchgear and protection

Heavy switchgear is dealt with in Chapter 34. The term 'industrial switchgear' is generally applied to that used for voltages not exceeding 600 V and controlling power from the l.v. side of a transformer to distribution boards which may include automatic circuit-breakers, or to fuse switches and contactors for motor-starting equipment. Consideration must be given to the prospective fault current at the point of installation, particularly as it may directly affect personnel, often unskilled and unaware of the danger should the apparatus fail to perform its function.

Today oil circuit-breakers are out of favour for operation at 415 V. Air circuit-breakers are still preferred by many engineers to moulded-case circuit-breakers, and for the higher current and fault levels they are essential.

38.11.1 Air circuit-breakers

For the larger industrial and commercial consumer the principal switchboard, and often subsidiary switchboards, will normally comprise a factory built cubicle assembly, incorporating conventional air circuit-breakers with ratings up to 4000 A. Air circuit-breakers now normally incorporate microprocessor electronic protection and management controls. Composite boards with a mix of air circuit-breakers, moulded-case breakers, fuse switches and contactor gear are also common. Segregation of components are via various forms of separation for the bus-bars, functional units and incoming or outgoing terminations. The forms of separation range from Form 1 to Form 4b, type 7.

38.11.1.1 Forms of internal separation

The Standard

Form 1 No Separation

Form 2 Separation of bus-bars from functional units

Form 3a Separation of bus-bars from functional units
Separation of functional units from one another
Separation of the outgoing terminals from the functional units
But
No separation of terminals from one another
No separation of terminals from the bus-bars

Form 3b Separation of bus-bars from functional units
Separation of functional units from one another
Separation of the outgoing terminals from the functional units
Separation of terminals from the bus-bars
But
No separation of terminals from one another

Form 4 Separation of bus-bars from functional units
Separation of functional units from one another
Separation of the outgoing terminals from the functional units
Separation of terminals from one another

Form 4 – Type 1
Separation of bus-bars by insulated coverings, e.g. sleeving, wrapping or coatings
Terminals in same compartment as functional unit

Form 4 – Type 2
Separation of bus-bars by metallic or non-metallic or non-metallic rigid barriers or partitions
Cables glanded elsewhere
Terminals in same compartment as functional unit

Form 4 – Type 3
All separation by metallic or non-metallic rigid barriers or partitions
The terminals for each functional unit have their own integral glanding facility
Terminals in same compartment as functional unit

Form 4 – Type 4
Separation of bus-bars by insulated coverings, e.g. sleeving, wrapping or coatings
Cables glanded elsewhere
Terminals *not* in same compartment as functional unit

Form 4 – Type 5
Separation of bus-bars by metallic or non-metallic rigid barriers or partitions
The terminals for each functional unit are separated by insulated coverings
Cables glanded in common cabling chamber
Terminals *not* in same compartment as functional unit

Form 4 – Type 6
All separation by metallic or non-metallic rigid barriers or partitions
Cables glanded in *common* cabling chamber
Terminals *not* in same compartment as functional unit

Form 4 – Type 7
All separation by metallic or non-metallic rigid barriers or partitions
The terminals for each functional unit have their own individual glanding facility
Terminals *not* in same compartment as functional unit

Where the switchboard manufacturer is not also the manufacturer of the air circuit-breakers, the customer should seriously consider the use of 'cassette' breakers. With this design each breaker is delivered to the switchboard builder complete in a sturdy 'box', i.e. the cassette. The breaker design will have been type tested in its cassette. Hence the user can be satisfied that the breaker will retain its certified test rating when fitted in a custom-built switchboard. This cannot be assumed with breakers offered to the switchboard builder in 'open cradle' form. At the same time, the switchboard builder can be absolved from having to carry out separate type tests to guarantee the requisite ASTA rating.

38.11.2 Moulded-case circuit-breakers

Of particular interest to the installation engineer are moulded-case breakers which find wide use in commercial and industrial applications. They are primarily intended for applications such as protecting main feeder cables, or acting as main circuit-breakers controlling large banks of other circuit-breakers. As such, their major function is to provide back-up protection to the sub-circuit protective devices. Overcurrent protection is a relatively secondary function. The moulded-case breaker is of more robust construction than the miniature circuit-breaker and can often be provided with auxiliary items such as extended operating handles for assembly in multi-switch cubicles, mechanical interlocking with adjacent breakers, motor-operated mechanisms and shunt trip coils. Current ratings up to 1250 A and short-circuit capacities between 25 kA to 150 kA are readily available.

38.11.3 Miniature circuit-breakers

Miniature circuit-breakers usually embody both overload and short-circuit current tripping devices, the overload usually by a thermal device, the short circuit by a magnetic one. A trip-free mechanism is incorporated so that the contacts cannot be held closed against a fault, and the thermal element prevents continuous rapid reclosing of a circuit when a fault persists.

The primary function of a miniature circuit-breaker is to protect an installation or appliance against sustained overloading and short-circuit faults, but it will also give protection against earth faults provided that the earth fault loop impedance is low enough. As a secondary function the miniature circuit-breaker can be used as an isolating switch or for local control switching. Today they are widely used in place of fuses in domestic installations where householders find them more convenient.

A circuit-breaker used to protect a sub-circuit should have a current rating matched to its load. The circuit-breaker rating must not exceed that of the cable; also, the current causing effective operation of the breaker must not exceed 1.45 times the lowest current-carrying capacity of any of the conductors in the circuit. (This requirement of WR16 also applies to other protective devices such as fuses.)

Miniature circuit-breakers are available with current ratings from about 0.5 to 100 A, with a fault capacity of 16 kA although values of 9/10 kA are more normal.

38.11.4 Fuses

The first edition of the IEE Wiring Regulations in 1881 contained a note reading: 'The fuse is the very essence of safety'. Today, over 120 years later, fuses have to face competition from other devices, but in their many modern guises, though very different from their counterparts of the 19th century, they still play an outstanding role in the protection of electrical circuits.

The HRC fuse is the common type for general power system use. It was first introduced in the 1920s, and in the intervening years its design and performance have steadily improved, until today it has an unrivalled reputation as a short-circuit protective device. Its breaking capacity and energy and current-limiting ability are superior to those of any other protective device.

The design of the HRC fuse is described in Section 34.2.2.2. The most important design feature is the fusible link, usually of silver or silver/tin alloy, but often with the addition of other metal inserts. This enables the designer to achieve the desired time/current and other characteristics, striking a balance between conflicting requirements of minimum fusing current, low let through I^2t and peak current on short circuit.

BS 88 was revised and up-dated in 1988, and is now identical to IEC 269 and BS EN 60269:1:1999, 'Low voltage fuses'. It makes significant progress in the international standardisation of LV fuses. The aim has been to achieve safe electrical installations, particularly from the viewpoint of avoidance of electrical shock, both in normal and fault conditions, and the provision of overcurrent protection to the electrical cables forming the fixed installation.

Some of the fuse link types used in some European countries have only partial range breaking capacity, i.e. they interrupt short-circuit fault currents, but are unable to interrupt overload currents safely. To distinguish these types from the much more widely used general-purpose fuse links, the concept of 'utilisation category' has been introduced in IEC 269, and in the revised BS 88:1988.

Each of the classes is identified by a two-letter code. The first letter indicates the breaking range of the fuse link:

g A full-range breaking capacity.
a A partial-range breaking capacity.

The second letter indicates the utilisation category:

G A fuse link for general application, including the protection of motor circuits.
M a fuse link for the protection of motor circuits.

The standards combine these letters to recognise three classes: i.e. gG, gM and aM.

Although harmonisation of the fuse dimensions as between the German, French, British and other systems has not yet been achieved, effectively there has been wide agreement on electrical parameters. This is a major achievement because it means that all general-purpose fuses, from whatever source, can be applied in the same manner. For example, for all fuses above 16 A rating, the fusing current shall not be less than $1.25I_n$, where I_n is the fuse rating.

Similarly, compliance with the specification results in a discrimination ratio between major and minor fuses of 1.6:1 based on pre-arcing characteristics. When fault current in an a.c. circuit starts to rise, the fuse element heats and begins to melt. This takes a finite *pre-arcing* time. As the element ruptures arcing occurs and shortly afterwards the break is complete, the arcing ceases and the current drops to zero. This is the *arcing* time. The sum of these two is the *total operating* time. In the past it was accepted that discrimination between major and minor fuses should be based on the vast majority of installation, under fault conditions with total operating times, but it is now argued that in modern fuses the arcing I^2t can be ignored. Even in three-phase circuits with relatively high power factors and fault levels up to 80 kA at 415 V, the 1.6:1 ratio is found to be valid, and this clearly provides economic benefits in any modern installation.

38.11.5 Prospective fault current

The prospective fault current at any point in a system is the current that would flow if there were a solid short-circuit at that point, and it represents the maximum current the protective device would have to interrupt. In practice, the actual current is usually much less than the prospective fault current because seldom does a solid three-phase short-circuit occur at the terminals of the protection device, and even short lengths of intervening cable reduce the fault current considerably. Nevertheless, knowledge of prospective fault current is necessary to ensure that the circuit-breakers, fuses, etc. in the system can deal with the fault current. If the fault level at the point is in excess of the rupturing capacity of the device, damage can be done to the installation and to the protective device itself, because of the amount of energy that passes before the current is completely interrupted. In such cases the protective device must be backed up by another breaker, or by fuses capable of interrupting the fault and before the 'let-through' energy has built up. The total 'let-through' energy is identified by I^2t, where I is the fault current and t is the time for which the current flows before complete interruption.

38.11.6 Discrimination

A major consideration in designing a distribution system is the problem of ensuring effective discrimination. Ideally, the protective devices should be so graded that, when a fault occurs, only the device nearest the fault operates. The other devices should remain intact and continue supplying the healthy circuits. It is usually possible to assess, with fair accuracy, how effective the discrimination will be between combinations of circuit-breakers and fuses, but there are several practical considerations to be taken into account: (i) manufacturing tolerances in components, and (ii) operating conditions such that the devices do not conform to the published data. Over a period of time a fuse may be subjected to high currents causing it to 'age', and this can have an adverse effect on its behaviour.

In any medium to large installation, the supply is subdivided for distribution, and hence there will be protective devices (sometimes several) between the final device and its source of supply. At the other end of the scale, as in a domestic ring circuit, there may be fuses further down the line backed by miniature circuit breakers. The primary objective is to ensure that the device in question will deal with faults up to its breaking capacity, the back up device taking over above this level. (These other protective devices may be circuit-breakers or fuses.) This, theoretically, is the ideal condition, but in practice—to allow for the discrepancies mentioned above—it is usually advisable to aim for back-up protection taking over at a fault-current level not exceeding about 70% of the circuit-breaker breaking capacity. The need to comply with this requirement will automatically set the upper limit to the current rating of the back-up device. The lower limit of current rating is usually set by the need to avoid loss of discrimination.

Where miniature circuit-breakers in sub-circuit distribution boards are concerned, the miniature circuit-breaker should be the device to operate for all fault levels up to 1000 A, i.e. the great majority of sub-circuit fault currents. If the zone in which the back-up fuse tends to take over from the circuit-breaker is below 1000 A, discrimination troubles may be experienced, but provided that it is not below about 700–800 A, operating experience indicates that little trouble will be met on this score. If the zone is above 1300 A, the installation should be relatively free, in practice, from any form of discrimination trouble.

If fuse time/current characteristics are available, the probable position of the take-over zone can be addressed quite readily. The total time taken by a miniature circuit-breaker to clear a short-circuit fault is usually about 10 ms. If, therefore, the prearcing or melting time of the back-up fuse—at a given level of fault current—is less than 10 ms, it is reasonable to assume that at that level of fault current a m.c.b. will not consistently discriminate against the back-up fuse. Thus, the current at which the fuse prearcing time/current characteristic crosses the 10 ms line will give a working guide to the position of the take-over zone.

The family of time/current characteristics shown in *Figure 38.10* is for a typical range of quick-acting HRC fuses. From this diagram it will be seen that the 63 A characteristic crosses the 10 ms line at about 1000 A. From this it is reasonable to assume that the take-over zone, i.e. the range of current over which there is a possibility of loss of discrimination, will probably extend from about 800 to 1000 A. For fault currents less than 800 A there should be no trouble with the back-up fuse. With an 80 A fuse the zone will probably extend from about 1100 to 2500 A. Hence, if a quick-acting HRC fuse is used for back-up, ideally it should not be rated at less than 80 A, and preferably at not less than 100 A.

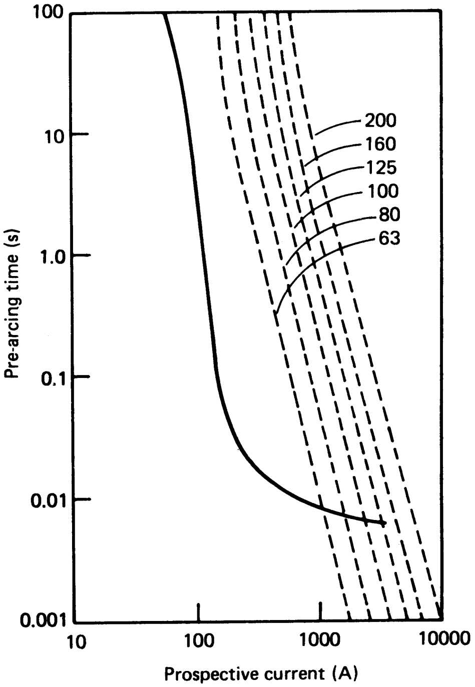


Figure 38.10 Time/current characteristics for HRC fuses and MCSs. HRC fuses - - - - -; HRC fuse; 30 A MCBs ——

The use of readily available computer programs to undertake a full discrimination study would enable the engineer to add, select, change and delete a device with reference to fault levels, in rush conditions and measuring the grading margin.

38.11.7 Motor control gear

A motor starter may be an individual item of equipment, or one of several in a multi-motor starter panel. In general, the supply authority lays down the limits of starting current permissible. Small motors are normally switched direct-on-line, but larger machines may require starting methods that limit the starting current and power factor.

38.11.7.1 Isolating switches

The Electricity at Work Regulations and the IEE Regulations WR16 require that every motor be provided with a means of disconnecting the motor and all its auxiliary equipment, including any automatic circuit-breaker. The isolating switch should be mounted nearby and is often incorporated in the same case as the starter itself. The switch must be capable of making and breaking the stalled current of the motor and under normal conditions it remains closed. The switch often incorporates auxiliary contacts which (among other functions) isolate the control circuitry in the starter, to ensure that the starter and its control circuit are isolated before the door is opened and maintenance work begun. The regulations do allow an isolating device to be located remote from the motor, but under these circumstances provision must be made so that the means of isolation can be secured in the open position.

38.11.7.2 Contactor starter

For the smallest motors, a manual mechanical switch is acceptable; however, the Regulations require that means must be provided to prevent automatic restarting after stoppage (due to drop in voltage or failure of supply) where unexpected restarting might cause danger. Where such a device is provided, it is logical that it should also be used to operate the switching contacts, and this is the reason for the popularity of electromagnetically operated contactor starters. Many other advantages, such as easy control by relays, limit switches and other light current devices, follow the adoption of contactor starters.

There are two basic types of contactor—the electrically held-in and the latched-in type. With both, the contactor should close when the voltage is as low as 85% of nominal. With the electrically held-in type, should there be a transient failure of the supply, the contactor will drop out. With certain types of drive this can be embarrassing, and a latched-in contactor may be preferred. With this type, the 'close' coil energises the contactor, which then latches in, and the 'close' coil is de-energised. The contactor is opened by energising an 'open' coil or by the operation of a protective device.

38.11.7.3 Motor protection

Motor starters are usually fitted with a trip device which deals with overcurrents from just above normal running current of the motor to the stall current. The aim should be for the device to match the characteristics of the motor so that full advantage may be taken of any overload capacity. Equally, the trip device must open the starter contactor before there is any danger of permanent damage to the motor.

Contactors are not normally designed to cope with the clearance of short-circuit conditions, and it is therefore usual for the contactor to be backed up by HRC fuses or by circuit-breaker.

The arrival on the scene of very compact motor starters and the need to provide proper back-up protection to them has posed a problem. BS EN 60947-4-1 (1992) (previously BS 4941) 'Motor starters', describes three types of co-ordination, the most onerous condition (type C) requiring that under fault conditions there shall be no damage to the starter or to the overload relay. The usual back-up device will be the HRC fuse. It is important that the user check with the manufacturer's catalogue to ensure that the correct fuse is used to secure this co-ordination.

38.11.7.4 Remote control of motors

When the starter is mounted remote from its motor, care must be taken in positioning the control push-buttons. Accidents can occur when the operator cannot view the motor from the control position; it may then be necessary to mount pilot lights adjacent to the motor to indicate to local personnel that it is about to start. In some cases, particularly with large complex machines, it may be necessary to give audible warning that the equipment is about to be started.

When an emergency stop button is located adjacent to the motor, it should be of the lock-off variety, so that when operated, not only does the motor come to a standstill, but also it cannot restart until the lock-off push-button has been released.

38.11.7.5 Braking and stopping machines

When it is necessary to stop a machine very quickly, some form of braking must be employed. An ordinary friction brake, the simplest device, should be held off against a spring or counterweight, so that it is applied automatically if power should fail. Electrodynamic braking is ineffective at slow speeds: it is therefore a useful addition rather than an alternative to a friction brake. Torque reversal (plugging) is very effective, but in many cases a reverse rotation cut-out will be required.

Care must be taken with the connection to the brake. Some motors tend to generate during run-down, and the brake circuit must therefore be interrupted positively to ensure that the friction brake is applied. A particular danger arises when rectified current is used for the brake solenoid on large machines. It may be found that the rectifier acts as a low-resistance shunt across the brake coil and prevents the magnetic flux collapsing quickly, making the brake action sluggish. In this case it is necessary for the main contactor to have separate contacts in the brake-magnet circuit.

When heavy machines are controlled by plugging, a problem will arise if the power supply should fail, for this would leave the machine uncontrolled, unless there is back-up emergency braking.

38.11.7.6 Limit switches

Limit switches are generally used to initiate a control sequence at the correct point in the mechanical duty cycle of a machine. For example, they may signify the end of a crane travel and initiate the stop sequence. Limit switches must be of robust construction—well protected against the ingress of fluids, dust or dirt. They are frequently used on machines expected to have a long life, and it is essential that the limit switches be equally generously designed. The current-carrying capacity of the electrical contacts should also be generously rated, since they may well have to handle in-rush current.

38.11.7.7 Inching control

When frequent inching is required, it is preferable that a separate contactor be employed. However, the motor circuit should still be taken through the overload protection device, to ensure that the motor is not allowed to overheat. At the same time the use of the separate contactor does restrict wear on the contacts of the main contactor.

38.12 Transformers

Most transformers used in industrial and commercial applications are either oil-filled or dry cast resin, natural-cooled and suited for mounting out of doors. The transformer is contained in a tank, plain or with cooling tubes or fins. Tanks may be equipped with small wheels, but more usually have a flat bottom for mounting on a concrete plinth. The modern transformer is reliable and tolerant of its location. In oil filled transformers, fires are infrequent but there is still some reservation with regard to siting oil-filled units indoors.

For such locations transformers with the mineral oil replaced by a non-toxic, biodegradable dielectric fluid, such as MIDEL, may be used. The use of polychlorinated biphenyls (PCBs) (Askerels) is now totally banned because of their toxicity an their perceived carcinogenic nature.

With the increasing need for transformers to be located indoors there has been a growth in demand for dry-type units, initially these were what used to be called class C type: i.e. with windings vacuum impregnated with silicone varnish, and with insulation and varnish selected to avoid the propagation of fire, and the emission of smoke and toxic fumes.

More recently, the demand has been for cast-resin transformers where each phase of the winding is encapsulated in epoxy resin. These units are usually supplied in ventilated sheet steel enclosures with lift-off access panels or lockable cubicle doors. Where access to a site is particularly difficult, these transformers can be dismantled before delivery and re-assembled *in situ,* but obviously this is not a favoured activity.

As an optional item axial flow fans can be supplied to boost the output of the cast resin transformer by up to 25%. This feature can be useful, particularly where one of two units operating in parallel has to be taken out of service for maintenance. The boosted output of the remaining unit enables it to take up the additional load, probably without affecting the total load normally being supplied from the two units. Using the fans for regular day-to-day use is not recommended.

Whatever type is supplied, and wherever located, heat loss from a transformer makes ventilation important. Even at an efficiency of 99% there can still be many kilowatts of heat to be dissipated in a confined space!

It is usual for a transformer to be delivered ready for service, and it is recommended that the installation engineer does not remove the lid unless there is evidence of some abnormality. If, however, there is doubt, the check testing below is carried out.

(1) *Voltage ratio*—apply a low-voltage three-phase supply to the h.v. winding and measure the l.v. output voltage.
(2) *Phase grouping*—connect one pole of the h.v. and l.v. windings together (*Figure 38.11*): then by applying a three-phase l.v. supply to the h.v. windings, voltages can be measured and the grouping determined from the results.

It will be readily seen that in both the above tests it is imperative that the test l.v. supply be connected to the h.v. winding.

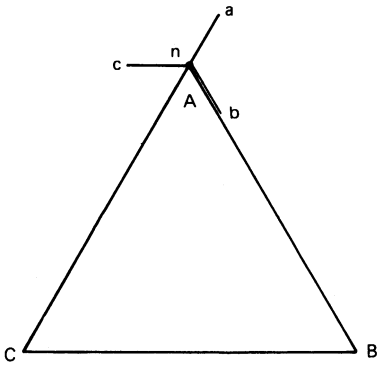


Figure 38.11 Checking the phase group of a transformer

It is seldom that a small or medium-sized transformer is shipped without oil; but if this should happen, or if the transformer has been in store, it may be necessary to dry out the windings before commissioning the unit. If carried out at the manufacturer's works, this would be done in a drying chamber, but on site it should be done by circulating currents in the transformer windings. It is then preferable that the oil be circulated through a filter plant at the same time. During the drying-out test, regular readings of the insulation resistance of the winding should be taken and checks made on the electric strength of the oil. The drying-out process is likely to take several days.

Where large transformers are concerned (say above 10 MVA), it would be foolish to proceed with any precommissioning or commissioning tests without enlisting the aid of the manufacturers.

38.12.1 Installation

Before connecting the transformer to the supply, several precautions are necessary. First, the tank should be efficiently earthed. In the case of a transformer with off-load taps, the correct tappings should be chosen.

With self-cooled units fitted with radiators, the valves at the top and bottom of the headers should be open. In the case of artificially cooled units, all the valves must be open and the correct quantities of oil and water should be circulating. Where the transformer is of the type employing forced oil circulation with a water-cooled cooler, the pressure of oil in the cooler must be greater than that of the water, so that, if there is any leakage, water will not be forced into the oil. In the case of transformers with fan coolers, these should be checked and tested, and the setting for the cutting-in of the fan checked.

Breathers, where used, should be fully charged with the drying agent. The jointing of the cable boxes should be checked, and where the transformer is fitted with bushings, these should be examined for cracks or chips, and if fitted with arcing horns, the gaps should be checked and set.

If the transformer is required to operate in parallel with another unit, it should be 'phased in', i.e. the phase sequence on the secondary, with the primary excited, should be identical with that of the other unit before the secondary is connected to the common outgoing supply. In a number of three-phase transformer connections there is a phase displacement of the secondary line voltage, as for example in a delta/star connected transformer. If the primaries of this be paralleled with the primaries of a star/star transformer, then it is not possible to run the secondaries in parallel, for the line voltages in the two cases have a phase displacement of 30°. A delta/star group may be paralleled with a star/delta or a star/interconnected star group, but not with a star/star or delta/delta group. A grouping commonly found in industrial and commercial work is the delta/star, either a Dy 1 or a Dy 11. In normal circumstances it is not possible to parallel Dy 1 and Dy 11 transformers, since, when the primaries are connected to the h.v. supply, there would be a 60° phase displacement between the l.v. connections. However, this can be corrected by crossing two of the h.v. connections on one of the transformers and the corresponding pair of l.v. connections.

After switching in and applying full voltage successfully, it is desirable to have the transformer operating in this no-load condition for as long as is practicable. The heat due to core loss warms the coils and the oil gradually; this minimises absorption of moisture and also allows trapped air to be removed by the circulation of the oil.

Inspection of naturally cooled transformers should be made annually; artificially cooled units should be inspected every 6 months. It is always advantageous to take a sample of oil and test it for electric strength. This gives an indication of the presence of moisture or other impurities in the oil. At all times oil levels should be maintained correctly and breathers should be recharged regularly.

38.12.2 Transformer protective devices (oil filled)

The practicable indication of the temperature rise of a transformer in service is the top oil temperature. There is a temperature differential between the winding and oil temperatures, and the winding temperature responds much more quickly to changes of load, but the winding temperature cannot readily be monitored.

On larger units it is usual to fit an instrument on the tank to read the sum of the oil temperature and an analogue of the winding gradient. The instrument comprises a normal dial thermometer, the bulb of which is surrounded by a heating coil, with thermal characteristics similar to those of the transformer windings and fed from the secondary of a current transformer. In this manner the dial reading gives an indication approximating to the temperature of the windings.

Winding temperature indicators are fitted with alarm contacts, so that warning can be given should the windings reach dangerous temperature.

The Buchholz relay is a mechanical device fitted in the pipe between the transformer tank and the conservator. It usually consists of two floats with contacts: one to an alarm circuit, the other to a trip circuit. Any breakdown in transformer insulation is accompanied by the generation of gas in the oil. A serious fault results in the rapid generation of gas, and as this rushed through the pipe, it operates the float and closes the trip contacts. Alternatively, if the fault develops slowly, gas is generated slowly but is sufficient to operate the alarm float and contacts.

38.13 Power-factor correction

Most industrial loads have a lagging power factor, and since most electricity tariffs are constructed to penalise low power factors there is a good commercial reason for installing power factor correction equipment.

Design is usually straightforward. There are three points where the correction can be applied; at the individual piece of plant; for a group of plant items (say in one workshop in an industrial complex); or at the main supply. Only the user can determine the optimum solution.

The common method of correction is by means of static capacitors, generally oil-impregnated and oil-cooled and with a paper dielectric. The loss in a capacitor is less than 2 W/kVAr, and the temperature rise does not exceed about 15 °C above ambient. Maintenance is negligible, and it is unnecessary to filter or replace the oil during the life of a capacitor. No special foundations are needed. Capacitors are available for direct connection to systems up to 33 kV at power frequency, and can be installed indoors or outdoors.

An alternative power factor correction method is the use of synchronous machines with over-excitation. Operated at a leading power factor, the machine can correct the lagging power factor of the rest of the system. However, the method is applicable only when the synchronous machine is required for a specific duty and when system conditions are such that the motor is not shut down while the rest of the system is still in operation.

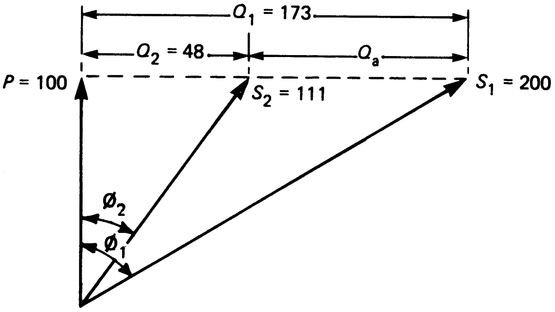


Figure 38.12 Power-factor correction

38.13.1 Capacitor rating

In order to evaluate the capacitor requirements, the system power factor must be known: it can be obtained from a power factor meter, or calculated from active, reactive and apparent power indicators.

Let a load of active power P, reactive power Q_1 (lag), apparent power S_1 and power factor cos ϕ_1 require correction to a lagging power factor cos ϕ_2, corresponding to P, Q_2 and S_2. The correction is shown in *Figure 38.12* with specific reference to a three-phase 415 V 50 Hz load of $P = 100$ kW and a power factor of cos $\phi_2 = 0.9$, both lagging. For the two conditions:

Power factor 0.5:
$P = 100$ kW $S_1 = 200$ kVA $Q_1 = 173$ kVAr
Power factor 0.9:
$P = 100$ kW $S2 = 111$ kVA $Q_2 = 48$ kVAr

The required capacitor rating is therefore $Q_\text{a} = \Leftarrow (Q_1 - Q_2) = 125$ kVAr. With a delta-connected 3-phase bank, each branch is rated at $125/3 = 42$ kVAr. *Table 38.2* gives the rating (in kilovars per kilowatt of load) to raise a given power factor to a selection of higher values. To obtain power factors approaching unity, the capacitor rating tends to be large, and it is usually uneconomic to correct to power factors greater than about 0.95.

For a rating Q_a (in kilovar) the capacitance C (in micro farads) required is:

$$C = (Q_\text{a}/2\pi f V^2)\ 10^9$$

where f is the frequency (in hertz) and V is the line voltage (in volts).

38.14 Earthing

Earthing an electrical system is the process of connecting all metalwork (other than the electrical power conductors themselves) to the main body of earth. The aim is to convey to earth any leakage of electrical energy to the metalwork without hazard to personnel or equipment. An earthing system has two distinct but related parts: (i) a low-resistance conductor bonding the metalwork, connected to (ii) an electrode or array of electrodes buried in the ground. A 'good' earth may be difficult to achieve. The main factors in (ii) are the form and configuration of the electrode assembly, and the electrical resistivity of the soil.

Table 38.2 Capacitor rating (kVAr/kW of load)

$\cos\phi_1$	Correction to $\cos\phi_2$				
	0.85	0.90	0.95	0.98	1.0
0.40	1.67	1.80	1.96	2.08	2.29
0.45	1.36	1.50	1.66	1.78	1.99
0.50	1.11	1.25	1.40	1.53	1.73
0.55	0.90	1.03	1.19	1.32	1.52
0.60	0.71	0.85	1.00	1.13	1.33
0.65	0.55	0.68	0.84	0.97	1.17
0.70	0.40	0.54	0.69	0.81	1.02
0.75	0.26	0.40	0.55	0.67	0.88
0.80	0.13	0.27	0.42	0.54	0.75
0.85	—	0.14	0.29	0.42	0.62
0.90	—	—	0.15	0.28	0.48

38.14.1 Earth electrode

The ideal electrode is a conducting hemisphere. Current entering the ground from the hemispherical surface flows in radial lines, with concentric hemispherical equipotential surfaces decreasing in potential from the electrode. Such an electrode, or radius r in soil of resistivity ρ, has a resistance to the main body of earth given by:

$$R = \rho/2\pi r$$

Most of this resistance is located near the electrode surface, where the current density is greatest.

Practical electrodes are usually rods, plates or grids, and have a non-uniform current distribution in close proximity, but the more remote form the electrode system, the closer does the current distribution resembles that of the hemisphere. Formulae for estimating the value of R for typical configurations are given below. In each case ρ is the soil resistivity.

Single rod of diameter d and buried length l:

$$R = \frac{\rho}{2\pi l}\left[\log_e \frac{8l}{d} - 1\right]$$

Multiple rods: it is seldom that a single rod can provide a low enough resistance and so arrays of rods have a relevance and are commonly used for substation earthing. The following formula is due to Tagg.[1]

For a group of n rods, each of resistance R_1, in a hollow-square configuration with a spacing s between rods:

$$R = \frac{R_1}{n}(1 + k\alpha)$$

where $\alpha = r_h/s$, where r_h is the radius of the equivalent hemisphere for one rod, and k is a factor depending on n.

Figure 38.13(*a*) gives values for r_h for 25 and 12 mm diameter rods buried to varying depths. Variations in diameter have little effect on the overall resistance.

Figure 38.13(*b*) gives values for factor k for varying numbers of rods.

Buried horizontal wire: buried wire or strip conductor is often used for earthing, particularly where ground conditions, as in hilly and mountainous country, make the driving of rods difficult or impossible. Formulae for the calculation of resistance values are complex, and it may be more helpful to quote values for particular conditions and leave the reader to make his own extrapolations.

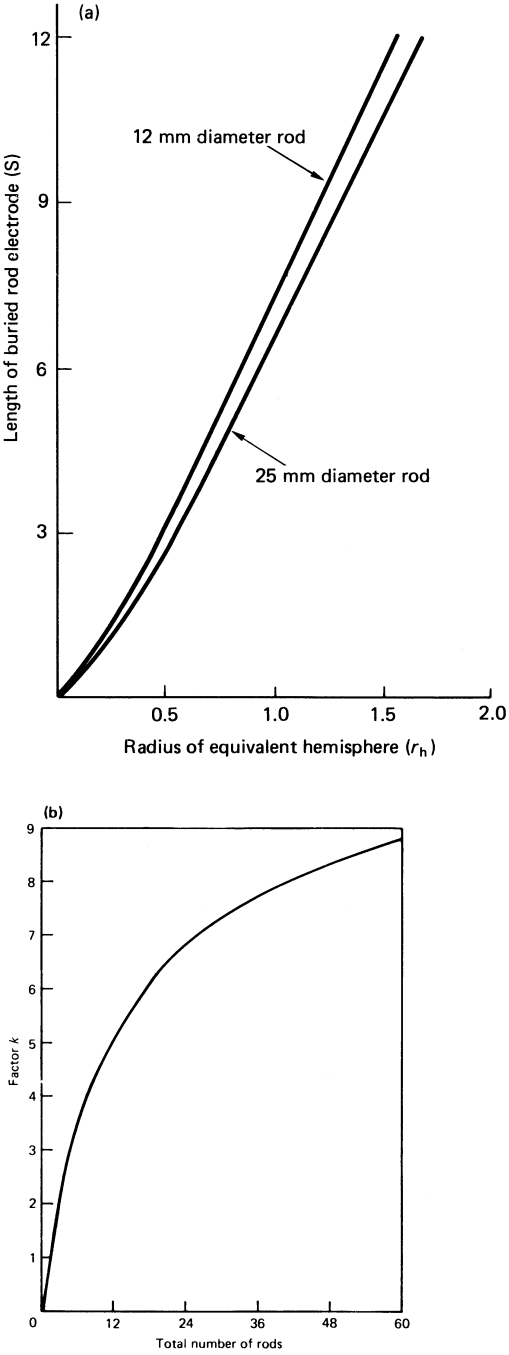


Figure 38.13 (a) Equivalent hemisphere radius *v.* length of rod electrode; (b) Factor *k* for hollow-square multiple-rod earthing array

For a conductor 2.5 cm in diameter, and buried about 1 m deep the resistance values shown in *Figure 38.14* would apply. Resistivity is assumed to be 10 000 Ω-cm. Within reason, neither the depth of burying nor the diameter of conductor make significant difference to the resistance.

Since it is not always possible to lay an unlimited length of conductor in one direction, alternative configurations are

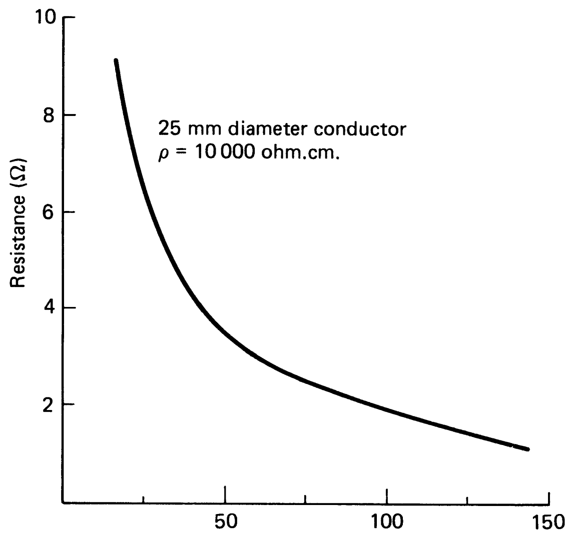


Figure 38.14 Resistance of buried horizontal conductor

often employed. For a total length of the above conductor of 150 m the resistance values would be as shown below. Note that for other values of ρςthe figures in curves and tables should be multiplied by ρ/10 000.

Arrangement	Resistance (Ω)
Straight wire	1.4
Right angle	1.45
Three-point star	1.49
Four-point star	1.6
Six-point star	2.0
Eight-point star	2.31

Plate electrodes: In the past considerable use was made of plate electrodes, but in general they are not as efficient or economical as rod or strip electrodes. When used, plate electrodes should be buried vertically with the top of the plate not less than 0.5 m below the surface. A simplified formula for the resistance of a plate electrode is:

$$R = \frac{\rho\varsigma}{223}\sqrt{\frac{1}{A}}$$

where *A*, the area of the plate is in square metres.

The addition of coke having a relatively low resistivity, to a distance of 250 cm all round an iron plate electrode may reduce the resistance to earth by up to 50%, but coke is liable to increase corrosion both by chemical and electrolytic action. It must not be used with copper electrodes, and cables should not be run within about 6 or 7 m of the plate electrode.

All the above formulae for resistance values of earth electrodes should be viewed with caution, but once the value of the soil resistivity has been established they do provide the engineer with guidance as to the best method to adopt in any particular application. The following section deals with soil resistivity and its measurement, but it should be recognised that a major problem in pre-determining a value for earth electrode resistance is that in practice soil is seldom homogeneous, and values can vary widely, even within the fenced area of one substation.

Computer programs are now available to be used to design complex earthing systems.

38.14.2 Soil resistivity

The quality of the soil (sand, clay, gravel, rock, etc.) and its homogeneity determine its resistivity. Typical values (Ω-m) are:

Rich arable soil, wet or moist	50
Poor arable soil, gravel	500
Rocky ground, dry sand	3000

The resistivity is greatly affected by (i) moisture content, (ii) temperature and (iii) the presence of chemical contaminants. Thus, the resistivity rises substantially when the moisture content is below 30% and also when the temperature falls below about 10°C; but the presence of salts in the soil lowers the resistivity. It is evident that when soil is subject to considerable seasonal changes of climate, earthing electrodes should be deeply buried. In practical terms this suggests the adoption of rod electrodes, which can be driven to a depth where the moisture content and the temperature are more stable.

38.14.3 Electrode installation

For toughness and resistance to bending, driven electrodes are of steel rod, either galvanised or with copper molecularly bonded to the surface (*Figure 38.15*). Rods of diameter 15–20 mm can normally be satisfactorily driven. For adequate depth, rods supplied in standard lengths can be connected by screwed couplers or special connectors. In an installation requiring few rods, manual driving is easy and economic, but for hard compacted soil, and where many rods are concerned, it is preferable to use a power hammer.

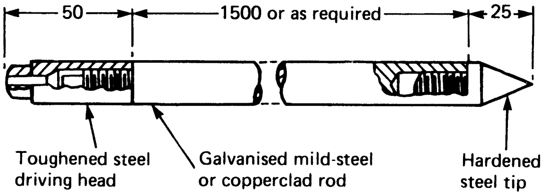

50
1500 or as required
25
Toughened steel driving head
Galvanised mild-steel or copperclad rod
Hardened steel tip

Figure 38.15 Typical earth-rod electrode

Table 38.3 Earthing-rod fusing currents (kA)

Rod section area (mm^2)		*Duration* (s)		
Copper clad	*Equivalent steel*	0.25	0.5	1.0
62	72	15	11	8
98	115	24	17	12
113	132	28	20	14
126	147	31	22	16
158	183	39	27	19
235	271	57	41	29
422	522	110	78	55

Each rod should be driven until a minimum resistance reading has been reached, as described later. If the overall resistance of a given rod assembly is not low enough, additional rods may be driven, spaced typically at double the depth.

An earth electrode system must resist corrosion. Both galvanised and copper-clad steel rods have good corrosion resistance. The system must also retain its ability to carry large currents repeatedly throughout its life (30 years or more). Individual rods can deal with currents up to the fusing values, indicated in *Table 38.3* for three typical durations.

Plate electrode This has lost favour, being more expensive to install, more liable to corrosion, and buried in top layers of soil where stable soil-resistivity conditions are not readily achieved.

Grid electrode Where the ground is rocky and the soil resistivity high, the only practicable method of obtaining an adequately low earth resistance may be by installing a grid of buried conductors.

Addition of chemicals In general, this is of limited application, as chemicals tend to leach out, but where earthing is difficult, gypsum (a form of calcium sulphate) may be employed, mixed into the top soil or spread on the surface. Its value lies in its relatively low solubility in water. With moderate rainfall, gypsum percolates through the soil and continues to keep down the earth resistivity for the life of the electrodes. It has minimal corrosive effect, but sulphates do attack concrete.

38.14.4 Resistivity and earth resistance measurement

The soil resistivity must be known for the design of an earthing system to be determined and the earthing resistance of the system must be checked after installation. Both measurements involve passing current through the soil and measuring the volt drop. Composite instruments are available. *Figure 38.16*(*a*) shows the measurement of *resistivity*. The test prods are pushed into the ground to a depth not exceeding $s/20$, where s is the spacing between prods. The greater the spacing s, the greater the volume of soil concerned in the measurement. By repeating the test with various values of s, the uniformity of the soil can be assessed: reasonably constant values indicate good soil homogeneity.

Testing *electrodes*, individually or in an array, is shown in *Figure 38.16*(*b*), using the same test set with C_1 connected to P_1. Prods C_2 and P_2 should be far enough from the electrode system for accurate readings of resistance to be obtained. Provided that there is an adequate distance between C_1 and C_2, there should be a zone within which the potential prod P_2 gives a constant reading.

38.14.5 Protective conductors

The earthing conductor connects the consumer's main earth terminal to the earth-electrode system, but such disconnection must require the use of tools.

Under the IEE Regulations WR16, the *circuit protective conductors* (formerly the 'earth-continuity conductors') are those that join the consumer's main earth terminal to all the exposed conductive parts of the system. The circuit protective conductor provides a low-impedance path for fault current, but must also ensure that no dangerous voltage can occur in metalwork in the vicinity of the fault. This is achieved by the connection of all extraneous metalwork to the main earth terminal: the provision covers such items as water and gas pipes, accessible structural steelwork, and central heating pipework, a requirement termed *main equipotential bonding*. The connections should be made as close as possible to the point of entry to the building concerned. Additionally, *supplementary bonding* is the connection of extraneous conductive parts that are accessible simultaneously with other conductive parts but are not electrically connected to the equipotential bonding: e.g. water taps, sink and bath waste outlets, central heating radiators.

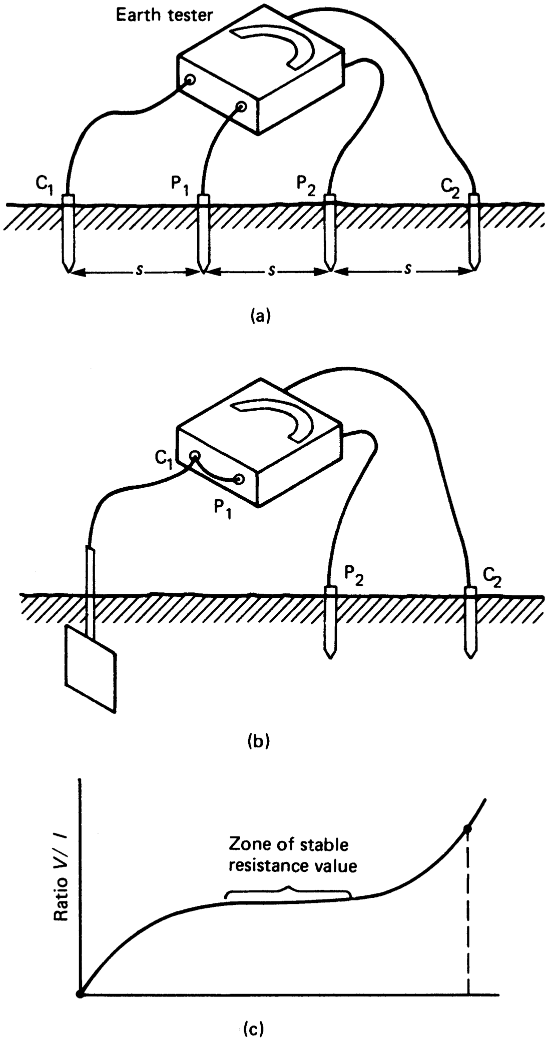


Figure 38.16 Measurement of (a) soil resistivity and (b) earthing resistance

The cross-sectional area (S in square millimetres) of the circuit protective conductor may be calculated as

$$S = (1/K)\sqrt{(I^2 t)}\Leftarrow$$

where I is the current (in amperes) for a fault of negligible impedance, t is the operating time (in seconds) of the disconnecting device, and K is a factor depending on the material of the conductor, the insulation, and the initial and final temperatures. The operating time is usually $t = 5$ s for fixed and $t =$ ~~0.~~4 s for portable equipment. A quick rule-of-thumb alternative to the formula is: S is the same as the cross-sectional area of the phase conductor where the latter is 16 mm^2 or less; $S = 16$ mm^2 for phase conductor sections between 16 and 35 mm^2; and S one-half of the phase conductor section for larger areas.

For main equipotential bonding conductors, the area shall be not less than one-half of the main earthing conductor of the installation, with a minimum of 6 mm^2 (10 mm for protective multiple earthing (PME). In general, supplementary bonding using 4 mm^2 bare or 2.5 mm^2 mechanically protected conductors should be acceptable.

38.14.6 System earthing

WR15 introduced a new designation for electricity supply systems to identify the type of earthing in use. WR16 continues this system, using a three or four letter code as follows:

First letter—'supply' earthing arrangement. T indicates that one or more points of the supply are directly earthed. I indicates that the supply is nowhere earthed, or is earthed through a fault-limiting impedance.

Second letter—'installation' earthing arrangement. T indicates that all exposed conductive metalwork is connected directly to earth. N indicates that all exposed conductive metalwork is connected directly to the earthed supply conductor (usually the neutral).

Third and fourth letters—earthed supply conductor arrangement. S indicates separate neutral and earthed conductors. C indicates combined neutral and earth in a single conductor.

In a TT system the exposed conductive parts of the installation are connected to an earthing electrode which is independent of the supply earth.

In a TN–S system the consumer's earth terminal is connected to the supply authority's protective conductor, usually the cable sheath and armouring, and so provides a continuous metallic path back to their earth electrode. The majority of installations with an underground supply are of this type.

In a TN–C system the neutral and protective functions are combined in a single conductor throughout the system. Typical is a system using earthed concentric wiring. If fed from a public supply authority, this system can only be used under special conditions and with the authority's permission.

In a TN–C–S system the supply neutral and protective functions are combined and earthed at several points, often with an earth electrode at or near the consumer's incoming terminals. The exposed conductive parts of the installation are connected to the neutral and to the supply neutral/earth at the consumer's main earthing terminal. This system is also referred to as *protective multiple earthing*.

In an IT system the exposed conductive parts of the installation are connected to the consumer's earth electrode, while the supply is either isolated from earth or connected to earth through an impedance. This form of supply cannot be used for public supply in the UK.

38.15 Inspection and testing

On completion and before energising, every installation must be inspected and tested to ensure that it complies with the appropriate regulations.

38.15.1 Inspection

WR16 calls for an inspection to be made of each installation before it is energised to ensure that the requirements of the regulations have been met. This must confirm that all electrical equipment is in compliance with the appropriate national standards, has been correctly selected and installed, and is in a safe working condition. The inspector/tester should understand how the installation functions and, to this end, he should ensure that his inspection covers such items as the identification and checking of the size of conductors, and the labelling of circuits, fuses, distribution boards and other switchgear. Finally, he should insist on the availability of diagrams, charts or tables describing the system under inspection.

38.15.2 Testing

Testing is required to ensure that the installation has been designed and installed (i) to give protection against electric shock during normal operation and during fault conditions, (ii) to ensure that it provides protection against overloads and short-circuits, and (iii) to provide protection against thermal effects—in particular, against fire and burns.

The IEE Regulations WR16 are quite specific regarding the items to be tested, and in particular they cite the following items which, where relevant, shall be tested.

Initial tests should be carried out in the following sequence:

(1) continuity of protective conductors, main and supplementary bonding;
(2) continuity of ring final circuit conductors;
(3) insulation resistance;
(4) site-applied insulation;
(5) protection by separation of circuits;
(6) protection against direct contact, by a barrier or enclosure provided during erection;
(7) insulation of non-conducting floors and walls;
(8) polarity;
(9) earth electrode resistance;
With the electrical supply connected re-check polarity before further testing using a suitable main voltmeter.
(10) earth fault loop impedance;
(11) operation of residual-current-operated devices.

Tests must be carried out in the correct sequence. For example, it is important that the continuity (and, therefore, the effectiveness) of the protective conductors be checked before insulation tests are carried out, because an open-circuited protective conductor coupled with a low insulation resistance reading could make the whole system live at the test voltage during the insulation test.

In general, the application of any voltage to a system prior to connection to the mains must be done with care, to ensure that no danger to persons, property or equipment can occur, even if the tested circuit is defective. There is an increasing use of electronic components in present-day installations, and these can be damaged by relatively low voltages. Therefore the insulation resistance tests should be undertaken before the components are fitted, or they should be removed before testing commences.

38.15.2.1 Continuity

To check continuity it is not sufficient that a measurement be made across the end terminals. In the case of final ring circuits two measurements need to be made. Appendix 15 of WR15 detailed these, but in the bid for harmonisation this appendix has been dropped from WR16. The appendix proposed two acceptable methods, both assuming that an outlet is installed at or near the mid-point of the ring.

In method 1 the resistance of each pair of like conductors is measured with the two ends separated. After reconnecting the two conductors, a measurement is taken between the distribution board and the mid-point. After measuring and deducting the resistances of the test lead, it should be found that the first reading is four times the second reading (see *Figure 38.17*(*a*)).

In method 2 no long lead is needed to make contact with the mid-point of the ring. The continuity of each conductor of the ring circuit is measured between the ends of the conductor when separated, as for method 1. The conductors are then reconnected at the origin, all the connections at the mid-point outlet are short circuited and the continuity is measured between the phase and neutral terminals at the distribution board. This value should be one-half of the previous reading. Finally, the continuity is measured between the phase and the earth (see *Figure 38.17*(*b*)).

Continuity tests in protective conductors, particularly steel conduit, require the application of a current approximately 1.5 times the normal design current, using a voltage not exceeding 50 V a.c. or d.c. (If d.c. is used, verify that no inductor is included in the circuit.)

It will be appreciated that the resistances being measured in the continuity test will be small (e.g. between 0.1 and 0.5 Ω), and it is suggested that a testmeter with a range between 0.005 and 2.0 Ω would be useful.

38.15.2.2 Earth electrode

Methods of testing earth electrodes are described in Section 38.14.4 and illustrated in *Figure 38.16*.

38.15.2.3 Insulation resistance

Insulation resistance tests are carried out to verify that the insulation of the conductors and electrical accessories is satisfactory. A low reading would indicate a deterioration in the insulation, and a very low reading a total failure.

The test meter should provide a d.c. voltage of not less than twice the nominal voltage of the circuit under test (r.m.s. value for an a.c. circuit), but the test voltage need not exceed 500 V d.c. for installations between 500–1000 V.

Insulation resistance testing involves test of insulation resistance between conductors of different polarity and between all conductors and earth, and the value must not be less than 1 MΩ. A large installation may be tested by dividing it into groups, each containing not less than 50 outlets; the minimum insulation resistance demanded of each group is 1 MΩ. The term 'outlet' includes every point and every switch, except where the switch is incorporated in a lighting fitting, socket outlet or power-consuming apparatus.

During the insulation test to earth of a whole installation, all fuses should be in place and all switches closed, including the consumer's main switch (if practicable). For the test (*Figure 38.18*) all conductors are bonded together at the load terminal of the main switch by means of a copper wire. The bonding wire is then connected to the 'line'

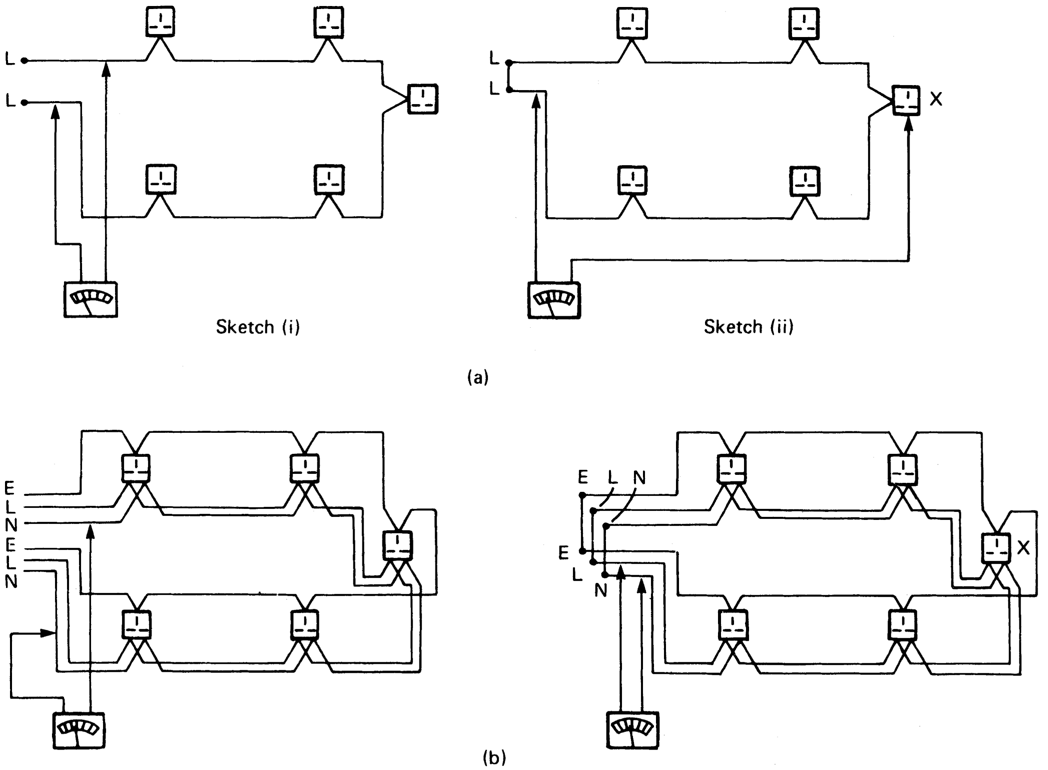


Figure 38.17 Measurement of the continuity of a ring circuit: (a) method 1; (b) method 2

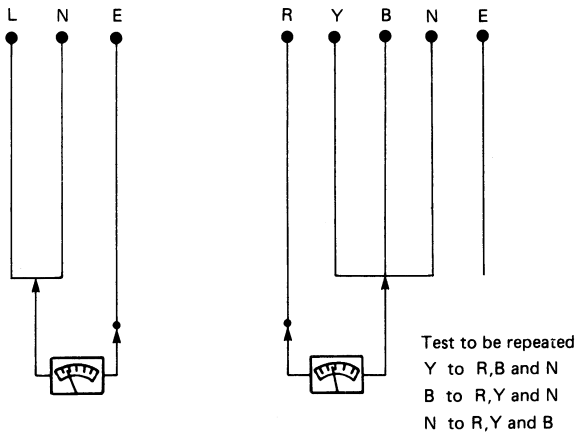


Figure 38.18 Insulation resistance testing

terminal of the tester, the other terminal being solidly connected to the consumer's earthing terminal. (When it is required to test to earth a live installation, the consumer's main switch must be opened and the supply from the mains disconnected.) Portable apparatus should not be connected to the system, as it can be tested separately. Any item of fixed apparatus may be disconnected if desired and tested separately. With the test set connected as described above, the insulation resistance is measured. Where two-way switches are included in the circuits, two test readings should be taken, one switch in each pair being changed over for the second test to include the strapping wires in the test.

If the insulation resistance value exceeds the specified minimum, the system can be passed; but if not, further sectional tests will have to be made to locate the faulty section.

For insulation testing between conductors, if practicable all switches and circuit-breakers should be closed, and all lamps removed, together with all fixed appliances.

Any equipment disconnected before the tests are done, and having exposed conductive parts, should itself have an insulation value of not less than 0.5 MΩ.

38.15.2.4 Polarity

It must be checked that all fuses and single-pole circuit-breakers and switches are in the 'line' side, that the wiring of plugs and socket outlets is correct, and that the outer contacts of concentric bayonet and screw lampholders are in the neutral (or earth) side.

38.15.2.5 Earth-fault loop impedance

If a fault of negligible impedance should occur from a phase conductor to earth, the supply voltage will provide a fault current through the earth-fault loop. The significance of the earth-fault loop impedance is that it determines the current, and this, in turn, determines whether or not the protective device will operate quickly enough to meet the requirements of the relevant regulations.

Take, for example, the arrangement shown in *Figure 38.19* and assume a fault to occur in an item of consumer's equipment. The fault current I_F will flow from the transformer, through the fuse and then via the casing of the equipment to the consumer's earth, and thence back through the earth path to the transformer, such that $I_F = V_s/Z$, where Z is the loop impedance. From Z calculations can be made for the overall performance of the system under fault conditions. For example, WR16 calls for all equipment connected to socket outlets to be cleared in 0.4 s or less when faulted. Other circuits supplying only fixed equipment are acceptable if the fault is cleared within 5 s. The regulations give tables of Z to meet these conditions.

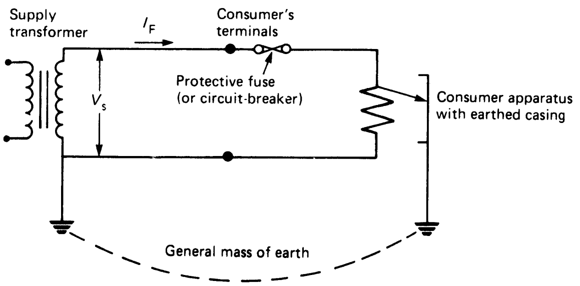

Supply transformer
I_F
Consumer's terminals
Protective fuse (or circuit-breaker)
V_s
Consumer apparatus with earthed casing
General mass of earth

Figure 38.19 Earth-fault loop impedance

In practical terms, a phase-to-earth loop can be tested by connecting a known resistor R between the phase and protective conductors when the system is live with R typically 10 Ω. Measuring the volt drop v across it gives the current $I = v/R$. If the supply voltage is V_s, then $V_s = I(Z + R)$ or $Z = (V_s/I) - R$, giving the required loop impedance.

The loop impedance must be tested using a *phase-earth loop tester*, which gives a direct reading in ohms. (The neutral-earth loop tester previously specified is no longe acceptable.) A dual-scale instrument covering 0–2 and 0–100 Ω is suitable.

38.15.2.6 Residual-current devices

Where the loop impedance is such that the required tripping times cannot be achieved, and in special cases (e.g. where the socket outlet in a house is used to supply gardening tools, or where the supply authority does not provide an earth terminal), a *residual-current* device must be fitted. It usually comprises coils on a magnetic circuit to carry the phase and neutral currents in opposing directions. In balanced conditions no magnetic flux is set up; but if a fault occurs on the system, even one of high resistance, the phase and neutral current imbalance induces an e.m.f in a third coil (*Figure 38.20*), tripping the circuit.

To test the device, an a.c. voltage not exceeding 50 V r.m.s. obtained from a mains-fed double-winding transformer, is applied across the neutral and earth terminals. The device should immediately trip.

Fault-voltage operated circuit-breakers may alternatively be used, but are subject to tripping when faults occur outside the protected zone. If used, however, the earth-fault loop impedance including earth-electrode resistance must not exceed 500 Ω.

Whether or not this simple go/no-go test is sufficient is debatable. Test sets are available to measure the effectiveness of the RCD which simulates an appropriate fault condition with respect to time.

38.15.3 Test gear

Because the testing requirements of WR16 are stringent, the UK Electrical Contractors' Association has suggested that the tester should equip himself with the following instruments, incorporating the features itemised.

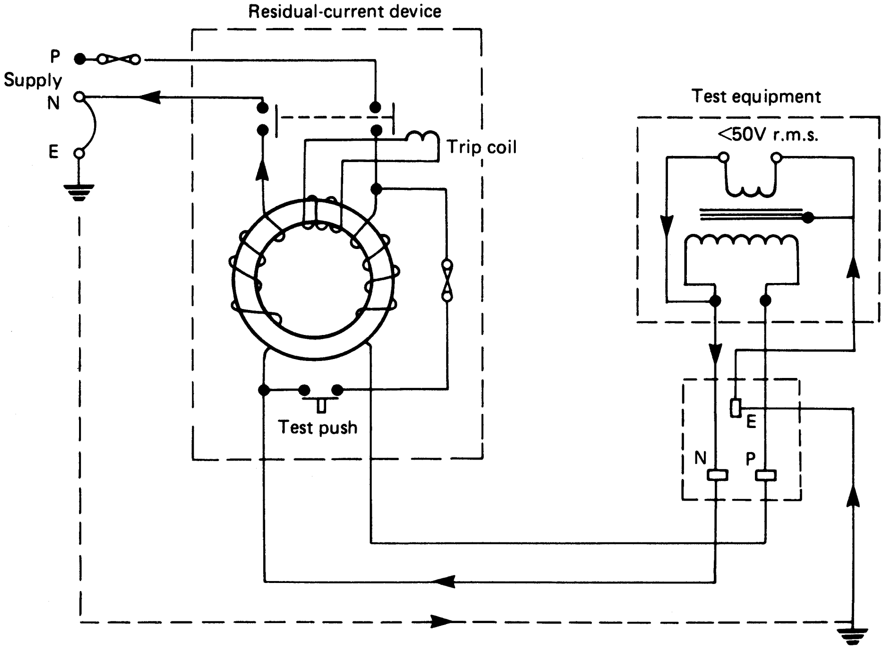


Figure 38.20 Residual-current device, showing test connections. The arrows show the path of the test current

Insulation-resistance/continuity tester—with 500/1000 V selector, analogue scale down to 0.01 Ω, battery power, push-button operation.

Digital milliohmmeter—with battery power, push-button operation, voltage protection indicator.

Polarity/earth-fault loop-impedance tester—with mains power, polarity indicator lamps, analogue low-resistance scale.

Residual-current-device/fault-voltage-device tester.

Earth-electrode/soil-resistivity tester.

Reference

1 TAGG, G. F., *Earth Resistance*, Newnes, London (1964)

Dedication

This section of the *Electrical Engineer's Reference Book* is dedicated to Jimmy Milne, BSc (Eng), C Eng, FIEE who has recently passed away and was not able to read my transcript of the updates from his original comprehensive work. I first met Jimmy when he was a Partner with Kennedy & Donkin Ltd, now Parsons Brinckerhoff Ltd, and I was an employee in the International Division. I am sure that all his professional colleagues will miss him.

Section H

Power Systems

39

Power System Planning

B J Cory DSc(Eng), FIEE, Fellow IEEE, CEng

Contents

39.1 The changing electricity supply industry (ESI) 39/3
  39.1.1 Central planning 39/3
  39.1.2 Vertical integration of generation and supply 39/3
  39.1.3 Deregulation and restructuring 39/3
  39.1.4 Least cost planning 39/3

39.2 Nature of an electrical power system 39/4
  39.2.1 Electricity supply 39/4
  39.2.2 Transmission 39/4
  39.2.3 Distribution 39/5
  39.2.4 Loads 39/5

39.3 Types of generating plant and characteristics 39/7
  39.3.1 Steam turbine 39/7
  39.3.2 Gas turbine and diesel 39/7
  39.3.3 Renewables 39/8

39.4 Security and reliability of a power system 39/9

39.5 Revenue collection 39/9

39.6 Environmental sustainable planning 39/9

39.1 The changing electricity supply industry (ESI)

Electrical power systems have developed over the 20th century to become the supplier of electrical energy to almost every household, factory, commercial building and activity in the developed world. Developing countries strive to emulate this achievement through attracting finance to build up their own infrastructure to stimulate their economies and to improve the quality of life for their inhabitants.

As with all industries, changing and improving technology through innovation, research and development has led to a change in the way power networks are planned and operated, initially as comparatively small local generation and supply networks, followed in the developed nations by more and more connections between local networks culminating in large continent wide grids with numerous large generating plant and components, high voltage systems and a supply to everyone requiring electrical energy. It has only been in the last 15 years from about 1985 that smaller operating units and private ownership of parts of the system has been seen to be an advantage.

39.1.1 Central planning

Power networks grew in capacity and extent in most industrialised countries through a centrally (government) controlled industry, often founded by treasury loans or sanctioned loans from the financial markets. This meant that with government involvement it was necessary to make a good case for the finance to be forthcoming to add to the present system or build anew to meet expected increase in electrical energy demand for all purposes. Consequently, centrally planned systems needed expert forecasting of demand at least 5 years ahead to design, manufacture and commission the necessary equipment and connections by cable or overhead line. Throughout the world the problem over many years was to achieve good forecasts which turn out to be dependent upon economic cycles and gross domestic product (GDP) which are well known for being extremely inaccurate, particularly in the 5 year time scale. The inevitable result has been that the ESI in most countries has been underdeveloped and never matches supply to demand, or overdeveloped where plant has been underutilised leading to inefficiencies and a higher price for energy than necessary. After 30 or 40 years from the 1940s to 1980s of attempting to make central planning work, the concepts of economic efficiency and plant installation by funding through the principles of the investment market—as in the oil and gas industries—was seen to provide a more acceptable way of achieving the desired objectives. This has produced a movement by many governments to deregulate and restructure the ESI such that central planning is no longer the norm, but private finance is encouraged to take the risk of getting a return on investment.

39.1.2 Vertical integration of generation and supply

Under central planning and government control, it was considered necessary to plan and operate the network as a unity such that the production, distribution and sale of electricity was either organised and run by one utility or each part had statutory obligations to ensure fairness and constraint on profit taking. Most countries, including the USA, ensured that utilities could not act as monopolies but only charge a reasonable tariff to obtain a regulated return on capital invested. It became the practice for utilities to serve defined areas or populations without fear of competition and to purchase energy from their own generation or interconnected power network. This was known as 'vertical integration' and prevented other suppliers encroaching on the supply territory of another utility. If tariffs were set by governmental edict or statutory control, there was obviously little incentive for utilities to innovate or compete on price. To overcome this obstacle, many industrialised countries with a mature and reasonably adequate energy infrastructure looked for ways of using market competition to reduce energy tariffs and to provide an improved service to the consumer. This has produced 'deregulation' of the ESI.

39.1.3 Deregulation and restructuring

These are the 'buzz' words for a world-wide movement to improve consumer service and reduce prices in mainly the industrialised nations. It also means that controlled planning is no longer required—the necessary improvements in service, quality and price to produce satisfied customers being obtained through 'market' forces. It was quickly realised that if generating plant could be allowed to compete country wide on price and the transmission and distribution network could deliver the commodity (energy) to any part of the country, then competition to supply individual consumers could also be encouraged. Consequently, the break-up of the generation utility monopolies by selling shares (if government owned) or trading generation ownership on the stock market would achieve the desired ends. On the supply side allowing suppliers to encroach into others' territory would also produce competition with (hopefully) low prices and improved service. This was the desired deregulation and restructuring with no government involvement in setting prices provided that 'level playing field' competition could flourish.

Unfortunately, the power system network which enables the generator owner to sell energy to the suppliers was almost a natural monopoly in that it would be prohibitively expensive for another entrepreneur to construct an alternative supply network. Hence a *regulator* would be needed to ensure that fair prices were charged by the owners of the power network for energy transportation. The net result appears to be that generation and supply through private or market investment has obviated any planning needs through government forecasting but the power system can be made adequate for consumers through risk management implemented through many utilities.

39.1.4 Least cost planning

With a deregulated ESI, no longer is there a planning authority but all investors need to plan for their own investments and expected returns. This implies that companies who are thinking of financing a venture into the ESI, must consider all options open to them and estimate the risk and return on investment by considering the following aspects:

- the minimum cost investment to meet the desired objectives over a given period;
- the robustness of the proposal against likely changes in market rates, social climate, environmental constraints, etc.;
- likely level of co-operation between the company and the customers being served; and
- the financial viability of the whole proposal, including the attitude of the shareholders, the regulator and the government.

In theory, any investment in new plant (generator, transmission facility, supply services) should not occur until the cost of alternative measures to achieve the same objectives is equivalent to new plant costs. In practice, new plant investment may be required because of new developments, economic boom, improved supply security, availability of new technology etc. It is against these criteria that new investment will need to be measured.

39.2 Nature of an electrical power system

39.2.1 Electricity supply

All countries now have available some supplies of electricity to connected consumers. In many industrialised countries a nationwide grid or distribution system is installed so that generating plant can be 'pooled' through interconnections to supply customers from industrial/factory complexes down to the smallest residential consumer, perhaps with a single light or TV set. As electrical energy is not easy to store, except by converting it to some other form of easily stored energy e.g. water pumped to a higher reservoir, the generating plant output must always match instantaneously the demand of the loads plus the losses (hopefully less than 10%) in transporting and delivering demanded energy (units of kWh). For many good reasons, most small consumers require their supply at a low voltage (230 V in Europe, 110 V in USA for example) whereas to keep losses low, electricity needs to be transmitted over any distance at a high voltage (400 kV in Europe, up to 700 kV in US/Canada). Generation, on the other hand, is most economically done at around 20 kV thereby requiring a step-up in voltage to the transmission system and a step-down in voltage for distribution to the myriad of small (mainly residential) consumers. This transformation is readily done by high efficiency *transformers* which require, due to Faraday's law, an alternating voltage at 50 Hz in Europe, Japan, Australia, etc. and 60 Hz on the American continent. To keep material costs to a minimum, transmission and distribution is best done using a *3-phase* system but only a *single-phase* supply is required by small consumers. Large and intermediate consumers such as industrial processes, factories, large buildings and hospitals etc. are most economically supplied at a higher voltage than to small consumers, at between 10 to 20 kV.

Consequently, many distribution systems consisting of step-down transformers, cable or overhead lines operate at this voltage and only the final, comparatively short connections to individual small consumers, operate at 230 V or 110 V, usually by tapping off from a 3-phase system.

39.2.2 Transmission

Part of a typical generation and transmission network is depicted as a single line diagram as in *Figure 39.1*.

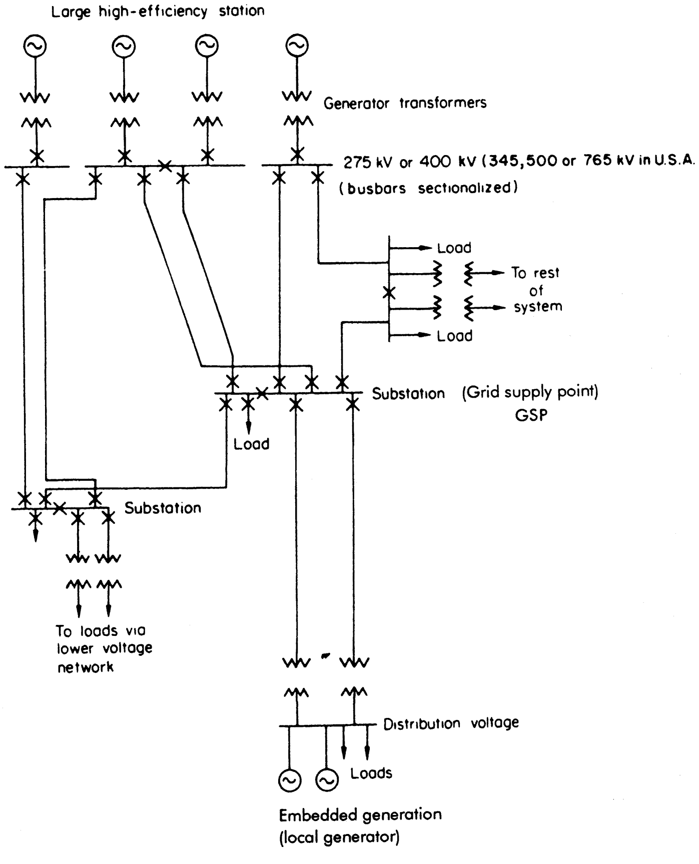

Large high-efficiency station
Generator transformers
275 kV or 400 kV (345,500 or 765 kV in U.S.A.
(busbars sectionalized)
Load
To rest of system
Load
Substation (Grid supply point) GSP
Load
Substation
To loads via lower voltage network
Distribution voltage
Loads
Embedded generation (local generator)

Figure 39.1 Part of a typical generation and transmission network. (Reproduced by kind permission of John Wiley & Son, Ltd)

It should be noted that generators are interconnected by a 3-phase system and it is essential that they run in synchronism with each other. If a generator cannot remain in synchronism due to a fault, then it must be disconnected by its circuit-breaker otherwise the whole system could collapse. In the figure, substations enable circuits to be switched and alternative routes are available if a circuit needs to be withdrawn for repair or maintenance. Distribution systems are fed from the high voltage network through step-down transformers and increasingly there are smaller generators 'embedded' in the distribution network adding to the combined energy output of the synchronised system.

In a deregulated ESI, the generators could be owned and operated by different utilities, the transmission lines and substations owned by other investors and the supplies to the distribution systems bought under contract to private distributors or suppliers.

Increasingly, for undersea connections or for connections between networks not in synchronism, high voltage direct current using semi-conductors as a rectifier one end and inverter the other is being used. Such connections should be considered as alternatives to a.c. connections.

39.2.3 Distribution

Medium and low voltage distribution systems vary in their design and layout depending upon the locality being served.

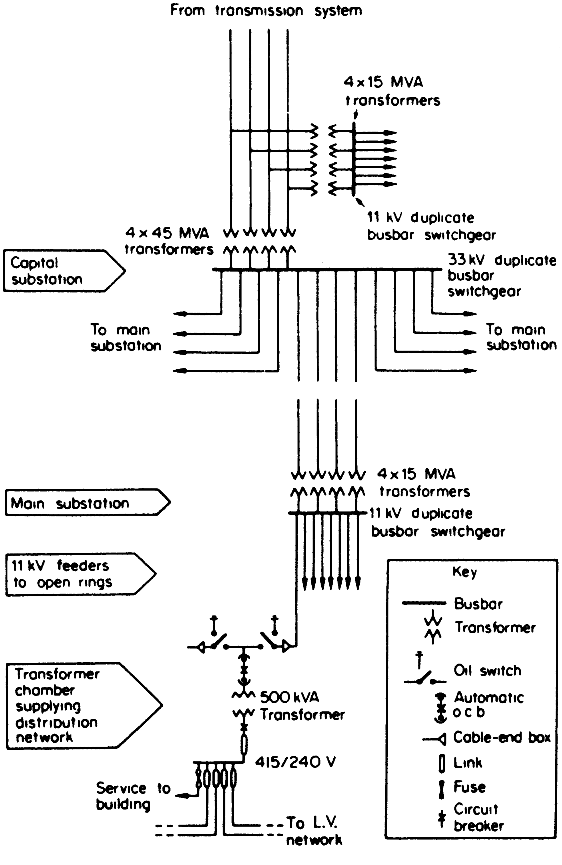


Figure 39.2 Typical arrangement of a supply to an urban network in UK. (Reproduced by permission of the Institution of Electrical Engineers)

In urban areas, where consumers are numerous and concentrated, an underground network with closely spaced step-down transformer substations are installed sized to meet the maximum expected demand after taking into account the average diversity of customer use. *Figure 39.2* depicts a typical urban network where each single line represents 3 phases contained within a single under road cable.

Because of the complexity of protecting the system if it was fully interconnected, it is usual to operate it as a radial system fed from the primary substations but allowing circuits to be supplied by alternative connections if any circuit is disconnected under fault or maintenance conditions.

For rural areas with sparsely sited consumers in farms or small villages, most supplies are stepped down by small pole-moounted transformers as close to the consumer as possible fed from a radial circuit. Fuse rather than relay operated circuit breaker protection is employed for cheapness and reclosing circuit breakers in the primary substation ensure that supplies can be quickly restored to healthy circuits once a fuse has blown because of a fault.

Figure 39.3 shows a typical rural distribution system in which section points allow manual reswitching to restore supplies after fuse disconnection. Overhead lines, either 3 phase or single phase, is the norm thereby allowing quick repair with a suitably equipped crew.

39.2.4 Loads

Consumer demand in a power system is often called a load and, of course, it will vary from hour to hour, day to day and season to season. Typical daily load curves aggregated over the whole England and Wales system are shown in *Figure 39.4*.

As previously remarked, the total generator output must match this demand and this requires the generators to be flexible. In practice, to maintain as high an efficient output as possible generators wish to run at full output or be offline so the system operator instructs plant to synchronise and desynchronise at pre-planned times worked out by some economic scheduling algorithm. At times when the demand is expected to fluctuate by ±5% or so within minutes (as could happen in a countrywide event when kettles or cookers are switched on/off in co-ordination by a TV programme) then a number of generators, particularly those with quick response, will be scheduled.

If the demand curve of *Figure 39.4* is plotted in descending order of magnitude as in *Figure 39.5*, the resulting diagram depicts a *duration curve*. Over a year of operation this curve indicates the *load factor* at which various kinds of generating plant can be expected to operate.

Normally the plant having the cheapest price per kWh would operate at base load and the peaking plant would operate on very small load factor around the peak demand. Other plant, depending on its characteristics and production costs, would be expected to run at the intermediate load factors, probably generating during the day and shutting down at night (known as *two shifting*).

In interconnected systems or power pools where energy trading is allowed, the operation of the system is much more complex. In this case, existing generating plant may find that it is unable to rely on base or intermediate load operation and must be installed and run according to its contracted output portfolio rather than in any economic sense. Plant without sufficient contracts to sustain their operation could therefore be isolated and eventually shut down, whereas newly installed plant with long term contracts could take its place. The availability of long term

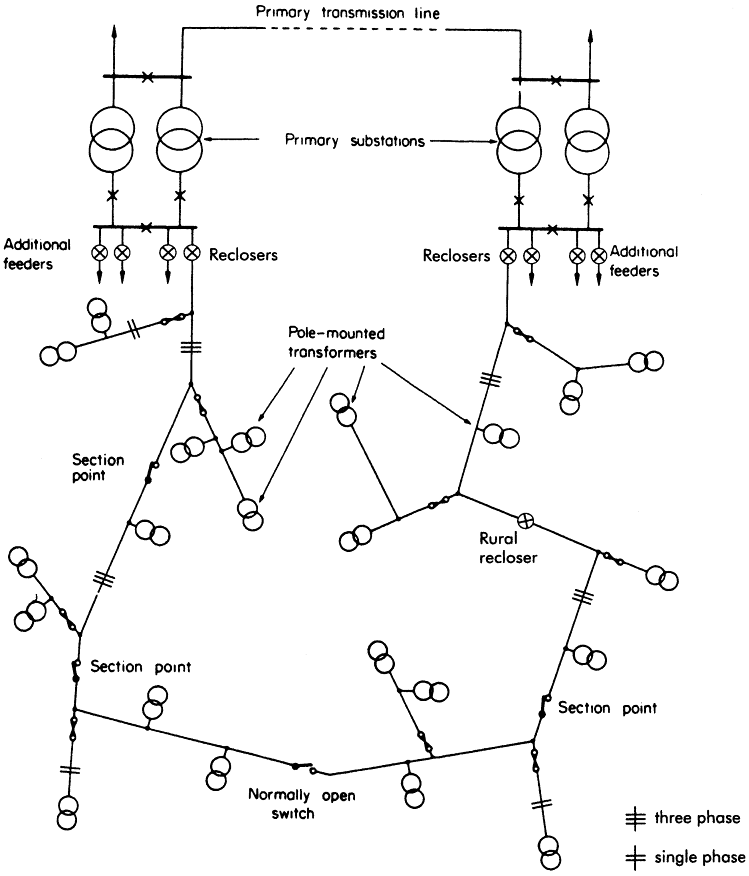

Primary transmission line
Primary substations
Additional feeders
Reclosers
Reclosers
Additional feeders
Pole-mounted transformers
Section point
Rural recloser
Section point
Section point
Normally open switch
three phase
single phase

Figure 39.3 A typical rural distribution system at 11 kV with step up and step down transformers, the latter protected by fuses

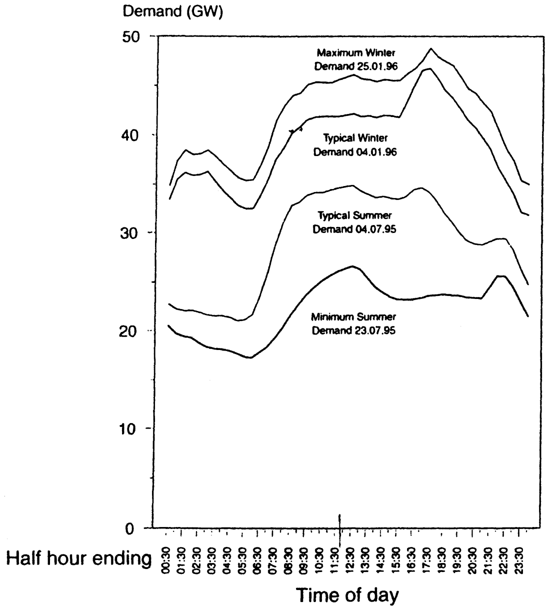


Figure 39.4 NGC summer and winter demands for 1995/96 (not weather corrected). (Reproduced by permission of NGC)

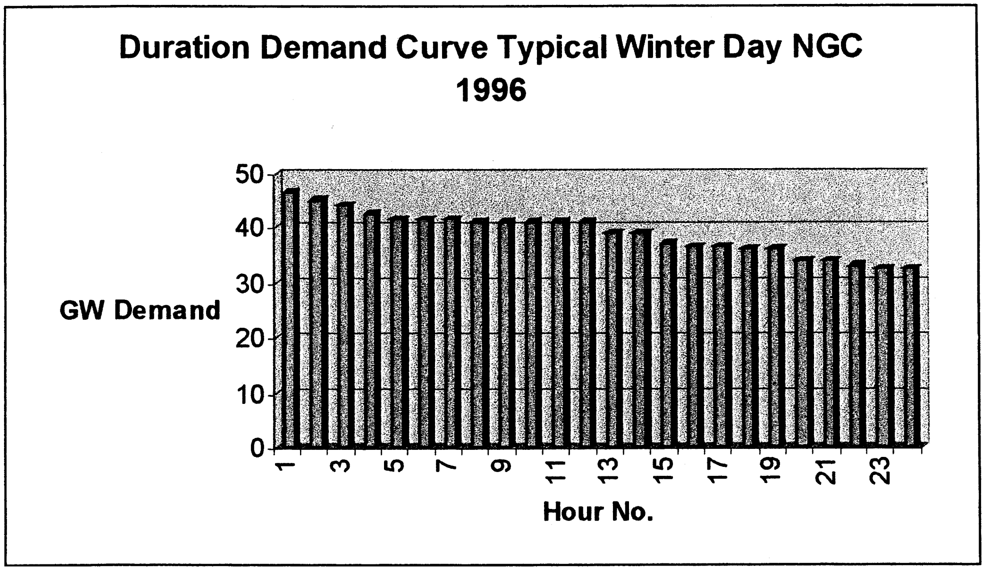


Figure 39.5 Demand duration curve for typical winter demand on 04.01.96

contracts on both the fuel supply side and the output energy side is likely to be the dominant feature of planning in the future.

39.3 Types of generating plant and characteristics

Details of particular types of generator are given in other sections of this volume. Here, the main characteristics for planning purposes are summarised to provide a comprehensive picture.

39.3.1 Steam turbine

The steam turbine, running at 3000 rpm for a 50 Hz output or at 3600 rpm for a 60 Hz output is the traditional work horse in most power systems.

It can be built in sizes up to 1200 MW when fuelled from coal, oil, gas, or nuclear sources. Coal fired plant requires quite expensive delivery, stocking and handling facilities compared with oil or gas thereby making it more difficult to finance initially although since the fuel is abundant in many countries and can perhaps be quarried rather than deep mined, the cost per MWh of output can be very competitive. Since most solid fuels contain sulphur, flue gas desulphurisation equipment may have to be statutorily fitted adding approximately 20% to the overall cost of the plant. The disposal of ash and gypsum (from the desulphurisation process) also has to be considered.

On the other hand, oil firing could also require some flue gas clean-up if sulphur is present but except in oil producing countries oil is relatively expensive compared to coal, but it is useful as a quick start (30 mins) source for peak lopping. Direct gas firing to raise steam is now a strong competitor to oil or coal, particularly if connection to a gas pipeline is inexpensive and gas can be used in a dual fired boiler to back-up coal firing dependent upon arbitrage costs. The problems of using nuclear reactors for steam raising are well known although some of the steam driven turbine generators are the largest units constructed.

All steam driven turbo-generators require cooling water to create a high vacuum in the low pressure turbine for maximum efficiency, some of which can reach 40% dependent upon the highest temperature and pressure that materials can tolerate. It should be noted that combined cycle gas turbines (see later) also depend for their high operating efficiency on pass-out steam turbines with cooling water condensers as in traditional steam raising plant. An abundant source of make-up water and water recirculation in the sea or inland lake for cooling purposes can often be an advantage compared to the need for large cooling towers or fan driven coolers in smaller plant.

39.3.2 Gas turbine and diesel

Since about 1980 the development of gas turbines running at 6000 rpm plus in 100–200 MW sizes because of the advances in high temperature blade material has enabled this combined cycle form of plant to generate at efficiencies of up to 60% and in comparatively small sizes. As this type of turbine is considerably smaller than a steam turbine of comparable power output, it can be started and stopped much more rapidly thereby producing the potential of being used as intermediate generation. As the site area necessary is also much smaller than required for steam boilers and turbines, it can be sited much closer to load centres or inside the curtailage of industrial complexes, thereby reducing the need for transmission to energy over some distance with consequent losses.

In the 1990s, combined cycle gas turbine (CCGT) had begun to provide the base load because of the comparative cheapness of gas as a fuel and the high operating efficiency.

As coal fired and nuclear plant had reached the end of its economic life, CCGT plant was being used more to supply base load, but during the early 21st century with the increasing installation of CCGT displacing older plant, intermediate (two shift) loading is becoming more usual. With two or more gas turbine driven generators whose waste heat raises steam for conventional generators, the flexible output obtained by varying the number of GTs is of prime importance to meet contracted varying output over the daily cycle. As the CO_2 produced by a CCGT with gas firing is only 55% of that of a similarly rated coal or oil fired plant, its popularity with environmentalists is evident. Note also that 10% overload for short periods is possible to meet short peaks and/or aid system recovery following a generation loss. Added to all these advantages is the comparatively low cost and short (2 years) installation time, particularly for CHP schemes, its use in power system planning is assured. An alternative to GTs in small CHP schemes (up to 50 MW) is that of diesel plant running at slow speed but overall efficiency of 90%.

39.3.3 Renewables

39.3.3.1 Hydro power and pumped storage

Until fuel cells or rechargeable batteries become economic and reliable, the well known water turbine driving a multipole generator at comparatively low speed is still a valuable fast response unit that can act as reserve or regulating output, particularly at peak periods. Combined with pumping action, either using the coupled turbine in reverse mode or a separately coupled pump, water can be stored in a high level reservoir at low energy cost periods for use during peaks or emergencies. Overall the combined pump-generate cycle can run at up to 70% efficiency, implying that provided the difference between peak and off-peak energy cost is 35% or more, a pumped storage plant can be run at a profit. Retrofitting hydro-plant with pumping equipment is another possible option, particularly as the first cost of hydro can be very high. Storage is essential if the renewable energy source is intermittent as described in the next subsections.

39.3.3.2 Wind generation

A popular renewable source of energy is from the wind, particularly in temperate zones where most days have a wind blowing. The economics of wind energy depend upon the range of wind speeds between the minimum 'cut-in' speed for the wind generator and the maximum speed that the wind blades are designed to withstand. A high average wind speed as usually found on exposed high ground is obviously an advantage, indicated on charts by 'isovent' contours. The visual impairment of any scheme needs the preparation of impact statements to satisfy the local environmental lobbies and the connections into what is essentially a rural distribution network requires careful design in conjunction with the distribution owner. Problems of voltage control, reverse current protection, safety of distribution personnel when sources of energy other than from the primary transformer are present in the network, all require consideration. Wind farms consisting of 20–30 wind turbines mounted on 30–50 m masts having 2 or 3 bladed sails, rotating at 100–150 rpm to produce up to 600 kW per turbine are now common. Reliability has increased and the cost per kW reduced as experience has been gained and selection of appropriate materials has improved. The output cost per kW in year 2000 is now, on average, twice that of coal or gas fired plant, but as the energy (when available) is fed directly into the local distribution system, thereby avoiding both transmission and the bulk of the distribution losses and use-of-system costs, the energy price to the small consumer is becoming competitive. It is possible that consumers are willing to pay a premium for energy said to come from a renewable source. In the UK, a good site can produce 1800 kWh of energy per kW installed in a year.

39.3.3.3 Solar power

In tropical areas and favourable temperate zones, power from the sun is becoming more and more flexible, either by using mirrors focused onto a steam raising boiler driving a conventional turbine or through the employment of solar panels with power electronic conditioning equipment to convert direct current into alternating current. The mirrors require angle positioning and tracking equipment for economic output and the area occupied can be comparable with that for a solar panel photo-voltaic array. Maintenance of mirrors and solar panels is required especially in dusty/desert zones. Installation of 10 MW or so are now becoming economically viable. Many of the considerations for wind energy installations apply to solar also.

39.3.3.4 Tidal, ocean temperature gradient and wave power

Tidal power in large estuaries where a suitable tidal range is available requires horizontal, two-way flow turbines for energy production. It is comparable in cost to hydro-electric schemes, implying that long construction times and up-front financing is necessary. It has not been a great success anywhere in the world because of the impairment of the natural environment likely to ensue. Using the temperature difference between surface water and that at 500–800 m deep could produce energy from a specially designed turbine at 3 to 4% efficiency but would require vast quantities of piped sea water needing large underwater structures not yet attempted. Wave energy, close to rugged shorelines is a prime example of an obvious and visual source if only it could be reliably and economically harnessed. The variations in wave height and strength over the seasons are well known and the ability to cater for the severest storm conditions by man-made structures is fraught with difficulty. Research into robust shore structures using the inflow and outflow of the wave energy to drive air through a reversible turbine is continuing.

39.3.3.5 Other renewables

There is some uncertainty as to exactly what should be regarded as a renewable resource. Wood or straw burning boiler plant for steam raising is acceptable—provided the wood is obtained from a managed forest where replanting on a scheduled basis is undertaken. Refuse incineration and landfill gas driven plant may or may not be classed as renewable depending upon the source of the bio-fuel. Such plants are increasing in number and extent and their assimulation into the local electricity supply system must be planned. Sources such as geothermal energy are valuable in suitable locations and as drilling becomes possible in deeper boreholes at reasonable cost, such 'renewable' sources using the heat of molten magma within the earth's crust become attractive.

39.4 Security and reliability of a power system

In industrialised countries, a high level of reliability of supply is expected amounting, on average, to not more than 60 min loss of supply (blackout) per year. In addition, voltage sags (dips), momentary interruptions or voltage surges should not cause interference with a consumer's installation. This condition is becoming much more onerous with the increasing use of power electronic equipment in which nuisance tripping can lead to shut-down of an entire industrial process. Consequently, considerable attention is now focused on *power quality* and the elimination of unwanted harmonics in the system. Many consumers with sensitive electronic equipment are encouraged to safeguard their vital equipment with an auxilliary uninterruptible power supply (UPS) using a battery or fly-wheel generator store. In general, the reliability of supply from generator to consumer depends upon the *availability* of generation plant, transmission and distribution equipment, the ability of the ESI to install sufficient generators to maintain a suitable margin of reserve capacity taking into account maintenance and forced outages and the design of the transmission system for security and continuity of supply to the consumer. On availability, all plant requires periodical planned outage for maintenance purposes and hopefully this aspect of planning will be done in conjunction with the system operator. Maintenance procedures are best developed from careful diagnostic measurements and statistical records including breakdown incidents plus advice from manufacturers. An increasing and desirable trend is towards more on-line diagnostic measurements from which possible breakdown occurences can be predicted, particularly if this leads to longer periods between maintenance or forced outages.

On the transmission system, networks should deliberately be designed to provide alternative paths for energy flow in the event of one (or sometimes two) of the *n* circuits being forced out because of a fault, lightning disturbance or vandalism. This is known as *n*-1 (*n*-2) security and could be a condition laid down by the regulator in granting a transmission licence. Incidentally, this 'overprovision' of circuits compared to that of a just 'adequate' system leads to fewer losses and hence reduced cost of operation.

On the distribution system, the expense of providing duplicate circuits is not justified except perhaps in densely populated areas such as cities and industrial complexes. Again statutory requirements or the regulator will lay down the 'target' frequency of interruptions and the time allowed for restoration, usually by designing the system for back-up re-switching through strategically placed isolators or circuit-breakers. The target 'restoration' time should account for proficient remote monitoring and alarming of supply conditions and the deployment of standby repair crews. Most countries have standards on demand levels and the security required dependent upon the MW disconnected through circuit outages.

39.5 Revenue collection

Of prime importance is the ability to meter and collect the dues from the various parties in the generation and supply chain. In a vertically integrated or government controlled power system, the most important revenue to be collected is from the final consumers. This is usually achieved through the reading of a customer meter, belonging to the power authority, and the sending out of a bill according to the agreement between the authority and the customer. Tariffs can be for units (kWh) consumed between the meter readings according to some scale, or, for the larger consumers, kVAh or kVArh and peak kW can be included. It must be emphasised that in a restructured ESI the final consumer may be supplied by an organisation who own no physical equipment but who make bulk purchases on the energy market and retail it at a negotiated price. For small (mainly residential) consumers this price must be published, but for larger (commercial or industrial) consumers this contract may be confidential.

The supplier must, of course, pay the transmission and distribution system owners for the delivery of the energy, thereby requiring not only metering in all parts of these systems but also an approved use-of-system charge as set by a government authority or a regulator. Finally the generators must collect revenue from the suppliers for their contracted energy as metered by an independent authority and assessed through an appropriate regulated formula. It is obvious that the measurement and collection of revenues is now a complex business requiring careful organisation and the taking of many millions of measurements each day for half-hour or hourly slots. Automated meter reading and formulaic calculation is absolutely necessary to reduce the costs of collection and billing if the final price of energy is to be restrained.

A continuing problem is the balance between introducing 'smart' meters with a communication channel between consumer and supplier or continuing with manual meter reading on a periodic basis perhaps interspersed with estimated readings. The former with all extra costs included can amount up to 20% of a residential consumer's bill although for larger consumers with bills of £1000 p.a. or more the cost of the new meter and associated communications can soon be recovered in the tariff. A meter reader can be expected to call on 100 to 150 customers a day dependent upon density of housing whereas a smart meter with two-way communication can be interrogated at least once a day and the payment notified immediately to the customer followed by direct debt to an agreed bank account. It was the practice to install pre-payment meters for poor paying or rapid turnover customers but the cash collection adds to costs and makes it difficult to give discounts and thereby reduce bills to the poorer members of society.

In developing countries where energy per household is extremely small, a fixed standing charge with fuse limitation of maximum demand is possible to save on collection costs. Where demand is for fixed and calculable times, such as street lighting etc., then an agreed monthly payment is preferable to avoid metering costs.

39.6 Environmental sustainable planning

No power system plan for a 'green-field' site or extension to an existing network can avoid careful and in depth consideration of the effect on the environment. It is expected that most governments have introduced or will introduce a strategy for reducing carbon dioxide and polluting gas emissions from electrical energy producers. This strategy not only requires energy producers to be more efficient than in the past but also expects the consumers—both large and small—to reduce considerably energy use of all kinds. Unfortunately, electrical energy is perceived to be the most polluting source because of the large power stations evident around the country and also because of the relative inefficiency of end use in all kinds of industrial plant, air conditioning and cooling, freezers, refrigerators and home heat losses. Estimates of possible reduction in CO_2 emission per kWh of electrical energy over the next 10 years are up to 20% and by careful design and encouraging awareness of the problem, more reductions could be possible.

Although pollutants produced by electrical energy production can be mitigated by design, the extra cost must be borne by the purchaser e.g. flue gas desulphurisation on a large power station can add 20% to its capital cost plus about 10% to the running costs, making it uneconomic compared to existing stations. One answer, now being introduced by many governments, is the purchase of an emission limit licence for each installation, which if not fully taken up each year can be traded like any other asset. The aim would be to control total emissions such that local atmospheric conditions could be stabilised and reduced.

For carbon dioxide emissions leading to global warming a similar 'carbon tax' on fuel use is proposed. Whilst the cost of CO_2 entrapment at source is tremendously expensive at present (+80% estimate for a power station) the tax would pay for more research and development leading eventually to power products such as renewables and fuel cells in which no CO_2 is emitted. The aim is to move towards non-polluting producers, including zero CO_2, and to use resources that will sustain the ecology of the planet and continually improve the quality of life for all.

References and further reading

1 WEEDY, B. and CORY, B. J., *Electric Power Systems*, Book-Wiley, 4th Edition (1998)

2 The National Grid Co. Plc., Seven Year Statements (published each year), NGC, National Grid House, Kirby Corner Rd., Coventry, CV4 8JY

3 *Modern Power Station Practice*, Volumes K & L, (*Power System Operation* & *EHV Transmission*), Books-Pergamon (1990)

40

Power System Operation and Control

H Glavitsch DrIng
ETH, Zurich

K Reichert DrIng
ETH, Zurich

F Peneder DiplIng
Asea Brown Boveri Cie, Baden

N Singh DrIng
ABB Power Automation AG, Baden

Contents

40.1 Introduction 40/3

40.2 Objectives and requirements 40/3

40.3 System description 40/4

40.4 Data acquisition and telemetering 40/6
  40.4.1 Introduction 40/6
  40.4.2 Substation equipment 40/7
  40.4.3 Data transmission 40/8
  40.4.4 Transmission system 40/9
  40.4.5 Communication channels 40/10

40.5 Decentralised control: excitation systems and control characteristics of synchronous machines 40/10
  40.5.1 Introduction 40/10
  40.5.2 Brushless excitation systems 40/10
  40.5.3 Static excitation systems 40/11
  40.5.4 Automatic voltage regulator and firing circuits for excitation systems 40/12
  40.5.5 Limiting the excitation of synchronous machines 40/12
  40.5.6 Control characteristics for synchronous machines 40/14
  40.5.7 Slip stabilisation 40/17
  40.5.8 Adapted regulator for the excitation of large generators 40/18
  40.5.9 Static excitation systems for positive and negative excitation current 40/20
  40.5.10 Machine models for investigating stability 40/21

40.6 Decentralised control: electronic turbine controllers 40/24
  40.6.1 Introduction 40/24
  40.6.2 The environment 40/24
  40.6.3 Role of the speed governor 40/25
  40.6.4 Static characteristic 40/26
  40.6.5 Parallel operation of generators 40/26
  40.6.6 Steam-turbine control system (Turbotrol) 40/27
  40.6.7 Water-turbine control system (Hydrotrol) 40/32

40.7 Decentralised control: substation automation 40/36
  40.7.1 Introduction 40/36
  40.7.2 Hardware configuration 40/36
  40.7.3 Software configuration 40/37
  40.7.4 Applications 40/37

40.8 Decentralised control: pulse controllers for voltage control with tap-changing transformers 40/38

40.9 Centralised control 40/39
  40.9.1 Hardware and software systems 40/39
  40.9.2 Hardware configuration 40/39
  40.9.3 Man–machine interface 40/40
  40.9.4 Wall diagram 40/41
  40.9.5 Hard copy 40/41
  40.9.6 Software configuration 40/42
  40.9.7 Memory management 40/42
  40.9.8 Input/output control 40/42
  40.9.9 Scheduling 40/42
  40.9.10 Error recovery 40/42
  40.9.11 Program development 40/42
  40.9.12 Inter processor communication 40/43
  40.9.13 Database 40/43
  40.9.14 System software structure 40/43

40.10 System operation 40/43

40.11 System control in liberalised electricity markets 40/44

40.12 Distribution automation and demand side management 40/44

40.13 Reliability considerations for system control 40/47
  40.13.1 Introduction 40/47
  40.13.2 Availability and reliability in the power system 40/47
  40.13.3 System security 40/48
  40.13.4 Functions 40/49
  40.13.5 Impact of system control 40/49
  40.13.6 Conclusions 40/49

40.1 Introduction

Power system operation and control are guided by the endeavour of the utility to supply electric energy to the customer in the most economic and secure way. This objective is underlined by the fact that the electric energy system is a coherent conductor-based system in which the load effects the generation without delay. Energy storage can only be realised in non-electrical form in dedicated power-stations, so that each unit of power consumed must be generated at the same time. It is therefore necessary to maintain all the variables and characteristics designating the quality of service, such as frequency, voltage level, waveform, etc.

The power system is operated continuously and extends geographically over wide areas, even over continents. The number of generators supplying the system and the number of components contributing to the objective are extremely high. However, the task of system operation is alleviated by the fact that there are not too many different types of component (if the intricacies of power-stations are excluded). In the transmission system the components that are quite easily managed are lines, cables, breakers, disconnectors, bus-bars, transformers and compensators. In the power-stations the generator commonly used is the synchronous machine. The prime mover is a turbine driven by steam or by water. Overlooking the details, there are similarities in the operation and control of turbines and generators that permit a certain uniformity of approach; however, many details have to be considered when it comes to design, failure modes, actual performance, quality of service, etc.

The mechanism of power flow in the system, which is a key issue in all considerations of operation under normal and disturbed conditions, is governed by the Ohm and Kirchhoff laws. Higher voltage at an appropriate phase angle will cause more power flow when a suitable path is available. To draw power from a system is extremely simple: the consumer's load is just connected to a three-phase bus-bar, where the voltage is regulated so that the load can be supplied. The system includes control actions by which the power is allocated to various generators, but there is little interaction in the transmission system.

For economic and secure operation, the actions are quite sophisticated. Control can no longer be performed in a single location by observing a single variable. The control system becomes a multivariable, multi-level system with real-time and prophylactic interventions, either manual or automatic. Thus, the control system is a hierarchical system where the levels and corresponding functions can be characterised as follows.

Decentralised control

(1) In the power-station: control of voltage, frequency, active and reactive power.
(2) In the substation: control of voltage (tap-changer on transformers), switching of lines and cables, protection.

Centralised control

(1) Regional control centre: switching, start-up and shut-down of generating units, unit commitment, reactive power control and load frequency control.
(2) Utility control centre: economic load dispatching, security assessment, load frequency control (automatic generator control), power exchange with other areas, security monitoring, security analysis and operational enhancement.

A hierarchical control system needs extensive communication. Hence there is an extensive telemetering and telecontrol system which provides data to the various control levels and executes control actions at a particular location that have been initiated in a control centre.

The following material is organised in such a way that the objectives and requirements of system operating and control are given and discussed first (Section 40.2): they underline the motivation for the development of modern complex systems. A way of presenting components and subsystems is given in Section 40.3. Data acquisition and telemetering are then treated (Section 40.4) these are prerequisites for power system control.

Decentralised control is divided into *excitation* control (Section 40.5), *turbine* control (Section 40.6) and *substation* automation (Section 40.7), which are the most important control functions at a local level. Pulse controllers for tap-changers are also considered (Section 40.8).

Centralised control is dealt with in Section 40.9, where the hardware and software aspects of computer-based system operation are considered. With the various systems and functions to hand, present-day system operation is characterised in Section 40.10. Changes in the system operation due to liberalised energy markets are also pointed out in the Section 40.11. The new focus on distribution automation and demand side management is touched upon in Section 40.12. Finally, the reliability of system control is considered in Section 40.13.

40.2 Objectives and requirements

The often-cited objectives of economy and security have to be considered in various time-scales and in different system conditions in order to achieve a systematic approach to power systems operation, particularly when computer-based systems are involved. Before we go into details, it should be noted that system operation as considered here is a problem within the framework of a given power system. Planning problems and problems of procuring the primary energy are omitted.

Any objective or requirement is derived from the basic task of supplying electric energy to the consumer with the least expenditure of economic effort, measured over a long period. Hence a variety of cost items and even some intangibles such as environmental effects are involved. These items range from generating costs, losses, the cost of outages and the cost of damages, to risks and the amount of emissions, etc. Stated thus, these items are still too general to be converted immediately into an objective upon which a control function could be built. Different objectives apply for decentralised and centralised control; moverover, within a centralised control system it is necessary to distinguish various conditions to which particular objectives apply.

The most widely used concept for the realisation of a systematic control approach in centralised control is the concept of *states*. A state of the power system is characterised by reserves, by the ability of the system to override a disturbance, by the presence or absence of overloads, etc. The starting point is a series of considerations concerning the security of the system. However, considerations of economy can be easily added.

The four states with some rough characterisations are as follows:

(1) *Normal* (*N*) *state*: the system has no overloads and a good voltage profile; it can withstand a line or generator outage and is stable.
(2) *Vulnerable* (*V*) *state*: the system has no overloads and a good voltage profile; line or generator outage causes

overloads and/or voltage droop; there is a low stability margin.

(3) *Disturbed* (*D*) *state*: the system still supplies its loads, but overloads are present and there is a low voltage profile.

(4) *Emergency* (*E*) *state*: the system cannot supply all of its loads, overloads are present, there is a low voltage profile and part of the system is disconnected.

There are inadvertent transitions between the states caused by faults, human error and dynamic effects. However, control actions will return the system to the normal state. A thorough understanding of system operation will show that appropriate objectives and corresponding control functions will effect this return in a logical manner. These mechanisms are illustrated in the schematic of *Figure 40.1*, which shows (upper part) a state diagram with various transitions. The lower part of the figure gives transitions associated with a number of objectives. These pertain to control actions only. It is clear that the objective of full load coverage has to be met first before the vulnerable state can be reached. Further, all vulnerable conditions have to be eliminated before economic dispatching can be initiated.

At the decentralised level the objectives are much simpler and easier to understand. A state concept is not necessary since the objectives are expressed in terms of errors or time sequences in a straightforward manner. As an illustration, let us consider protection and excitation control.

In protection the objective is to keep fault duration or outage to a minimum. The fault itself cannot be avoided, but the adverse effects, possibly leading to a disturbed or emergency state, can be. Thus protection supports the aim of security on the decentralised level.

Excitation control maintains the voltage at a given location of the system and supports the stability. It is the stable voltage, with all its implications, that is at stake. A stable voltage is a very important prerequisite for system security. For its realisation many detailed considerations are required since it is system dynamics that determines the performance of excitation control. Beyond that, an excitation system has many monitoring functions, i.e. limit-checking and even protective functions. The objectives of decentralised control, which become recognisable in terms of set-points, time periods and the like, must be co-ordinated. This co-ordination is performed either in the planning or operational planning phase (e.g. for protective relays), or in real time (e.g. for economic dispatching).

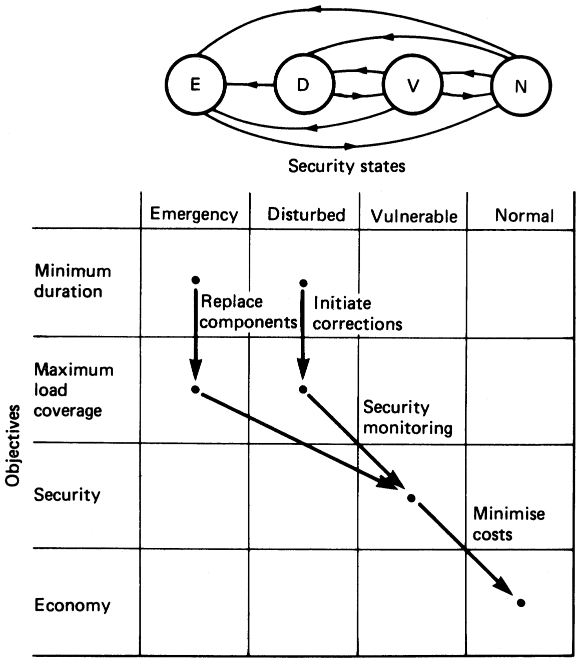


Figure 40.1 Security states and objectives: state diagram (above) and controlled transitions (below) following given objectives

In power system operation and control it must be recognised that the control functions have a certain time range in which they are effective. Thus the corresponding objective has its validity within this time range only. As *Figure 40.2* shows, there is a complete hierarchy of functions ordered in terms of their effective time range. This hierarchy in time is also responsible for the functional hierarchy of the control system.

40.3 System description

In describing an electric power system we must distinguish the network (responsible for transmission and distribution) and the generating stations, as well as the loads.

The network is amenable to description by topological means such as graphs and incidence matrices as long as the connection of two nodes, by a line or cable, is of importance. Thereby it is implied that the connection is made by a three-phase line which is itself described by differential equations. This topological description is always inherent and necessary. In routine work, however, it may not always be obvious what the true background is. Hence it is worthwhile to consider these topological means as a starting point.

Graphs and incidence matrices are equivalent and describe the way in which nodes are connected. A graph is a pictorial representation, whereas the incidence matrix is a mathematical formulation which can be interpreted by a computer. An example will illustrate the correspondence between these two descriptions. *Figure 40.3* shows a graph derived from an electrical network. The connections between nodes have been marked by arrows; this fixes the way of counting and thereby orientates the graph. *Figure 40.4* is the incidence matrix corresponding to the graph. The way of counting is indicated by the sign of the entries. Both the graph and the matrix describe the same network. The graph, however, is much easier to grasp.

The incidence matrix is the base for two processes, both of which are important for systematic description. The first is the reading of network data, wherein the way in which lines are connected is already implied. The second is the establishment of admittance and impedance matrices.

Consider the graph in *Figure 40.3* where the nodes have a particular numbering. This numbering permits the specification not only of a connection between two nodes but also of the type of connection. Assume that the connections consist of simple series impedances. Then the following description, which is machine-readable, is possible:

Connection	Impedance	Connection	Impedance
1–2	$0.05+j0.68$	3–6	$0.06+j0.75$
2–3	$0.06+j0.75$	3–5	$0.03+j0.35$
1–4	$0.03+j0.35$	4–5	$0.04+j0.50$
1–5	$0.04+j0.50$	5–6	$0.05+j0.65$
2–5	$0.05+j0.70$		

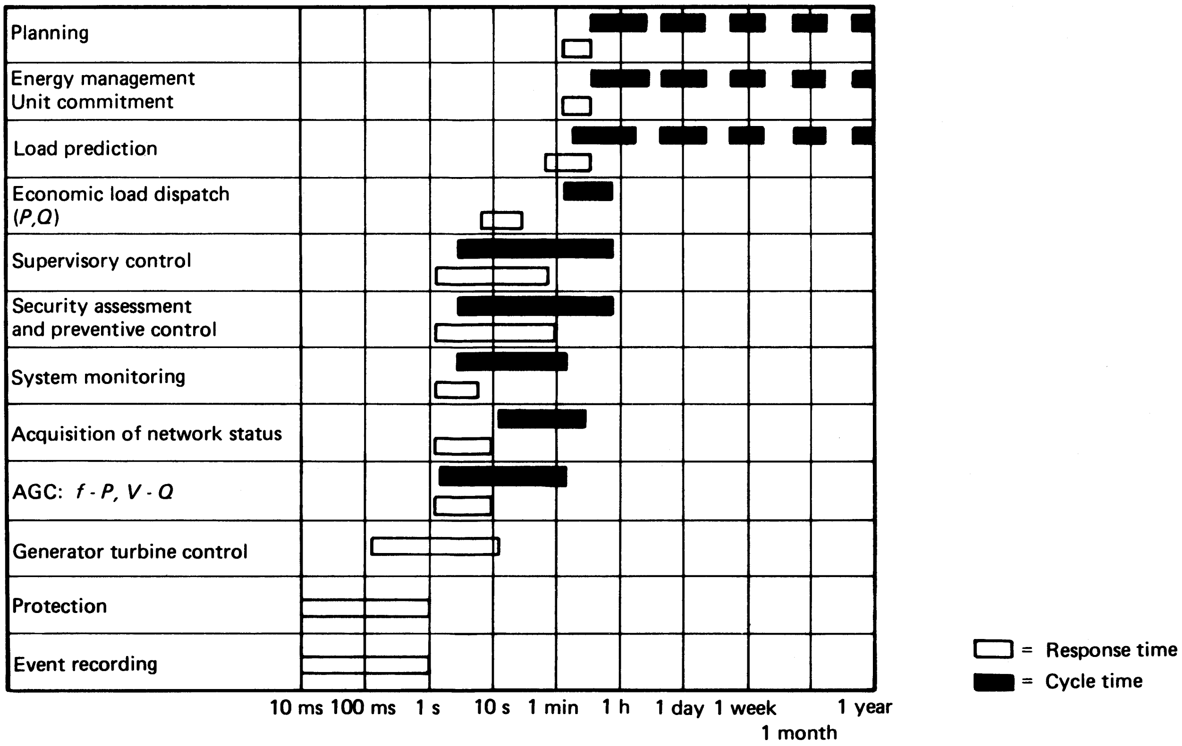


Figure 40.2 Response times and cycle times for plant management, monitoring protection and control functions. AGC, Automatic generation control; *P*, active power; *Q*, reactive power; *f*, frequency; *V*, voltage

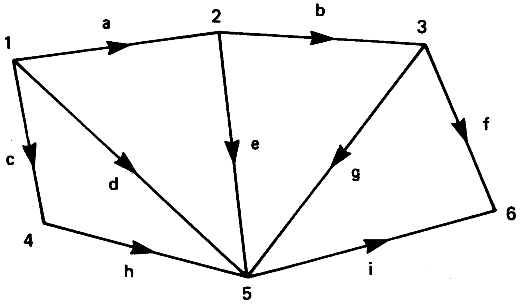


Figure 40.3 Oriented graph describing the structure of a six-node network: 1–6, nodes; a–i, lines

	1	2	3	4	5	6
a	+1	−1				
b		+1	−1			
c	+1			−1		
d	+1				−1	
e		+1			−1	
f			+1			−1
g			+1		−1	
h				+1	−1	
i					+1	−1

Figure 40.4 Nodal incidence matrix **A** corresponding to the graph in *Figure 40.3*. The matrix establishes relationships between line and nodal quantities (currents and voltages). The arrow leaving a node determines the positive sign of the entry (+1)

The first column specifies the nodal connection, the second the complex impedance. Each line of the list corresponds to a connection or a line of the network.

The generation of the nodal admittance matrix can be explained by a multiplication of the primitive admittance matrix $_{\mathrm{p}}$(diagonal matrix) by the corresponding incidence matrix and its transpose from left and right respectively:

$$Y = A^{T} Y^{p} A$$

The nodal admittance matrix contains all the information necessary (topology, impedance) to describe the network.

Loads and generators are connected to the nodes of the network. The description of the nodal constraints depends on the way the network is employed in an analysis or decision-making process.

For the purposes of load-flow calculations the specification is done in terms of *PQ*- or *PV*-nodes, where *PQ* means that the active power *P* and reactive power *Q* at the node are constant (i.e. independent of voltage), and *PV* means constant active power and constant voltage.

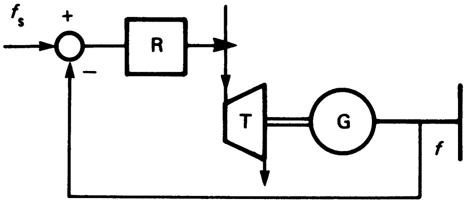


Figure 40.5 Schematic diagram of a generator and a speed governor: G, generator; T, turbine; R, regulator/speed governor; *f*, measured frequency; f_s, set-point of frequency

For dynamic analysis more complex models for the nodal description have to be added. On the load side, the frequency dependence or, if necessary, a set of different equations is needed. For the generating unit, differential equations are comprehensive, but not always transparent. Hence, a block diagram is used to specify forward and feedback paths, wherever appropriate. As an example, the signal flow and the generation of torque within a turbogenerator are given by the block diagram in *Figure 40.5*. The dynamics of a synchronous machine would also be amenable to such a description, but in practice a description by differential equations (two-reaction theory) is usually preferred. Details of controllers and regulations are described by block diagrams, as is common in control engineering.

40.4 Data acquisition and telemetering

40.4.1 Introduction

To be able to supervise and control a power system, the network control engineer requires reliable and current information concerning the state of the network. This he obtains from the power flows, bus voltages, frequency, and load levels, plus the position of circuit-breakers and isolators; the source of these data being in the power stations and substations of the network. As these are spread over a wide geographical area, the information must be transmitted over long distances. Thus the electrical parameters and switch states must be converted into a suitable form for transmission to the control centre. A telecontrol system is also required to transmit these data using communication channels which were historically limited in capacity and were shared with other facilities such as speech, protection and, perhaps, also telex communication. Speed of data transmission was therefore often limited, and the telecontrol network was configured to optimise the use of the available bandwidth on the carrier channel. The modern communication is based on fibre optical or wireless communication networks, which provide a powerful broad band communication. The classical power line carrier communication technology has also been improved with the use of digital technology. It is almost mandatory that all the new HV and EHV lines be laid with fibre optical core in the ground wire, as the incremental cost is quite low. The trunk routes may have dedicated logical links for speech, data and protection signal in addition to the data and voice network for telecommunication purposes.

The telecontrol system comprises a master station communicating over communication channels with remote terminal units (RTUs) located in the power stations and switching stations. Normally the RTUs are quiescent, i.e. they only send data after a direct interrogation from the master station; thus more than one RTU can be connected to a transmission channel as the channel can be time shared. A telecontrol network can be configured as a point to point, star (radial) or multi-point (part line) system as illustrated in *Figure 40.6*. The transmission can be either duplex, half-duplex or simplex (in the case where data traffic is unidirectional). The newer generation of RTUs can remain quiescent and can report the data on their own initiative using a slave to master communication protocol. The use of standard telecontrol protocols is also being promoted to facilitate use of equipment from different suppliers. International Electrotechnical Commission (IEC) has recommended IEC870-5-xxx group of protocols, which allow certain amount of openness in the system. The state of the art RTUs can communicate over data communication networks providing a large bandwidth with good reliability.

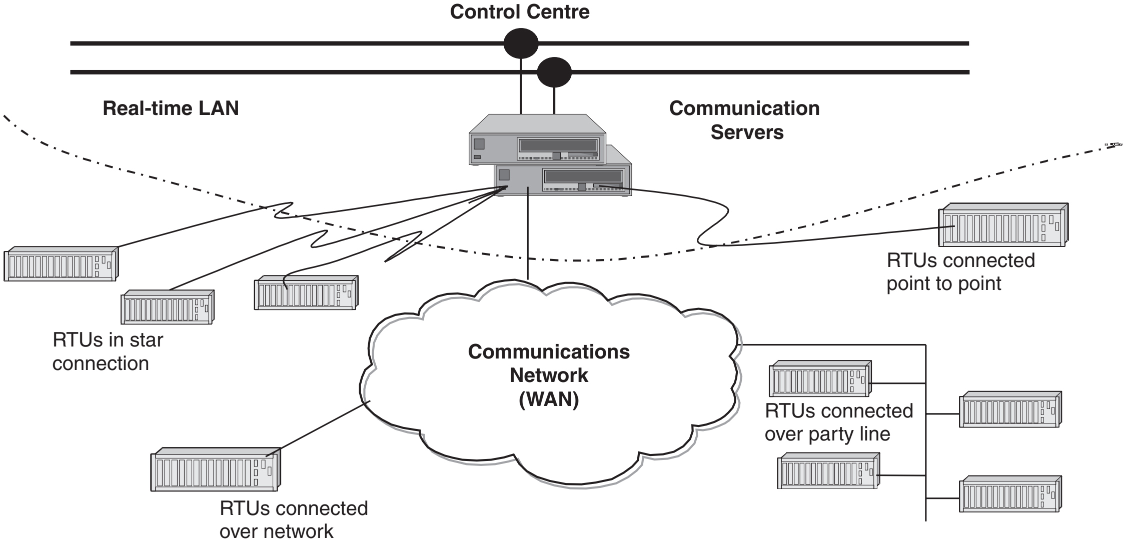

Control Centre
Real-time LAN
Communication Servers
RTUs connected point to point
RTUs in star connection
Communications Network (WAN)
RTUs connected over party line
RTUs connected over network

Figure 40.6 Transmission network configuration

40.4.2 Substation equipment

The data-acquisition equipment in a substation is usually a microprocessor-based RTU, equipped with both programmable read-only memory (PROM) and random-access memory (RAM). The data-acquisition program and the transmission program are loaded into the PROM, which is non-volatile (not corrupted during a power supply failure). All the system modules are connected to the system bus (see *Figure 40.7*) which is structured to facilitate data transfer from the information source to its destination, under the control of the program. The microprocessor function is monitored by the watch-dog module. The RTU is configured to suit the substation requirements. A typical configuration is shown in *Figure 40.8* with the input/output for one high-voltage (h.v.) circuit only. The input modules fall into three main categories: digital input, analogue input and pulse input.

40.4.2.1 Digital inputs

Digital input modules are used for inputting contact states indicating breaker positions, isolator positions and alarm contact operation. The contact states are continually monitored but are not transmitted unless a contact state has been changed. Under program control, the module is cyclically interrogated to see whether it has a change of state to report: if the reply is negative, the program moves to another module; if it is positive, the change of state is

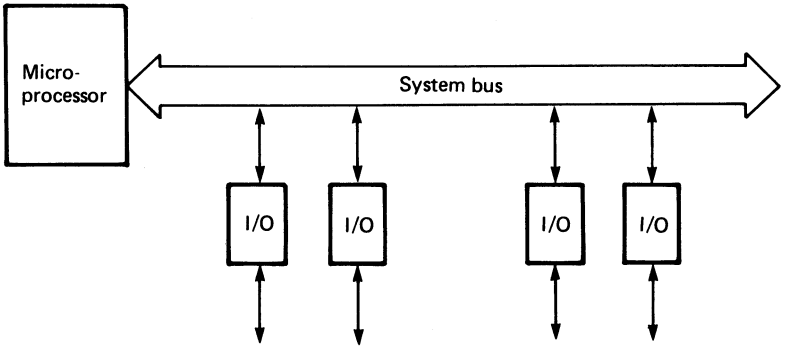

Micro-processor
System bus
I/O
I/O
I/O
I/O

Figure 40.7 RTU bus structure: I/O, Input/output

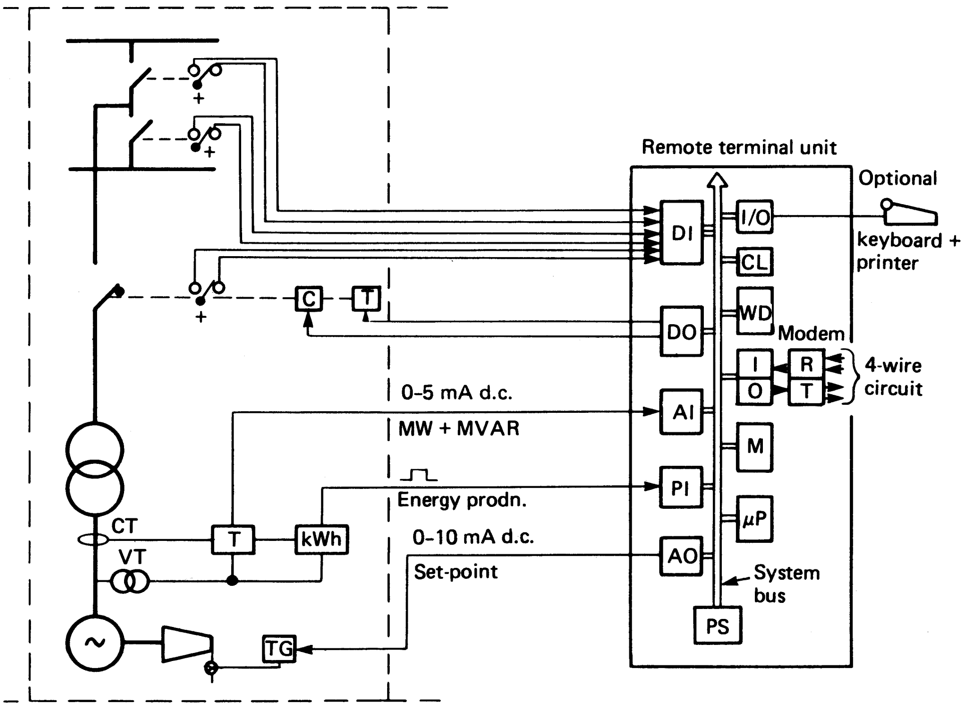

+
+
+
C
T
0–5 mA d.c.
MW + MVAR
Energy prodn.
0–10 mA d.c.
Set-point
CT
VT
T
kWh
TG
Remote terminal unit
Optional
keyboard + printer
DI
DO
AI
PI
AO
I/O
CL
WD
Modem
I
O
R
T
4-wire circuit
M
μP
System bus
PS

Figure 40.8 Example of the configuration of a RTU: DI, digital input; DO, digital output; AI, analogue input; PI, pulse input; AO, analogue output; μP, microprocessor; M, memory; CL, clock (optional); I/O, serial input/output interface; WD, watch-dog; PS, power supply; TG, turbine governor; CT, current transformer; VT, voltage transformer; T, transducer; CT, close, trip; R/T, receiver/transmitter

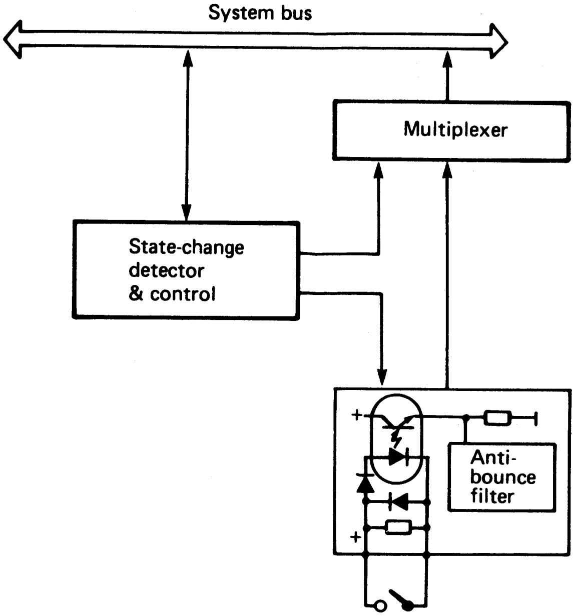


Figure 40.9 Digital input circuit

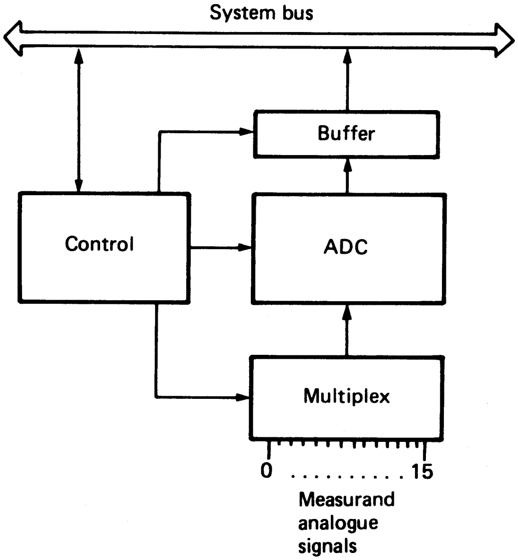


Figure 40.10 Analogue input module

written to the memory. The input circuits are decoupled from the power system equipment by electromagnetic relays or opto-couplers and are filtered to eliminate the effect of contact bounce. A digital input circuit is illustrated in *Figure 40.9*.

40.4.2.2 Analogue inputs

The electrical measurements are converted from current-transformer and voltage-transformer levels to analogous direct-current (d.c.) milliampere signals by suitable transducers. These convert voltage, current, active and reactive power, frequency, temperature, etc., into proportional d.c. signals which are fed into an analogue input module. This module (*Figure 40.10*) has a multiplexer for scanning multiple inputs (usually 16) and an analogue-to-digital converter (ADC) which converts the d.c. signal into a digital code (usually binary or BCD). The local scanning rate of the measurands is less than 0.5 s, with an accuracy of conversion for normal requirements of 1% for bipolar (7 binary bits + sign) values and of 0.5% for unipolar values. For higher accuracy requirements an accuracy of 0.1% is used (11 bits).

40.4.2.3 Pulse inputs

Energy or fuel consumed can be measured by integrating meters giving a proportional pulse output. The pulses are integrated in a pulse input module which is periodically scanned, e.g. at 15 min intervals. The consumption over that period is then the difference between the last two readings. Special measures are taken to ensure that the reading does not interfere with the pulse integration or vice versa.

40.4.2.4 Digital output

Digital output modules convert a coded message, received from the control centre, into a contact output, e.g. to trip or close a circuit-breaker or to raise or lower a tap-changer. Obviously, the transmission of these signals must be very secure, and checking that the correct output is given must be rigorous. Echo checks are made to ensure that the correct code has been received by the module, and additional circuits can be added to detect stuck contacts and to prevent more than one command being output at a time.

40.4.2.5 Analogue output

Set-points are output via analogue output modules that receive a coded message which they convert into a d.c. analogue signal; this is output e.g. as a turbine-governor setting.

40.4.2.6 Clock

An optional feature of the RTU is a clock module, which enables the events in the substations to be arranged and transmitted in chronological order and the time at which an event happened to be added. The clock is capable of giving a 10 ms time resolution in event sequence recording and the time in hours, minutes, seconds and hundredths of a second.

40.4.2.7 Printer

A further option can be included to print the alarm and event sequences locally at the substation by adding memory, serial–parallel interface, a printer and additional program. If the memory capacity is limited, the output is either coded or abbreviated; however, with adequate memory the print-out can be in clear text.

40.4.3 Data transmission

The substation data are locally scanned and memorised in the RTU ready to be transmitted to the control centre. When an instruction is received from the master station, the processor prepares to send the data to the control centre. To every message containing data, it adds check bits that enable the master station to ensure that the message has not been corrupted during transmission. Similarly, to any message output from the master station, check bits are added to enable the RTU to verify that the received message has not been corrupted. The data are then output to the communication channel via the parallel-to-serial converter and the modulator–demodulator (modem) which converts d.c. pulses into keying frequencies in the voice-frequency (v.f.) range. Normally two v.f.-range frequencies are used to represent '1's and '0's of the message code and to modulate the carrier of the transmission channel. At the receiving end, the modem converts the keying frequencies into d.c. pulses and the serial–parallel converter module reassembles the message into a parallel word including the check bits. The message can then be checked for errors induced during transmission.

There are various methods of checking for errors in a message telegram. The error-detecting capability of a checking system is often described as a Hamming distance. The Hamming distance between two binary words is the number of bit positions by which they differ, which is undetectable by the error-checking system. For most purposes in telecontrol, the minimum Hamming distance is 4, i.e. all 3-bit errors are detected. An example of a modified Hamming code error-checking system is shown in *Figure 40.11.*

The advantage of this system is that in addition to detecting the same number of error patterns as equivalent error-checking systems of the same size, the introduction of odd parity in one check-bit position and even parity in the others forces at least two changes in bit settings in the telegram. This considerably reduces the risk of loss of synchronism when the word is being read. The Hamming distance of this code is 4: i.e. all patterns of 3-bit errors are detected, as well as all odd numbers of bit error patterns and over 92% of all even numbers of bit error patterns.

40.4.4 Transmission system

The data transmission system must be designed to allow all the RTUs to transmit their data to the control centre in a reasonable time without imposing impractical demands on the communication system. The data transmission can be divided into several parts, each of which can comprise one or more RTUs. Each part system is completely independent of the others; thus they effectively operate in parallel.

Three basic systems are possible: point-to-point, radial or multi-point (see *Figure 40.6*). Combinations of radial and multi-point are also possible.

With point-to-point systems, the transmission can be simplex (where the RTU sends data continuously to the master station) or half- or full-duplex (where data can be transmitted in both directions).

With radial or multi-point systems, the dialogue between the master station and the RTU is entirely controlled by the master station to avoid two or more RTUs sending data simultaneously. The master station transmits instructions to the RTU, which, in complying with the instruction, acknowledges it. The RTU takes no initiative in transmitting data and sends data only after receiving a specific instruction to do so from the master station. Thus all data in a part system are transmitted from each RTU in turn and response times are a limiting feature of a part system configuration.

There are two basic forms of data to transmit: even-controlled data which appear at random intervals, and data that must be transmitted cyclically as they are liable to be continually changing. The cyclically transmitted data, e.g. measurands which require frequent updating at the control centre, take up most of the transmission time; however, as the RTU is quiescent, periodically it must be asked whether it has any spontaneous data to transmit. To save transmission time, this is made in a 'broadcast' interrogation to all RTUs in the part system simultaneously. If no RTU answers positively, the normal cycle sequence is resumed; however, if one or more RTU answers that it has spontaneous data to send, the normal cycle is interrupted and each RTU is interrogated in turn until all the spontaneous data have been transmitted and received at the master station. In order that the detection of spontaneous data (e.g. a breaker trip) is transmitted in a reasonable time, the spontaneous data interrogation is interjected between the transmission of groups of measurands (say 16). An example of a transmission sequence is given in *Figure 40.12.* The spontaneous data can be sent in chronological order of events taking place and also with the time at which each event happened.

The response times are dependent on the transmission speeds, and a choice of speed is available. Standard speeds of 50, 100, 200, 600 and 1200 bit/s are available within the v.f. band. The choice of speed is then dependent on the response times required and the bandwidth available on the communication channel. As an example of response times, a part system comprising four RTUs each having measurands would have a measurand and indication interrogation scan time (with no changes of state to transmit) of approximately 6.5 s, using a 200-baud channel. To decrease this time, either the number of RTUs in the part system

|← Message telegram →|

Bit No.	15	14	13	12	11	10	9	8	7	6	5	4	3	2	1	0
Bit significance	M2	M1	L	D7	D6	D5	D4	$\overline{H4}$	D3	D2	D1	H3	D0	H2	H1	H0
$\overline{H4}$	×	×	×	×	×	×	×	●								
H3	×	×	×	×					×	×	×	o				
H2	×	×			×		×			×	×		×	o		
H1	×		×		×		×		×		×		×		o	
H0	×	×	×	×	×	×	×	×	×	×	×	×	×	×	×	o

Figure 40.11 An example of message telegram organisation: ●, odd parity check; o, even parity check; ×,⇐bits supervised by Hamming bits; D, data bits; L, message sequence complete; M, message-type definition; H, Hamming bit

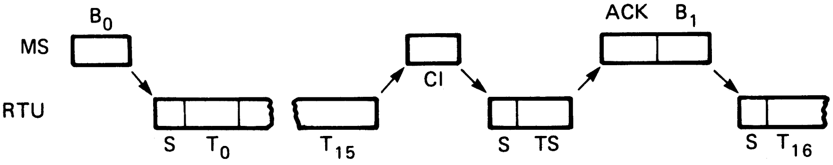


Figure 40.12 A typical transmission sequence: B_n, call for transmission of a group of measurands; S, start character; T_n, values; CI, call for transmission of a change of state; T_s, change-of-state message; ACK, acknowledgement of receipt of meassage; MS, master station

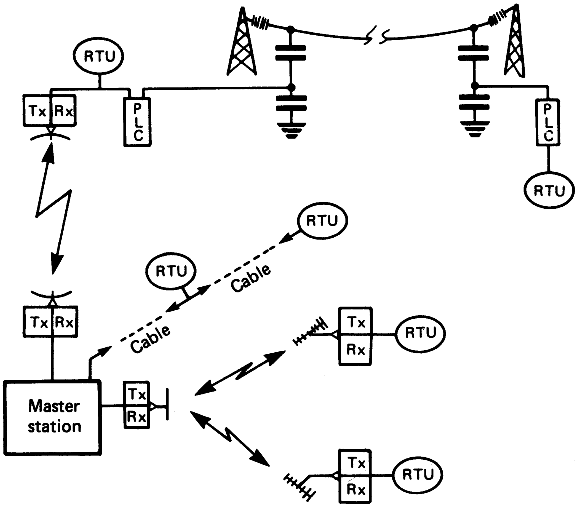


Figure 40.13 An example of a data transmission system. Tx, Rx, radio transmitter and receiver; PLC, programmable logic controller

would have to be reduced or the speed of transmission would have to be increased.

40.4.5 Communication channels

The transmission of the frequency-shift keying signals requires a four-wire circuit communication channel, which can be telephone-type cable, power line carrier equipment, radio transmission or any combination of these media in series or parallel. *Figure 40.13* gives an example of a possible combination of different types of communication channels.

Often the telecontrol signal must share the v.f. band with other transmissions such as speech, telex, protection or even another telecontrol part-system transmission. An example of the frequency multiplexing of the v.f. band is given in *Figure 40.14*. The multiplexing equipment is normally available with the power line carrier. The frequency-shift keying takes place within the frequency bandwidth indicated in *Figure 40.14*: e.g. for a 200-baud modem the lower keying frequency is −90 Hz and the upper is +90 Hz of the channel centre frequency. The total bandwidth required is 360 Hz.

40.5 Decentralised control: excitation systems and control characteristics of synchronous machines

40.5.1 Introduction

A synchronous generator operating on an interconnected grid requires a fast-response excitation system to ensure its stable operation. This system must be able to adjust the level of magnetic flux in the generator according to the grid network requirements. Since the parameters of the generator will fall into a fairly narrow range according to the type of generator (salient pole or turbo), the excitation system must be designed to bring the generator flux (i.e. excitation current) to the required level in the shortest possible time, notwithstanding the long time-constants of the generator transfer functions.

Brushless excitation equipment has been developed to meet these requirements on various types of synchronous generator—particularly on smaller turbogenerators—and has given excellent results in service. The control characteristics of this system, however, are dictated largely by the exciter and the rotating diodes rather than by the voltage regulator (d.c. exciters are not used in modern excitation systems).

For large generators a static excitation system represents the best solution since, in principle, it imposes no limit on the ceiling voltage. In the per-unit system the ceiling voltage is the ratio of the maximum excitation voltage to the excitation voltage for nominal open-circuit voltage at the generator terminals. Values of 10 or 15 p.u. are quite often used. A great advantage of static excitation is that the total field-voltage range is available with practically no time delay. This speeds up not only field forcing but also field suppressions when required, by change-over to inverter mode.

40.5.2 Brushless excitation systems

Brushless excitation systems are mainly employed either where the control requirements are not too stringent or where the atmosphere is chemically aggressive. The main generator excitation is provided by an auxiliary exciter generator mounted on the main generator shaft.

This exciter is a salient-pole synchronous generator with a three-phase winding on the rotor and a d.c. excitation winding on the stator. The alternating currents produced by the rotor winding are rectified by rotating diode bridges mounted on the shaft, the resulting d.c. being fed to the field winding of the main generator. The excitation for the exciter stator winding is supplied by a thyristor regulator which comprises voltage regulators and limit-value controllers. Field suppression of the exciter machine is also carried out by the regulator (*Figure 40.15*).

The controlled rectifiers and the electronic circuitry of the regulator are supplied either from a permanent magnet (p.m.) generator on the main shaft or through a transformer fed from the main generator terminals. In the latter case, compounding equipment is available for the required selective switching (short-circuit duration in excess of 0.5 s) which is then combined with the standard regulators.

In its basic form the regulation system contains the continuously active elements for automatically controlled operation, which are:

(1) a voltage regulator;
(2) a reference-value potentiometer for remote adjustment;

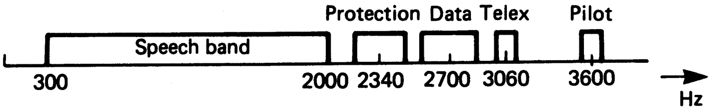


Figure 40.14 Allocation of communication functions to a 4 Hz channel

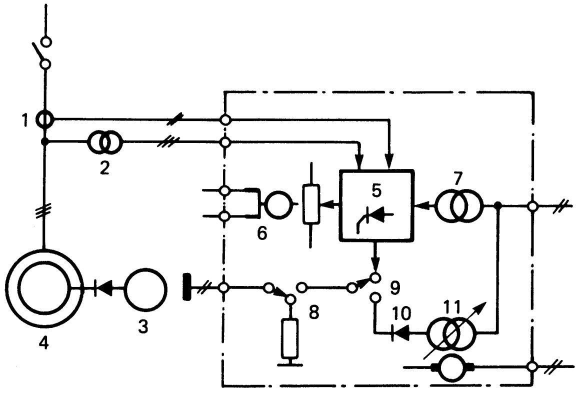


Figure 40.15 Voltage regulation system with manual control facility: 1, current transformer; 2, voltage transformer; 3, exciter; 4, generator; 5, voltage regulation; 6, reference value; 7, supply transformer for the voltage regulator; 8, field switch/discharge resistor; 9, change-over switch; 10, rectifier; 11, variac with motor drive

(3) a supply transformer;
(4) a field-breaker and discharge resistors; and
(5) diode monitoring consisting of the manual control device, i.e.
  (a) a variac transformer,
  (b) a diode bridge, and
  (c) a change-over switch (auto to manual).

40.5.3 Static excitation systems

On generators with difficult regulation requirements a rotating exciter should not be used. Here the excitation is supplied either directly from the generator terminals via a transformer and a controlled rectifier bridge (*Figure 40.16*) or by an auxiliary generator.

The main components of a static excitation system are:

(1) an excitation transformer (or auxiliary generator);
(2) a static converter;
(3) a field suppression system;
(4) control and firing circuits; and
(5) a protection and monitoring system.

The capacity of the excitation system is dependent on the generator specification and on the requirements of the power system. The rating of the excitation transformer or exciter in particular is determined by the maximum continuous current rating of the rotor winding and the required ceiling voltage. In designing the converter circuits, which have much shorter thermal time-constants, the ceiling current and short-circuit current capacity must also be considered.

The voltage of the transformer secondary (or of the auxiliary generator) is determined by the maximum ceiling voltage, V_c of the excitation system, and its current by the maximum continuous current I_{fm} of the generator field winding. A simple approximation to the rating of a three-phase excitation transformer is $1.35\ V_cI_{fm}$.

The static converter comprises one or more fully controlled thyristor bridges, which are almost always cooled

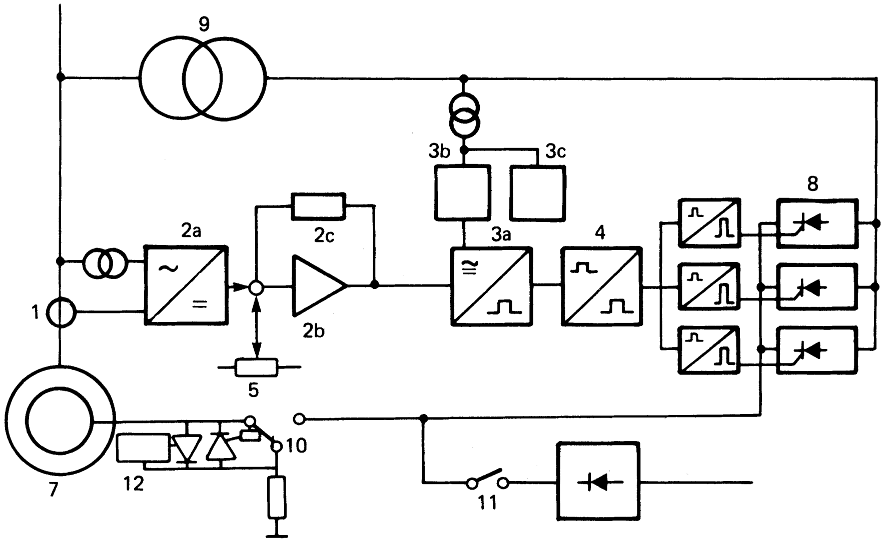


Figure 40.16 Block diagram of a voltage regulator with static excitation (without manual control): 1, instrument transformer; 2a, transducer; 2b, voltage regulator; 2c, PID filters; 3a, firing-angle control; 3b, filter; 3c, voltage relay for control of excitation field flashing; 4, pulse amplifier; 5, reference-value setting unit; 7, generator; 8, main excitation rectifier; 9, excitation transformer; 10, field suppression device; 11, field flashing device; 12, overvoltage protection

by forced air. The number of parallel bridges and the number of series elements in each bridge branch are determined by the technical data (current rating and inverse voltage) of the thyristor bridges, the maximum exciting current, the transient overload (due, for example, to a system fault) and the ceiling voltage. An extra redundant bridge is normally included (in parallel) so that, even if one bridge fails, the full operational requirements can still be met. To ensure selectivity, it is necessary to have at least three bridges in parallel. This is because a faulty thyristor together with a healthy branch can form a two-pole short circuit that persists until the fuse in series with the defective element blows. If only two parallel bridges are provided then the short-circuit current in the healthy branch is shared by only two thyristors; the fusing integral of both the corresponding fuses may then be exceeded, resulting in an excitation trip.

The rectifier can have either a block structure or be arranged phase by phase. In the block arrangement, one or more bridges are grouped together and gated from a common pulse amplifier. Blocks are normally provided with individual fans, but common cooling is also possible.

The arrangement of the converter in phase groups reduces the risk of an interphase short circuit. All parallel thyristors of one phase are brought together in a single rack. A separate cooling system for each rack is inappropriate in this arrangement and common cooling is employed, consisting of two fans each capable of supplying the full cooling-air requirement.

With the phase-by-phase arrangement, the excitation transformer is usually composed of three one-phase units; this will always provide a higher degree of safety against interphase short circuits.

The field suppression system comprises essentially a two-pole field-breaker and a non-linear field discharge resistor. The component ratings are so chosen that neither the permissible arcing voltage at the breaker nor the maximum permissible field-circuit voltage is exceeded, even with the maximum possible field current (following a terminal short circuit on the generator). A 'crowbar' is included in the discharge-equipment cubicle as additional overvoltage protection. This comprises two antiparallel thyristor groups which are triggered by suitable semiconductor elements such as breakover diodes (BODs). If, for instance, a voltage is induced in the rotor field circuit owing to generator slip, this voltage can rise only to the threshold of the BOD element. If this value is reached, the element triggers the thyristors, and the rotor circuit is connected to the discharge resistance, allowing a current to flow which immediately causes the induced voltage to break down.

40.5.4 Automatic voltage regulator and firing circuits for excitation systems

The voltage regulator for brushless exciters as well as for static excitation systems constitutes the central control element in a modern plant. This unit uses a voltage transformer to measure the actual value, rectifies it and compares it with the reference value. The difference-voltage is fed to an amplifier with a lead–lag filter (the response characteristics of which must be carefully adjusted to suit the particular generator and grid parameters) and the amplified value is brought to the gate control unit. Using additional reactive current compensation it is possible to adjust the reactive current behaviour of the generator with regard to the network.

The voltage regulator has normally to be supplemented with parallel-connected limiters which function before corresponding generator protection relays are activated. Depending on requirements, rotor-current, load-angle and stator-current limiters may be employed. Their features are described later.

If the limitation control causes a reduction in the excitation current (on over-excited operation), time-delay elements are incorporated to allow high transient currents. For underexcited operation there is no time delay.

Stabilising equipment can be used to damp power oscillations that may arise under exceptional network conditions.

The voltage regulator may be supplemented with an overriding system for controlling the power factor or the reactive power flow, as required. All voltage regulators are provided with manual control (changeover from closed-loop to open-loop control). A separate redundant gate control unit is often provided (double-channel type). In case of a fault in the 'automatic' channel, a follow-up control system ensures a smooth transition to manual operation.

The amplified difference signal is transformed into pulses with the appropriate firing angle by the gate control unit. The firing pulses are amplified in an intermediate stage and led to the individual output stages which are assigned to the various converter units. The output stages shape the pulses to the steep slopes necessary to ensure simultaneous firing of all parallel thyristors. The pulses from the output stages are fed via impulse transformers to the thyristor gates.

The transfer function of the automatic voltage regulator (AVR) including the gate control system is

$$\frac{\Delta V_{\mathrm{f}}}{\Delta V_{\mathrm{g}}} = A\,\frac{(1+\mathrm{p}T_1)(1+\mathrm{p}T_2)}{(1+\mathrm{p}T_3)(1+\mathrm{p}T_4)}\,\frac{1}{1+\mathrm{p}T_{\mathrm{M}}}\,\frac{1}{1+\mathrm{p}T_{\mathrm{E}}}$$

where T_{M} is the measuring time-constant, T_{E} is the exciter time-constant, $T_1, \ldots, T_4$ are the equivalent lead–lag time-constants, A is the amplification, and p is the operator $\mathrm{d}/\mathrm{d}t$.

40.5.5 Limiting the excitation of synchronous machines

Ever-increasing demands on power system reliability led to the development of ancillary equipment (limiters), to inhibit the tripping of protection gear when this was not wanted and to make, within the permissible limits, better use of the synchronous machine. These limits are clearly depicted in the power chart of a turbogenerator shown in *Figure 40.17*.

The limit of active power output (line CE) is determined by the prime-mover, and is disregarded. A factor connected directly with excitation current, however, is the permissible temperature rise in the rotor winding, represented by the arc CD with centre A corresponding to the non-excited condition. This thermal limit is defined by the ageing of the insulation. It may therefore be exceeded for a short time—a requirement essential for the stability of the power system.

In the underexcited mode, operation of the synchronous machine is limited by the airgap torque necessary for the active power transfer, the steady-state condition being represented by a definite, permissible rotor displacement angle (line AE). This limit has mechanical and dynamic characteristics, and calls for instantaneous intervention as soon as it is exceeded, to prevent the generator from falling out of step.

Another important condition governing the limits to be introduced is the requirement that these do not interfere with the normal operation of the voltage regulator, and that neither intervention nor return to voltage regulation leads to disturbances.

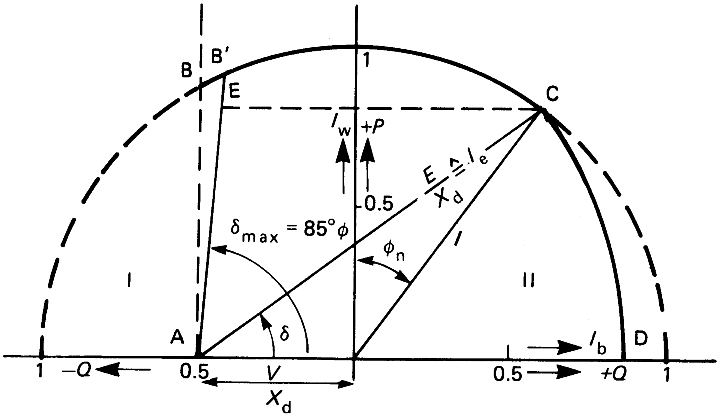


Figure 40.17 Current and power diagram of a turbogenerator ($I = 1$ p.u.; $V = V_k = 1$ p.u.; $X_d = X_q = 2$ p.u.; cos $\phi = 0.8$): I, underexcitation; II, overexcitation. AB, Practical stability limit; AB′, steady-state stability limit; BC, limit of stator temperature rise; CD, limit of rotor temperature rise; CE, active power limit; E rotor e.m.f.; I_b, reactive current; I_e, excitation current; I_w, active current; P active power; Q reactive power; V_k, terminal voltage; X_d, direct-axis reactance; X_q, quadrature-axis reactance; δ, load angle; ϕ_n, rated phase angle

Electronic load-angle and current-limit controllers were introduced some time ago, and these are now standard components of all good voltage regulation equipments. Practical experience has made possible the addition of improvements and refinements.

Consider the static excitation equipment commonly used for large generators (*Figure 40.18*). Normally, the voltage regulator holds the generator voltage (and thus, indirectly, the reactive power output) at a constant level. The static converter adjusts the excitation current so that the reference voltage remains a function of the reactive current. At the same time, the rotor-current limit controller measures the excitation current in the static converter supply and compares it with the relevant reference limit. The load-angle limit controller continuously forms from the generator current and voltage a signal that is proportional to the angle, and compares it with the limit.

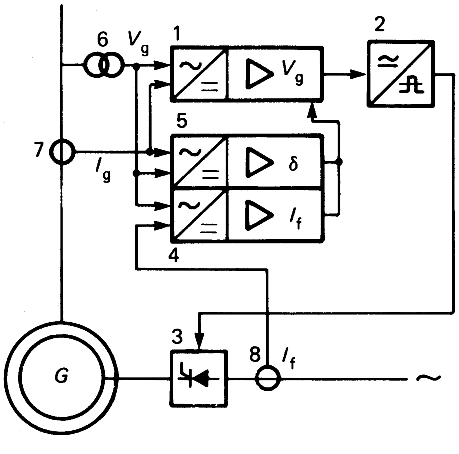


Figure 40.18 Block diagram of the automatic voltage regulation system for a synchronous generator: 1, voltage regulator; 2, gate control unit; 3, static converter; 4, rotor-current limiter; 5, load-angle limiter; 6, voltage transformer; 7, generator-current transformer; 8, excitation-current transformer; G, synchronous generator, V_g, generator voltage; I_g, generator current; I_f, rotor current; δ, rotor angle

The two limiters operate as parallel controllers, i.e. their signals completely replace the voltage as controlled variable when their output variables become smaller or greater than the voltage control signal. Thus, the change in specific dynamic behaviour of the control circuit for each limit function can be fully taken into account. It will be shown that unwanted side-effects occurring on signal take-over can be avoided entirely.

40.5.5.1 Rotor-current limitation

Every exciter is capable of delivering a maximum current appreciably greater than the continuous value based on thermal considerations. This capacity for overexcitation is necessary to provide the additional reactive power demanded during selective clearance of faults and to maintain a synchronous torque even when the voltage level has fallen.

Although it safeguards the winding insulation against damage from prolonged overexcitation, rotor protection gear (such as overcurrent and overtemperature relays) disconnects the generator even when disconnection is not desired. Use of a controller to limit the excitation current to an acceptable value during operation would substantially improve the availability of the generator. For safety reasons, however, a separate protective device must be retained to safeguard the winding against overloading.

The demands to be met by the rotor-current limiter are best explained with reference to the voltage and excitation current when a short-line fault occurs. The voltage regulator reacts to the drop in voltage with surge excitation. The controller is not intended to impede this, but merely to determine the ceiling current I_{fm}. Although the clearance normally takes place in a few hundred milliseconds, the longest duration (back-up protection) of 2–3 s will be assumed. When, after the fault has been cleared, nominal voltage is attained with a permissible continuous excitation current, the thermal controller will not intervene. However, when because of, for example, breaker failure the fault is not cleared, or when generated reactive power no longer suffices to maintain the voltage level, overexcitation will continue to be present. The limit controller now aims to reduce the excitation current before any of the protection gear is tripped. A further important aspect is that this also causes the short-circuit power of the system to be reduced, which, in turn, will diminish the extent of the damage resulting from failure of a breaker or other protective gear.

The condition in which the maximum continuous excitation permissible is present might persist for a long time, and experience has shown that further short circuits appear relatively frequently during this time. When these secondary disturbances occur, it becomes necessary to deliver the maximum reactive current to the system once more, especially to retain system stability. This means that limitation has to be cancelled again for a short time. The characteristic which trips the reset is the steep drop in voltage, $\mathrm{d}V/\mathrm{d}t$.

A further requirement results form the fact that today most of the large excitation equipment is fed from the generator terminals. The necessary ceiling excitation current is normally already available at 90% of the rated voltage. This means that at 120% of the rated voltage, a ceiling current which is 33% higher than the necessary value will flow when an instantaneously acting excitation limiter is not provided.

The limitation signal can be introduced as a reference-value variation of the voltage regulator. However, two facts suggest applying the signal to the voltage regulation output for use as dominant parallel limitation. These are: (i) the limitation is absolute and thus fully effective even when the actual value for the voltage regulator is lost (voltage-transformer fuses blown); and (ii) the limit controller can be adapted to the dynamic conditions of the field-current controller and become independent of voltage-regulator response.

As has been mentioned, limitation to the permissible continuous excitation current should be delayed in order to first give the voltage the maximum support possible. The permissible continuous current in the excitation circuit is determined by the thermal stressing of the field winding, the supply transformer, or the thyristors or diodes. The overload capacity of these elements can usually be represented by integration of the current. A slight overrun of the set limit for the continuous current is then tolerated correspondingly longer. It can be argued, however, that the ceiling excitation will be required only as long as fault clearance is still taking place, in which case a fixed timing element is provided. In both cases, it is usually desired that ceiling excitation should be available again in the event of secondary disturbance.

Based on previous experience, improved rotor-current limit controllers were developed to cater for the many varied applications met with in practice; these controllers can be easily adapted to the particular protection method used.

A maximum-current limit controller, instantaneous acting and always available, ensures that for terminal-fed excitation the desired ceiling current is not exceeded over the entire operating voltage range.

In the case of the simple integrator mode, the speed at which reduction to the limit value takes place is relative to the magnitude of the overcurrent. The integrator is reset only when the limit for the continuous rotor current is underrun.

With the switching mode, reduction to the limit begins after a set delay T_{v} of several seconds. Every time there is a sharp drop in voltage, maximum-current limitation begins anew and the timing element is reset.

A method frequently used in the combined mode: here, integration of the overcurrent is combined with resetting of the integrator by each succeeding voltage drop.

The actual current can usually be measured with a current transformer in the alternating-current (a.c.) power supply prior to rectification by the thyristors or diodes. A d.c./d.c. converter makes connection to a shunt also possible.

When partial failure of components in the excitation circuit causes the permissible continuous current to be reduced, change-over to a pre-set second current limit set-point is possible.

40.5.5.2 Load-angle limitation

For a number of years it has been standard practice to provide protection gear that detects excitation failure or loss of generator synchronism and trips the generator breaker. Inadequate excitation can, for example, also be caused by a change in the system configuration due to a fault, by reference-value failure or by other faults in the voltage regulator. The load-angle limiter obviates unnecessary disconnection by the under-excitation protection device in all these cases.

The limit angle is set intentionally to a value considerably smaller than the pull-out limit. For turbogenerators, for example, the angle lies between 70°⇐and 85°.⇐For salient-pole machines the actual stability must be taken into account and a suitable response limit determined. Thus, an adequate angle difference exists for the acceleration of the generator during clearance of a nearby short circuit. This reserve is directly related to the time allowed for clearing the fault.

The principle of analogue simulation of the phasor diagram using stator voltage and current is applied. This method is notably simpler than the direct measuring processes. In the event of a transmitter on the shaft providing the rotor position, the same device may be employed for further evaluation. The angle between the two simulated phasors, rotor voltage E_{p} and the infinite bus voltage V_{s}, is converted into a d.c. signal in a solid-state angle discriminator. When transients appear, the angle generated by the simulation leads the true load angle slightly. This is advantageous as this limiter should intervene at an early stage. Such a procedure has been shown to be reliable in practice.

40.5.5.3 Stator-current limitation

The stator-current limit, which is governed by thermal considerations, normally lies beyond the operating range of the synchronous generator. Thus, stator-current limitation is employed only in special cases. To illustrate this point, actual applications of a stator-current limit controller, usually in conjunction with limit control of the load angle and excitation current, are given below.

(1) With gas turbo-sets, for example, the useful output can be raised at times to cover peak demands, so that the limit is governed partly by the stator current.
(2) In coastal areas it is necessary to reduce the system voltage at certain times of the year because of salt vapour; there is thus an increase in stator current for the same power.
(3) In the case of reactive power compensators the under-excitation limit can be attained only with a stator-current limiter.
(4) With synchronous motors, a practical utilisation limit can appear that will be detected approximately by a stator-current limit controller.

40.5.6 Control characteristics for synchronous machines

As far as the system of synchronous generator and network is concerned, there is a strong similarity between the speed and active-power control system on the one hand, and the voltage and reactive-power control system on the other.

To understand control by characteristic it is useful to compare the two systems (*Figure 40.19*). As long as the set is operating at no load, or on isolated duty, the appropriate value is controlled by the corresponding control system, i.e. the speed (frequency) or generator voltage. This is no longer the case, however, in parallel operation with a live network. Frequency and voltage are already present and can be changed by the set only to a limited extent. Secondary controlled variables are involved here, and it is imperative that they are controlled in parallel operation. These secondary variables are the active and the reactive power.

On what is this two-fold nature of the control system based? In isolated duty, only one control system is acting on any control loop. In parallel operation, however, many loops are coupled through the common controlled variable. The primary control task is distributed over a large number of control points. Selective stable distribution is of decisive importance. The overall task here is to maintain frequency and to produce active load, or to maintain voltage and to produce reactive load. The familiar solution is to provide the regulator with a drooping characteristic, i.e. the set-point of the primary controlled variable drops as the secondary controlled variable rises. The point of intersection of this characteristic of the network rating defines the corresponding operating conditions. Any set-point can also be set for the secondary controlled variable by parallel displacement of the characteristic. The gradient of the characteristic is usually expressed as a droop of the form $S_{\mathrm{n}} = \Delta n/n$ or $S_{\mathrm{v}} = \Delta V/V$, i.e. as the percentage deviation in the primary controlled variable that is necessary to bring the secondary controlled variable from zero to the rated value.

While the frequency of the network is the same throughout, in the case of voltage this applies only to the imaginary voltage of an infinite bus-bar. The true network represents roughly a variable 'mountain landscape' of voltages defined by its line impedances and feed-in or feed-out at a large number of junction points. This true network must first be accessible by reducing it to a simple equivalent circuit (*Figure 40.20*).

Each individual generator feeds through transformer and line impedances into the 'finite bus-bar' formed by the sum of all the other generators. As can be seen from the phasor diagram in *Figure 40.20*, which may be considered as qualitative, the products of the reactances and reactive components of the current are decisive for the magnitude of the voltage drop $|\Delta V|$. The drop in active voltage and the phase displacement of the voltage phasor can here be ignored. The reactances between the generator terminals and the infinite network thus produce a natural drop in the reactive current, whereas the drop of the speed-control system is always synthetic. The generator reactance X_{D} is within the control loop and therefore need not be taken into account for quasi-steady-state phenomena. The effect of the natural drop in reactive current, and its increase or decrease by artificial means, is the main theme to be discussed from various viewpoints in the following.

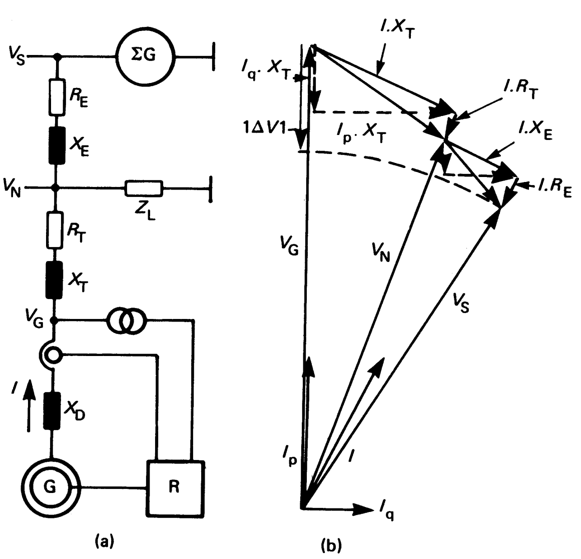


Figure 40.20 Impedances and voltage drops: (a) equivalent circuit diagram; (b) phasor diagram. V_s, Voltage of infinite system; V_N, system voltage; V_G, generator voltage; I, generator current; I_P, active component; I_q, reactive component; X, external (short-circuit) reactance; X_T, transformer reactance; X_D, direct-axis reactance of generator; E_E, R_T, active resistances; Z_L, load impedance; G generator; ΣG, sum of all other generators; R, voltage regulator

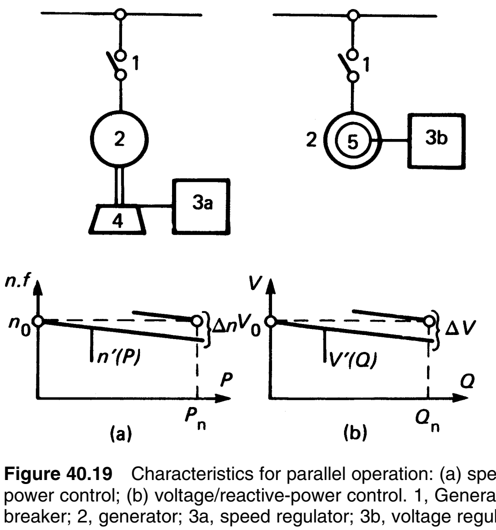


Figure 40.19 Characteristics for parallel operation: (a) speed/active-power control; (b) voltage/reactive-power control. 1, Generator breaker; 2, generator; 3a, speed regulator; 3b, voltage regulator; 4, turbine; 5, rotor winding; 6, network; P, active power; Q, reactive power; n, frequency; V, voltage

40.5.6.1 *Parallel operation in a power-station*

Parallel operation in a power-station is illustrated in *Figure 40.21*. As virtually all large generators are operated in unit connection, the transformer reactance between the terminals and the h.v. bus-bar causes a natural reactive current droop of 8–12%. This accurately defines the reactive power output. Conditions are very different in the case of low-voltage (l.v.) generators, which must operate in parallel direct at the machine terminals. Here there is no natural droop, and even the smallest difference in the voltage values will lead to undesirable mutual exchange of reactive power between the machines, without resulting in a defined distribution of demand. It is not until an artificial droop is introduced by applying reactive current in the measuring loop of the voltage regulator that the reactive load distribution becomes stable; initially it is immaterial whether the generators themselves define the voltage at a purely consumer network, or coincide with the voltage given by a live network.

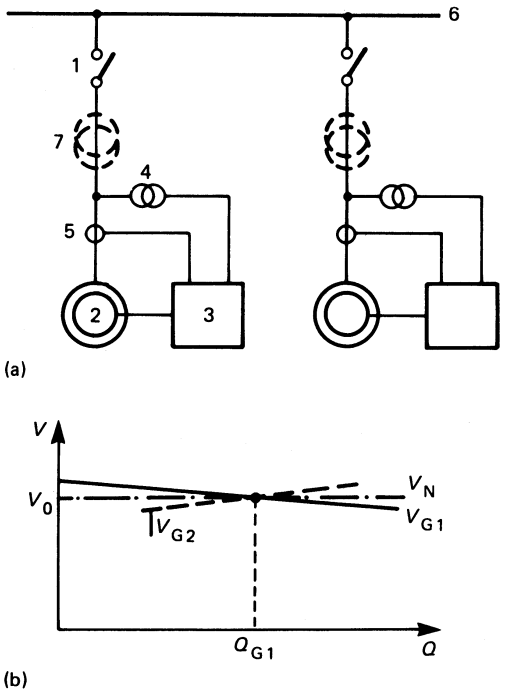


Figure 40.21 Parallel operation in a power-station: (a) basic circuit diagram; (b) characteristics. 1, Generator breaker; 2, generator; 3, voltage regulator; 4, voltage transformer; 5, current transformer; 6, bus-bar; 7, unit-connected transformer; V_N, system voltage; V_{G1}, V_{G2}, generator voltage characteristics; V_0, no-load voltage; Q_{G1}, reactive power output of generator 1

Electronic controllers have no dead-band and permit closed-loop gains of between 100 and 200 in accordance with a proportional range of 1–0.5%. Consequently, in isolated duty a satisfactory distribution over the parallel generators is achieved even with a 3–4% artificial droop. For reasons discussed later, there are various factors that make higher droop values more suitable for interconnected systems. The polygon connection used previously in some power-stations for distributing the reactive load has therefore lost its significance to a large extent because total elimination of changes in the steady-state voltage had to be bought at the cost of complicated circuitry.

40.5.6.2 Principle of current bias

The principle of phasor addition of a current-proportional voltage to the generator voltage has been known in various forms for some time (*Figure 40.22*). All three-phase voltages should be measured in each case to keep ripple and filter time-constants small. As shown in *Figure 40.22*, a one-phase current bias can be applied. In the case of controllers for large generators, the extra cost for symmetrical three-phase bias is justified.

The summation is made such that the overexcited reactive component of the current increases the actual voltage. This method of falsifying the actual value in relation to an unchanged set-point results in a drooping characteristic. In a steady-state condition the effect is the same as that of a reactance. A subtractive current bias causes a rising characteristic, i.e. a compounding. The effect of a true reactance can be reduced by compounding. In simple circuits the reactive current droop is accompanied by a slight compounding in the active current, but in most cases this is useful and under no circumstances is it a disturbing influence.

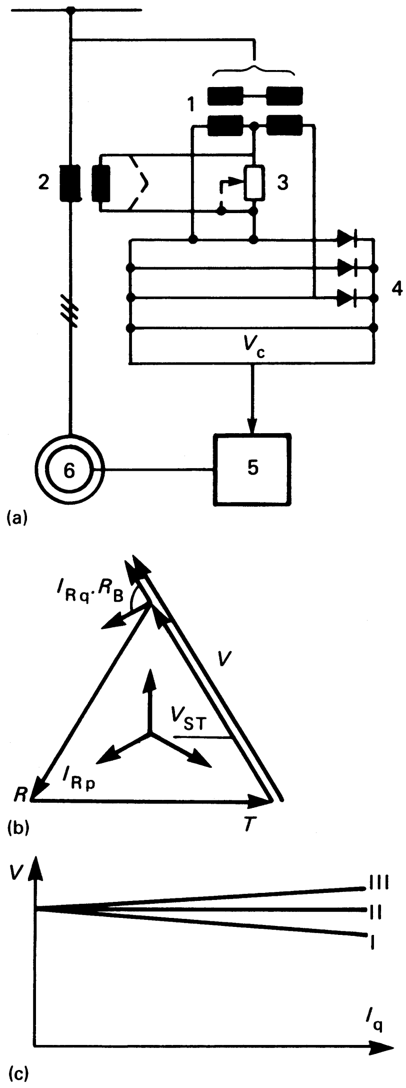


Figure 40.22 Current bias: (a) basic circuit diagram; (b) phasor diagram; (c) characteristics. 1, Three-phase voltage transformer; 2, current transformer; 3, load resistance R_B; 4, rectifier; 5, voltage regulator; 6, generator; V_c, V, r.m.s. voltage; I_{Rp}, active current in phase R; I_{Rq}, reactive current in phase R; I_q, reactive current, superexcited; I, reactive current droop; II, no effect; III, reactive current compounding

It should also be noted that the current bias curve is not entirely linear but slightly progressive. This is of virtually no significance. The current bias can be varied between zero and maximum by altering the load resistance. Nearly all voltage regulators are issued with this equipment.

As mentioned above, the current phasor is chosen in most cases so that an almost purely reactive load effect is created.

However, an active-load effect or a mixture of the two effects can be achieved by selecting a different phase angle. Consequently, in medium-voltage networks, additional active current compounding has shown itself to be a suitable means of improving the operating behaviour.

40.5.7 Slip stabilisation

Power transmission between synchronous machines and the load centre must remain stable even over long distances and under complex conditions. Experience has shown that certain network conditions can cause excessive hunting. Analysis indicates that hunting can be effectively reduced only by the introduction into the voltage control circuit of transient stabilising signals derived from the machine speed/frequency or produced by a change in the electrical output of the generator. Knowledge of the origin of the instability and the means of intervening in the controlled synchronous machine are essential for applying additional signals.

If the synchronous machine is connected to a rigid network through a reactance, it can be easily seen from the phasor diagram that the terminal voltage of the machine and the electrical torque alter with the rotor angle and with the main flux. The changes in terminal voltage, torque and main flux are given by differential quotients. In order to be able to make a quantitative statement for a given duty point, small changes are observed and the transmission functions and their representative factors are linearised about the duty point (index 0). This gives simple expressions, and a block circuit diagram (*Figure 40.23*) can be drawn which characterises the response of the synchronous machine at the rigid network. Let

$$\Delta M_{\mathrm{dE}} = k_1 \Delta\delta\psi + k_2 \Delta E'^{\Leftarrow}_{\mathrm{q}}$$

$$\Delta E'_{\mathrm{q}} = k_3 \Delta E_{\mathrm{p}} + k_4 \Delta\delta\psi$$

$$\Delta V_{\mathrm{t}} = k_5 \Delta\delta\psi + k_6 \Delta E'^{\Leftarrow}_{\mathrm{q}}$$

$$k_1 = \Leftarrow \frac{\Delta M_{\mathrm{dE}}}{\Delta\delta\psi} = \Leftarrow \frac{V_0 \cos\delta_0}{x_{\mathrm{q}} + x_{\mathrm{e}}} E_{\mathrm{q0}} + \Leftarrow \frac{x_{\mathrm{q}} - x'^{\Leftarrow}_{\mathrm{d}}}{x'^{\Leftarrow}_{\mathrm{d}} + x_{\mathrm{e}}} V_0^2 \sin\delta_0 \frac{1}{x_{\mathrm{q}} + x_{\mathrm{e}}}$$

$$k_2 = \Leftarrow \frac{\Delta P_{\mathrm{d}}}{\Delta E'^{\Leftarrow}_{\mathrm{q}}} = \Leftarrow \frac{V_0 \sin\delta_0}{x'^{\Leftarrow}_{\mathrm{d}} + x_{\mathrm{e}}}$$

$$k_3 = \Leftarrow \frac{\Delta E'^{\Leftarrow}_{\mathrm{q}}}{\Delta E_{\mathrm{p}}} = \Leftarrow \frac{x'^{\Leftarrow}_{\mathrm{d}} + x_{\mathrm{e}}}{x_{\mathrm{d}} + x_{\mathrm{e}}}$$

$$k_4 = \Leftarrow \frac{x_{\mathrm{d}} - x'^{\Leftarrow}_{\mathrm{d}}}{x'^{\Leftarrow}_{\mathrm{d}} + x_{\mathrm{e}}} V_0 \sin\delta_0$$

$$k_5 = \Leftarrow \frac{\Delta V_{\mathrm{t}}}{\Delta\delta\psi} = \Leftarrow \frac{V_{\mathrm{d}}}{V_{\mathrm{t}}} V_0 \cos\delta_0 \frac{x_{\mathrm{q}}}{x_{\mathrm{q}} + x_{\mathrm{e}}} - \Leftarrow \frac{V_{\mathrm{q}}}{V_{\mathrm{t}}} V_0 \sin\delta_0 \frac{x'^{\Leftarrow}_{\mathrm{d}}}{x'^{\Leftarrow}_{\mathrm{d}} + x_{\mathrm{e}}}$$

$$k_6 = \Leftarrow \frac{\Delta V_{\mathrm{t}}}{\Delta E'^{\Leftarrow}_{\mathrm{q}}} = \Leftarrow \frac{V_{\mathrm{q}}}{V_{\mathrm{t}}} \frac{x_{\mathrm{e}}}{x'^{\Leftarrow}_{\mathrm{d}} + x_{\mathrm{e}}}$$

The relationship between change in torque and change in rotor angle in a closed voltage-regulator loop is therefore

$$\frac{\Delta M_{\mathrm{dE_2}}}{\Delta\delta\psi} = \Leftarrow \frac{k_1 k_3 k_5 + k_2 k_4 (1 + pT_{\mathrm{E}}) \Leftarrow}{1/k_2 + k_6 k_{\mathrm{E}} + \mathrm{p}(T_{\mathrm{E}}/k_2 + T'^{\Leftarrow}_{\mathrm{d_0}}) + \mathrm{p}^2 T_{\mathrm{E}} T'^{\Leftarrow}_{\mathrm{d_0}}}$$

The factors, $k_1, \ldots, k_6$ have a considerable influence on the damping of the synchronous machine and its behaviour, as follows.

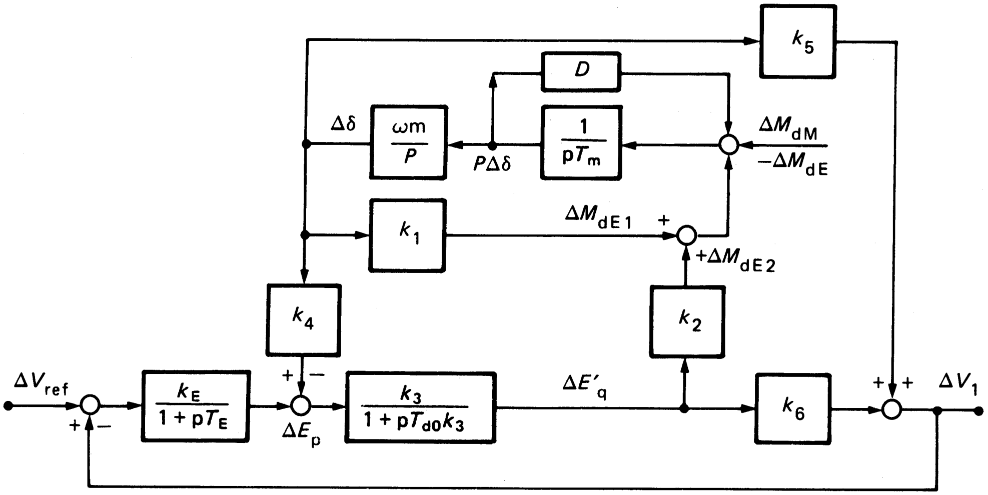


Figure 40.23 Block circuit diagram of a synchronous machine connected to a rigid network: *D*, damping constant of synchronous machine; *k*, k_{E} proportionality factors; M_{dE}, electric torque; M_{dM}; mechanical torque; $T'^{\Leftarrow}_{\mathrm{d0}}$, T_{E}, time-constants; T_{m}, moment of inertia; ω_{m}, angular velocity of synchronous machine; p, operator d/d*t*; other symbols, as for *Figure 40.17*. For linearisation of transfer functions assuming small changes:

$$k_1 = \Leftarrow \frac{\Delta M_{\mathrm{dE1}}}{\Delta\delta\psi} \quad E'_{\mathrm{q}} = \text{constant}$$

$$k_2 = \Leftarrow \frac{\Delta M_{\mathrm{dE2}}}{\Delta E'^{\Leftarrow}_{\mathrm{q}}} \quad \delta\psi = \text{constant}$$

$$k_3 = \Leftarrow \frac{\Delta E'^{\Leftarrow}_{\mathrm{q}}}{\Delta E_{\mathrm{p}}} \quad \delta\psi = \text{constant}$$

$$k_4 = \Leftarrow \frac{\Delta E'^{\Leftarrow}_{\mathrm{q}}}{\Delta\delta\psi} \frac{1}{k_3}$$

$$k_5 = \Leftarrow \frac{\Delta V_{\mathrm{t}}}{\Delta\delta\psi} \quad E'_{\mathrm{q}} = \text{constant}$$

$$k_6 = \Leftarrow \frac{\Delta V_{\mathrm{t}}}{\Delta E'^{\Leftarrow}_{\mathrm{q}}} \quad \delta\psi = \text{constant}$$

k_1 is generally positive and can assume negative values only at correspondingly large network reactances, but it is then no longer possible to transmit stable power. In the normal range, however, this factor has a stabilising effect, i.e. damping.

k_2 is always positive and has a stabilising effect.

k_3 is independent of the rotor angle and responds only to changes in excitation.

k_4 is negative and reduces damping in the system, an effect that can be markedly reduced by high gain in the voltage regulator. The negative electric torque generated by k_4 is synchronous with the torque achieved with k_1 and is compensated by this.

k_5 comprises two components. With large system reactances x_e, i.e. where the rotor angle is large anyway, the expression can become negative. A torque that reduces damping is generated and is further enlarged by the gain at the voltage regulators. Active-load hunting commences with large amplitudes and leads to the machine losing synchronism and being shut down. Only with slip-stabilisation equipment can the effects of this component be eliminated and stable power transmission restored.

k_6 is independent of rotor angle and therefore of no significance as far as hunting is concerned.

It can be seen from the block circuit diagram that the damping-reducing influence is introduced through the difference between set-point and actual values at the voltage regulator. Provided that the synchronous machine is not equipped with static excitation, it is recommended to feed the stabilising signal to the mixing point of the control amplifier. Where static excitation equipment is provided, mixing is also possible direct at the entrance of the gate control system of the power stage of the excitation equipment. To damp the rotor oscillations, it is essential that more active power is delivered at the stator terminals when the rotor is accelerated, and less when it is braked.

There are various means of measuring speed and acceleration, and their corresponding effect on the rotor motion. One simple method is to measure the active power. Assuming that the prime-mover power is constant, changes in the active power cause corresponding acceleration or deceleration of the rotor. Measuring the active power fluctuations gives the first derivative of speed. The actual speed can be determined by integration. The accuracy of the method is adequate for this purpose. However, the reaction to load changes at the drive end is entirely false. Any danger to operation can safely be eliminated by limiting the effect of the stabilising equipment on the voltage control and by introducing an additional signal that is also proportional to frequency.

Thus the use of Δf, i.e. the angular frequency ($\Delta\omega$), instead of integration of ΔP represents a considerable advance in this field. The base signal combination ($\Delta\omega\psi$ plus ΔP) arrangement therefore gives the best arrangement for damping of oscillations (*Figure 40.24*). Test results have fulfilled all expectations in respect of the stabilising signal. Peaks in the signal, which occur when breaker operations take place in the network, are of such short duration that they have no effect upon the excitation. The load/frequency system combines the advantages of load and frequency measurement without any detrimental side-effects.

40.5.8 Adapted regulator for the excitation of large generators

The excitation of a generator is in principle controlled by automatic voltage regulators, and improvements have been achieved in several respects. The original purpose of the feedback arrangement was *voltage regulation*. This basic requirement is important for maintaining stable synchronous operation of the generators in the power system and for controlling the voltage supplied to customers. Further advantages have been obtained since *static* (power semiconductor) *excitation* systems have been used. Through these fast-acting devices, the regulator may contribute to the *transient stability* after faults (keeping the generator from falling out of step) and to the *damping* of electromechanical rotor oscillations. Furthermore, stabilising signals have been introduced with pre-set amplification gains as a compromise to various operational points. These features are now provided by many commercial regulators. The adapted regulator improves the performance one step further. The main result is the good damping provided for a wide region of operating points, and the smooth voltage regulation.

For better understanding of the benefits of an adaptive control on a network, the one-machine system is presented in *Figure 40.25*. The generator is connected to the 'infinite bus' V_∞ over two transmission lines with circuit-breakers. V_∞ is an ideal three-phase voltage source, and the lines are represented by pure reactances. X_e denotes the equivalent reactance between the generator and V_∞. This configura-

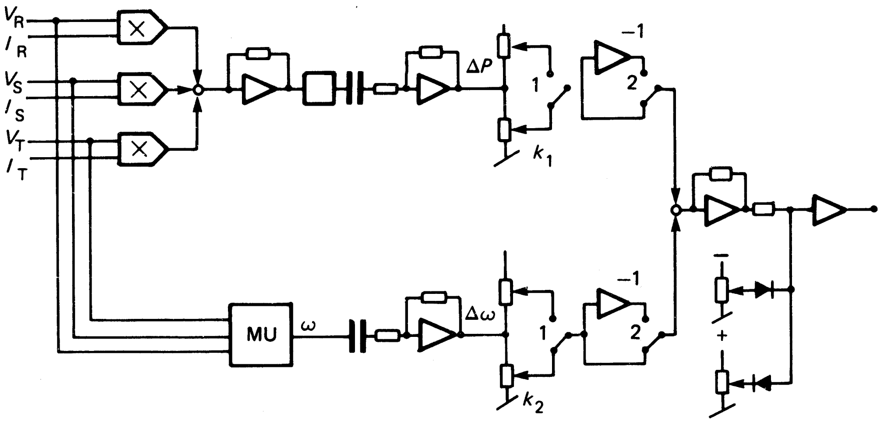


Figure 40.24 $\Delta\omega + \Delta P$ base signal combination arrangement for damping of oscillations, MU

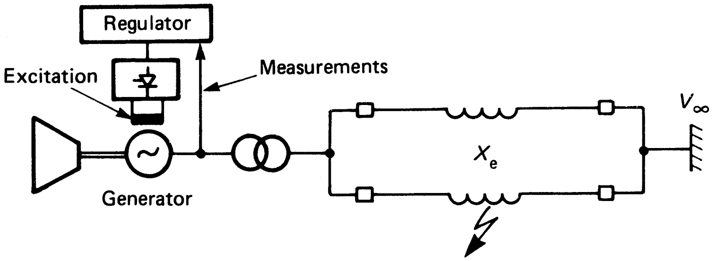


Figure 40.25 Power system model

tion corresponds to the realistic case of a remote power-station supplying a distant load centre.

The control quality of the synchronous machine is assessed in terms of the steady-state behaviour and the dynamic performance in the presence of disturbances. Two kinds of *disturbance* must be considered.

(1) short circuits in the network, line switchings;
(2) changes of operating point (power, voltage, network parameters).

Therefore in the design of the voltage regulator the following performance criteria must be considered.

(1) Regulation of the generator terminal voltage with regard to: (i) smoothness (no ripple during steady-state operation); (ii) speed of response (to avoid overvoltages after load or topology changes); and (iii) accuracy.
(2) Ability to keep synchronism after a fault. This requirement means simply that the excitation voltage should be at the maximum during faults and during dangerous rotor accelerations. The excitation may return to normal after the first peak of the rotor swing.
(3) Damping of rotor oscillations.
(4) Criteria (1)–(3) for a wide region of operating points of the generator (power loading, voltage).

The voltage regulator used here corresponds to the present-day conventional regulator as described earlier. However, the stabilising signals derived from power and frequency measurements are introduced via variable gain factors (*Figure 40.26*). P and ω feedback are regarded as additional signals, the purpose of which is to produce damping (stabilisation) of rotor oscillations.

The gains for the power and frequency feedback are determined by using the D-decomposition technique. D decomposition (or domain separation) is a method based on the characteristic equation of the linearised model. It produces curves of constant damping in the plane of the gains G_1 and G_2, at a certain operating point (P=constant, Q=constant) and constant line impedance X_c. The method is combined with an optimisation procedure to generate optimal gains (optimal damping of the dominant poles). However, as already stated, the gains that are adequate for the best damping of rotor oscillations depend on the operating point of the generator. Owing to the non-linearities in the system, the operating point itself may vary in time, depending on the loading, voltage and network topology (X_c).

Three quantities are sufficient to define the operating point: active power P, reactive power Q and the reactance X_e of the network. P and Q are measurable directly, while X_e can be identified from local measurements by system response. When the three quantities are known, the linearised model of the generator is fully defined (assuming $V_\infty = 1$ p.u.) and the adequate gains can be adjusted automatically.

However, two situations must be considered: steady-state operation and transient operation (short circuit, rotor oscillations). In the steady state, the gains G_1 and G_2 should be reduced for better voltage control, but they may be larger during transient operation.

The resulting adaptation scheme is shown in *Figure 40.27*. The entries of the look-up table are computed off-line. G_1 and G_2 are adjusted whenever major events occur or the parameters have drifted significantly. Updating of the gains is only permitted once every few seconds in order to secure the stability of the adaptation. The identification of X_e is based on a simple curve-fitting procedure. The adaptation scheme (*Figure 40.27*) is implemented on a digital

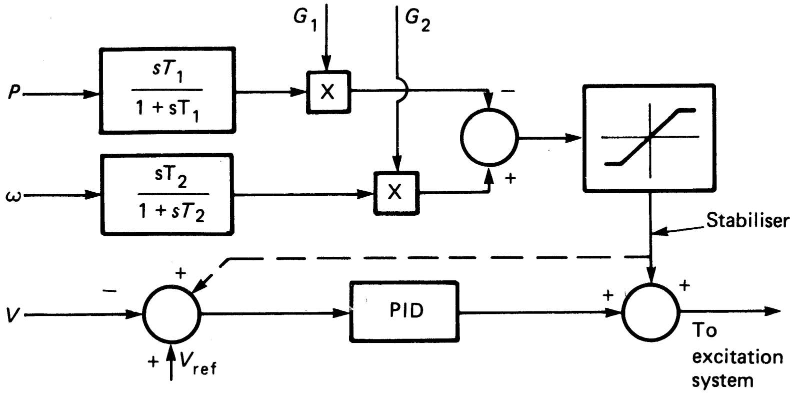


Figure 40.26 The analogure regulator (present-day Brown Boveri Cie (BBC) regulator). - - -, Alternative connection required by some customers

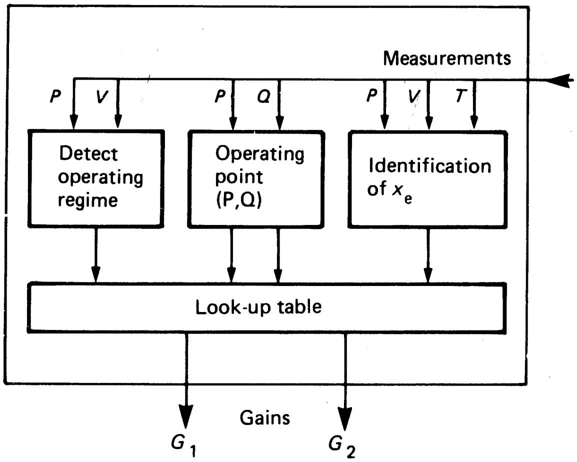


Figure 40.27 The adaptation scheme

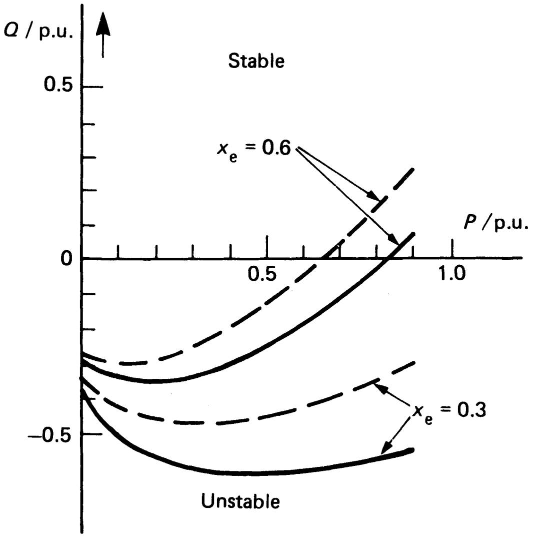


Figure 40.28 Steady-state stability limits of the generator: — —, Unadapted regulator; ——, adapted regulator

microprocessor. The gains G_1 and G_2 are then applied to the analogue regulator of *Figure 40.26*.

The need for adaptation at critical operating points is demonstrated in *Figure 40.28*. Two sets of gains are used, each designed for a specific operating point. The graphs show that the gains designed for one case may not be used in the other. When the wrong gains are used, instability occurs after a disturbance or in the form of self-induced oscillations. *Figure 40.28* shows the region of stable operating points for two typical values of X_e. The stability boundary is evaluated using the linearised model.

40.5.9 Static excitation systems for positive and negative excitation current

Hydroelectric power-plants are usually remote from the load centre, and the power produced is transmitted over long three-phase transmission lines to the consumer. Depending upon the active power to be transmitted (light- or heavy-load operation), the transmission line will supply or consume reactive power, which must be absorbed or supplied by generators at two ends of the line or by reactive current compensators (var sources). Owing to the great distance of transmission along the line, the excitation equipment of such synchronous machines must be designed for high ceiling voltages and negative excitation currents in conjunction with stability of operation. Static excitation equipments with controlled static converters connected in antiparallel and subject to circulating current are capable of satisfying these requirements for hydroelectric generators and synchronous compensators; they improve the capacity of these machines for absorbing reactive load, while retaining optimum control and operating characteristics not merely for steady-state operation but also for dynamic and transient conditions in the network. The circuit diagram for an excitation device is shown in *Figure 40.29*.

The maximum continuously permissible positive excitation current is defined by the maximum continuous load of the synchronous machine for a given power factor. The short-term loadings due to ceiling current and load-independent short-circuit current at the rotor in the event of a fault must also be superimposed upon this value. The maximum continuous current of an excitation device for hydroelectric generators is therefore about 1.5 times the rated excitation current, if redundancy design is ignored. On the other hand, the maximum negative excitation current that must be applied to maintain the voltage of the synchronous machine under extreme capacitive load is determined solely by the design, and therefore by the machine parameters, of the synchronous machine.

A salient-pole generator of terminal voltage V operating at a load angle δ with an internal field electromagnetic force (e.m.f.) E_p supplies a reactive power Q given by

$$Q = \frac{E_p V}{X_d}\cos\delta - \frac{V^2}{X_d}\left(1 + \frac{X_d + X_q}{X_q}\sin\delta\right)$$

where X_d and X_q are the direct- and quadrature-axis reactances, respectively. If the load angle δ and the excitation E_p are zero, the synchronous machine takes continuously from the network a reactive power, the value of which is determined by the direct-axis reactance X_d: i.e. $Q = -V^2/X_d$. The supply of reactive power by the network is then just equal to the absorption capability of the synchronous machine. If no negative excitation current can be supplied and the capacitive network load increases further, self-excitation will occur, and the machine must be disconnected. Since, however, the synchronising torque $M_e = \mathrm{d}P/\mathrm{d}\delta$ begins to decrease only with negative excitation beyond the value corresponding to $E = -V[(X_d - X_q)/X_q]$, negative excitation up to this value can be introduced with quick-acting, regulators, the reactive power absorption rising to $Q = -V^2/X_q$.

The ratio X_d/X_q for salient-pole hydrogenerators lies in the range 1.3–1.4, and the no-load excitation current I_{f0} corresponds to the generator terminal voltage. The maximum negative exciting current is therefore $I_{f0}(X_d/X_q - 1)$, which approximates to $0.4I_{f0}$. Thus the equipment for imposing negative excitation needs to be designed for only about 40% of the excitation on no-load; it does, however, ensure an increase of about 40% in the reactive power absorption at rated load and frequency. If the reactive load supplied by the connected network rises above $Q_c = -V^2/X_q$, the synchronous machine will no longer be capable of holding the voltage, and the definitive self-excitation condition, which cannot

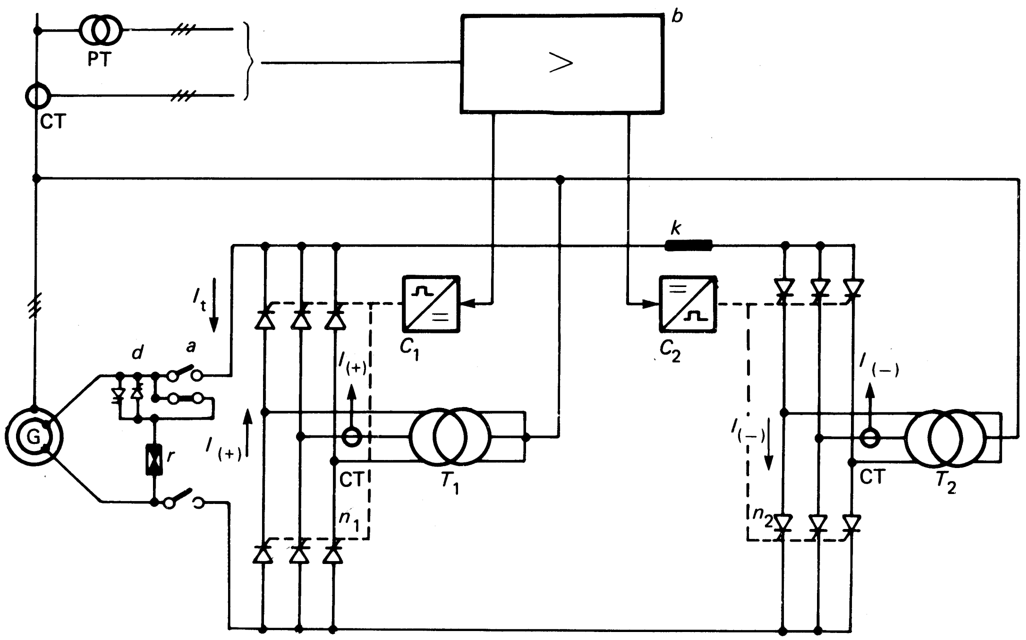


Figure 40.29 Circuit diagram for excitation device: CT, current transformers; G, synchronous generator; I_f, field current of synchronous generator; PT, voltage transformer; T_1, T_2, transformers; a, field suppression switch; b, control electronics; C_1, C_2, firing-angle devices; d, crowbar; k, reactor; n_1, static converter for positive excitation current; n_2, static converter for negative excitation current; r, non-linear de-energising resistor

be controlled by any excitation current or regulator, will be initiated.

If the operating conditions now change specifically with regard to the frequency, the limitations on the use of negative excitation currents must be carefully investigated, since the design of the means for compensating the transmission system will be affected. For instance, with a synchronous machine operating on a long line, if a sudden drop occurs in the active load at the consumer end of the line then, in addition to the charging power for the network, a frequency rise also occurs. This leads both to a rise in the quadrature-axis reactance of the machine and to a fall in the capacitive network impedance. This implies, however, that maintaining the voltage is rendered more difficult by the square of the frequency change. The maximum negative excitation that can be used must in this case be reduced in inverse proportion to the frequency.

In order to decouple the two d.c. circuits (to absorb the different voltage–time areas) and to avoid short-circuit balancing currents, a reactor is incorporated in the d.c. intermediate circuit. The balancing current between the two bridges is regulated to a minimum, but it does provide a guarantee that the two converter sets always carry current; therefore changeover from negative to positive excitation current can be effected virtually without loss of time.

40.5.10 Machine models for investigating stability

Generally, when investigating stability, one examines the consequences of the following disturbances.

(1) Short circuit (one- to three-phase) in the network with subsequent complete (three-phase) or partial interruption of the load flow between generator and network (rapid reclosure).
(2) Partial or complete interruption of load flow between generator and network (load rejection).
(3) Small fluctuations in the load flow (static stability).

It is always assumed that the generators are connected to an infinite power system via a transformer and a network reactance X_{n} (single-machine problem).

The major parameters for studying stability are:

(1) the generator parameters H, X_{d}, $X'_{\mathrm{d}},{\Leftarrow}X''_{\mathrm{d}}$, $X'_{\mathrm{q}},{\Leftarrow}(X'_{\mathrm{q}}{\Leftarrow})$, $X''_{\mathrm{q}},{\Leftarrow}T'_{\mathrm{d}},{\Leftarrow}T''_{\mathrm{d}}$, $(T'_{\mathrm{q}}{\Leftarrow})$, $T''_{\mathrm{q}},{\Leftarrow}X_{\mathrm{p}}$, X_{re}, $E={\Leftarrow}f(I_{\mathrm{F}})$;
(2) the transfer functions of the voltage regulator and turbine controller;
(3) the unit-transformer reactance X_{tr} and the network;
(4) the nature of the fault (one-phase, three-phase, short circuit, etc.), the fault location and fault time t_{F}.

The generator parameters are determined by the construction of the machine. Evaluation of the data for a variety of machines reveals the following trend: for a given frequency f, the inertia constant H diminishes with increase in rating S and a decreasing number p of pole pairs:

$$H ={\Leftarrow}\frac{GD^2}{4}\frac{\omega_{\mathrm{s}}^2}{2S}$$

where $GD^2/4$ is the moment of inertia referred to diameter, and $\omega_{\mathrm{s}} = 2\pi f/p = 2\pi n_{\mathrm{s}}$ is the synchronous angular velocity. Only a lower limit of H can be specified, because its magnitude depends considerably on the turbine inertia.

The reactances X_{d} and $X'_{\mathrm{d}},{\Leftarrow}$the stator resistance R and the short-circuit time-constant $T'_{\mathrm{d}}{\Leftarrow}$show a tendency to become larger with increasing power rating. However, these relationships can be very strongly influenced by structural features and design layout.

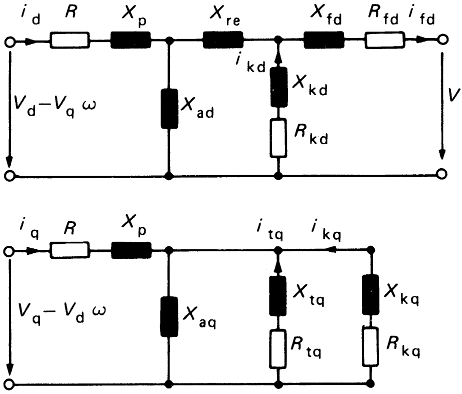


Figure 40.30 Equivalent circuit for model 1 of a synchronous machine (for symbols, see text)

Experience shows that the method of calculation and the machine model influence the results more strongly with high-speed transients than with slow ones.

40.5.10.1 Synchronous machine models

Six different models of synchronous machine are generally used. Saturation is disregarded and the same equations of motion are used for all models.

Model 1 This model has six windings (three damper windings). Its equivalent circuit diagram is shown in *Figure 40.30*. The corresponding equations are as follows.

The network equations are

$$\begin{aligned} V_d^G &= V_d^N + R_e i_d + L_e(di_d/dt) - \omega L_e i_q \\ V_q^G &= V_q^N + R_e i_q + L_e(di_d/dt) + \omega L_e i_d \end{aligned} \qquad (40.1)\Leftarrow$$

in which $\omega L_e = X_{tr} + X_N$.

The following equations are used for the synchronous machine:

Flux–current relationships

$$\begin{aligned} {}_d &= (X_p + X_{ad})i_d + X_{ad}i_{kd} - X_{ad}i_{fd} \\ {}_{kd} &= X_{ad}i_d + (X_{ad} + X_{re} + X_{kd})i_{kd} - (X_{ad} + X_{re})i_{fd} \\ {}_q &= (X_p + X_{aq})i_q + X_{aq}i_{tq} + X_{aq}i_{kq} \\ {}_q &= (X_p + X_{aq})i_q + X_{aq}i_{tq} + X_{aq}i_{kq} \\ {}_{tq} &= X_{aq}i_q + (X_{aq} + X_{tq})i_{tq} + X_{aq}i_{kq} \\ {}_{kq} &= X_{aq}i_q + X_{aq}i_{tq} + (X_{aq} + X_{kq})i_{kq} \end{aligned} \qquad (40.2)\Leftarrow$$

Voltage equations

$$\begin{aligned} d\ {}_d/dt &= -V_d^G + {}_q\omega + R i_d \\ d\ {}_q/dt &= -V_q^G - {}_d\omega + R i_q \\ d\ {}_{fd}/dt &= V_{fd} - R_{fd}i_{fd} \\ d\ {}_{tq}/dt &= -R_{tq}i_{tq} \\ d\ {}_{kd}/dt &= -R_{kd}i_{kd} \\ d\ {}_{kq}/dt &= -R_{kq}i_{kq} \end{aligned} \qquad (40.3)\Leftarrow$$

Mechanical equations
Torque:

$$M_B = {}_q i_d - {}_d i_q \qquad (40.4)\Leftarrow$$

Power:

$$P = V_d^G i_d + V_q^G i_q \qquad (40.5)$$

$$Q = V_q^G i_d - V_d^G i_q \qquad (40.6)$$

Equation of motion:

$$2H(ds/dt) = M - M_B \qquad (40.7)\Leftarrow$$

$$d\theta/dt = s \qquad (40.8)\Leftarrow$$

$$\omega = 1 + s \qquad (40.9)\Leftarrow$$

The quantities L, X, R, , i, V, M and ω are per-unit quantities.

From the input data

$$R, X_p, X_c, X_d, X_q, X'_d, X''_d, X''_q, T'_d, T'_q, T''_d, T''_q, S_N, V_N, f, I_{fd0}, V_{fd0}, p, H$$

of which T, S_N, V_N, f, I_{fd0}, V_{fd0} and H are not per-unit, the constants X_{ad} and X_{aq} of the synchronous machine can be calculated in the following manner:

$$X_{ad} = X_d - X_p \qquad (40.10)\Leftarrow$$

$$X_{aq} = X_q - X_p \qquad (40.11)\Leftarrow$$

$$X_{re} = \frac{X_{ad}(X_{re} - X_p)}{X_d - X_{re}} \qquad (40.12)\Leftarrow$$

$$X_{td} = \left(\frac{X_d - X_p}{X_d - X_{re}}\right)^2 \frac{(X_d - X_e)(X'_d - X_e)}{X_d - X'_d} \qquad (40.13)\Leftarrow$$

$$X_{kd} = \left(\frac{X_d - X_p}{X_d - X_{re}}\right)^2 \frac{(X'_d - X_e)(X''_d - X_e)}{X'_d - X''_d} \qquad (40.14)$$

$$X_{tq} = \frac{(X_q - X_p)(X'_q - X_p)}{X_q - X'_q} \qquad (40.15)\Leftarrow$$

$$X_{kq} = \frac{(X'_q - X_p)(X''_q - X_p)}{X'_q - X''_q} \qquad (40.16)$$

$$R_{fd} = \left(\frac{X_d - X_p}{X_d - X_{re}}\right)^2 \frac{(X_d - X_e)^2}{X_d - X'_d}\,\frac{1}{T_{fd}\omega_B} \qquad (40.17)\Leftarrow$$

$$R_{kd} = \left(\frac{X_d - X_p}{X_d - X_{re}}\right)^2 \frac{(X'_d - X_e)^2}{X'_d - X''_d}\,\frac{X''_d}{T''_d X'_q \omega_B} \qquad (40.18)$$

$$R_{kq} = \frac{(X'_q - X_p)^2}{X'_q - X''_q}\,\frac{X''_q}{T''_q X'_q \omega_B} \qquad (40.19)$$

$$R_{tq} = \frac{(X_q - X_p)^2}{X_q - X'_q}\,\frac{1}{T_{tq}\omega_B} \qquad (40.20)\Leftarrow$$

$$T_{td} = T'_d \frac{X_d}{X'_d}\left(T_{kd} - T''_d \frac{X'_d}{X''_d}\right) \qquad (40.21)\Leftarrow$$

where

$$T_{kd} = T''_d \frac{X'_d}{X''_d} \times \frac{(X_d - X_{re})(X'_d - X''_d) + (X'_d - X_{re})(X''_d - X_{re})}{(X'_d - X_{re})^2} \quad (40.22)$$

$$T_{tq} = T'_q \frac{X_q}{X'_q} - \left(T_{kq} - T''_q \frac{X'_q}{X''_q}\right) \quad (40.23)$$

$$T_{kq} = T''_q \frac{X'_q}{X''_q} \times \frac{(X_q - X_p)(X'_q - X''_q) + (X'_q - X_p)(X''_q - X_p)}{(X'_q - X_p)^2} \quad (40.24)$$

$$\omega_B = 2\pi f \quad (40.25)$$

The reference values are:

V_N for voltages V_d^G and V_q^G

S_N for powers P and Q

$I_N = S_N/\sqrt{3}V_N$ for current i_d and i_q (40.26)

$M_N = S_N/\omega_B$ for torque M (40.27)

I_{fd0} for excitation current i_{fd}

V_{fd0} for the excitation voltage

In Equations (40.1), R_e and L_e are respectively the resistance or the inductance of the positive-sequence system between generator and infinite network, i.e. $\omega L_e = X_{tr} + X_N$.

If the site of the short circuit is between the generator and the infinite network, there are ten first-order differential equations to integrate.

Model 2 This model has only one damper winding in the quadrature axis, i.e. $i_{tq} = 0$. The relationships $_{tq} = f(i_q, i_{tq}, i_{kq})$ in Equations (40.2) and d $_{tq}/dt = R_{tq} i_{tq}$ in Equations (40.3) do not apply. Instead of Equations (40.16) and (40.19) we have

$$X_{kq} = \frac{(X_q - X_p)(X''_q - X_q)}{X_q - X''_q} \quad (40.16a)$$

$$R_{kq} = \frac{(X_q - X_p)^2 X''_q}{(X_q - X''_q) T''_q X_q \omega_B} \quad (40.19a)$$

Equations (40.20), (40.23) and (40.24) do not apply.

Model 2 also takes account of the transformation terms, but has only two damper windings. In this case only nine differential equations have to be integrated.

Model 3 This, like model 2, has one damper winding in each of the direct and quadrature axes. The transformation terms in the stator equations do not apply, i.e. Equations (40.3) become

$$V_d^G = \psi_q \omega - R i_d$$
$$V_q^G = -\psi_d \omega - R i_q$$
$$\mathrm{d}\ _{fd}/\mathrm{d}t = V_{fd} - R_{fd} i_{fd} \quad (40.3a)$$
$$\mathrm{d}\ _{kd}/\mathrm{d}t = -R_{kd} i_{kd}$$
$$\mathrm{d}\ _{kq}/\mathrm{d}t = -R_{kq} i_{kq}$$

Model 3 also has two damper windings. The transformation terms are disregarded. There are only five differential equations to integrate. It is found from experience that with this configuration the time step can be much larger than with models 1 and 2.

The equations of motion (40.7)–(40.9) remain unchanged. In addition to M_B in Equation (40.4), however, the braking torque M_{Bs}, caused by the transformation terms when switching operations occur, must be introduced into Equation (40.7):

$$2H(\mathrm{d}s/\mathrm{d}t) = M - M_B - M_{Bs}$$

if the movement process is to be represented correctly.

Model 4 This model has no damper windings. The basic equations therefore become

$$_d = (X_p + X_{ad}) i_d - X_{ad} i_{fd}$$
$$_{fd} = -X_{ad} i_d + (X_{ad} + X_{fd}) i_{fd}$$
$$_q = (X_p + X_{aq}) i_q \quad (40.2b)$$
$$V_d^G = \psi_q \omega - R i_d$$
$$V_q^G = -\psi_d \omega - R i_q$$
$$\mathrm{d}\ _{fd}/\mathrm{d}t = V_{fd} - R_{fd} i_{fd} \quad (40.3b)$$

In the absence of damper windings, the asynchronous damping torque (M_{B2}) also must be included in the Equation of motion (40.7):

$$2H(\mathrm{d}s/\mathrm{d}t) = M - M_B - M_{Bs} - M_{B2}$$

Instead of Equations (40.10)–(40.24) we have the expressions

$$X_{ad} = X_d - X_p \quad (40.10^*)$$
$$X_{aq} = X_q - X_p \quad (40.11^*)$$
$$X_{fd} = \frac{(X_d - X_p)(X'_d - X_p)}{X_d - X'_d} \quad (40.13a)$$
$$R_{fd} = \frac{X_d - X_p}{X_d - X'_d} \frac{X'_d}{T'_d X_d \omega_B} \quad (40.17a)$$

The transformation terms are disregarded. Thus only three differential equations need to be integrated.

Model 5 This model has no field winding and no damper winding. The base equations for this model are:

$$_d = X'_d i_d - E_f \quad (40.2b)$$
$$_q = X_q i_q$$
$$V_d^G = \psi_q \omega - R i_d$$
$$V_q^G = -\psi_d \omega - R i_q \quad (40.3c)$$

The mechanical behaviour is described by Equations (40.4) and (40.7)–(40.9). E_t is calculated in terms of the initial values of i_d, i_q and V_q^G:

$$E_t = V_q^G + R i_q + X'_d i_d$$

On the stator, conditions are represented by one winding each in the direct and quadrature axes, their reactances being different. The effect of excitation appears as a current source in the equivalent circuit of the direct-axis winding. Only the two differential equations for the mechanical system have to be integrated.

Model 6 This model, like model 5, has neither field winding nor damper windings. Equations (40.2b) contain $X_{\mathrm{d}}'^{\Leftarrow}$ instead of X_{q}. Thus Equation (40.4) becomes

$$M_{\mathrm{B}} = \Leftarrow E_{\mathrm{t}} i_{\mathrm{q}}$$

Model 6 resembles model 5 except that it has identical reactance in the direct and quadrature axes.

40.5.10.2 Comparison of the models

The essential difference between models 1 and 2 on the one hand, and models 3–6 on the other, is the allowance made for the transformation effect. Models 1–3 have damper windings, while models 4–6 do not. The stability of a synchronous machine after disconnection of the short circuit is determined chiefly by the accelerating torque M_{A}, or braking torque M_{B}, that was acting on the rotor during the short circuit. Accelerating torque M_{A} is composed of driving torque M and braking torque M_{B}: i.e. $M_{\mathrm{A}} = \Leftarrow M - M_{\mathrm{B}}$. The braking torque M_{B} arising during the short circuit and oscillations must therefore be correctly represented in each model.

The braking torque M_{B} acting after the fault has occurred consists of the following component parts:

(1) a braking torque M_{B1} corresponding to losses in the stator and network resistance R;
(2) an asynchronous braking torque M_{B2} corresponding to losses in the rotor field and damper windings; and
(3) a decaying braking torque M_{B3} introduced by the transformation terms in the machine and network equations—this torque becomes effective at every change of state.

M_{B3} is mainly responsible for the well-known 'backswing' effect. The synchronous machine is then braked initially in the first 20 ms after the short circuit. However, the extent of the backswing depends on the site of the short circuit, because torque M_{B2} incorporates the network reactances between the machine and short-circuit location. In some instances, therefore, a short circuit across the terminals may be less of a stability problem than a distant short circuit.

Direct account is taken of all the braking torques M_{B1} to M_{B3} only in the case of models 1 and 2. M_{B1} is present with all the models. With model 3 there is also torque M_{B2}.

The following conclusions can be drawn.

(1) The transformation terms in the Park equations (models 1 and 2) and in the network equations must be taken into account (i) when power is low, (ii) when the short circuit is in the vicinity of the generator, and (iii) if the braking torque M_{B3} brought about by the transformation terms cannot be introduced into the mechanical equations.
(2) The backswing effect occurs only when the site of the short circuit is close to the generator. From this it follows that under certain circumstances a short circuit remote from the generator can be more dangerous, as regards stability, than a short circuit nearby.
(3) It is known that the individual synchronous machine models behave very differently.

In many instances, especially in the case of transient stability problems, the simpler models (4 and 5) exhibit the same behaviour as model 2 or 1 with transformation effect. Nevertheless, with the simple models it is particularly important to represent damping correctly. Models 4 and 5 are therefore particularly suitable for simulating synchronous machines with constant excitation voltage that are located far away from the short circuit. Investigation shows also that the widely recommended model 6 (constant voltage beyond transient reactance) yields results that are pessimistic.

If it is assumed that the constant H decreases, and reactances X_{d}, $X_{\mathrm{d}}'^{\Leftarrow}$and X_{t} increase, with rising generator capacity, then the critical short-circuit duration will be smaller as generator capacity rises. Studies show, however, that the behaviour of even large synchronous machines can be described with the known models of synchronous machines, and that there is usually no need to make allowance for the transformation terms.

40.6 Decentralised control: electronic turbine controllers

40.6.1 Introduction

The turbine equipped with a controller represents an important item of decentralised control as its controller governs the active power input to the power system. For many reasons, both historical and technical, this controller is kept separate from excitation control, although a computer would be able to fulfil the combined function.

Modern turbine controllers are electronic and have to meet a series of functions and requirements. Frequency control is just one of these; safety functions, monitoring, limit checking, etc., are equally important. The most important requirements are the following:

(1) automatic start-up and shut-down of the turbine;
(2) application of a load–frequency control system;
(3) application of an external reference input;
(4) adjustable droop of the speed controller;
(5) good adaptability to the given operating condition;
(6) high-response sensitivity;
(7) Co-ordinated operation of the turbine–generator set and all other systems in the power-plant;
(8) short reaction time between controller and control valves; and
(9) short closing time of the control valves.

Nowadays turbine controllers are composed of electronic modules or functional groups. Most of these can be used in the control system of steam, gas or water turbines. However, the control schemes of such turbines vary widely. We therefore describe both a modern steam-turbine control system and a system for a Francis water turbine. First, however, we explain in simple terms the basic functions of frequency and load control. We derive the elementary block diagram of a frequency controller, explain set-point control, and set out the basic concept of droop. Our treatment concludes with a consideration of frequency control in a multi-machine system.

40.6.2 The environment

The turbine driven by water or steam is connected to a synchronous generator. The synchronous generator has the important property of being able to align itself with other generators when the armatures are interconnected and the rotors are excited. When this alignment is achieved, i.e. when peaks and zero crossings of the induced voltages occur at the same instant, 'synchronous' operation is realised. In greater detail, there are small differences between the mechanical positions of the rotors due to the individual loadings, which may reach 45°. We do not consider different numbers of poles, but assume that all generators have the same number of poles.

In synchronous operation all generators have the same frequency or speed. When the frequency changes, all generators change their speed. Hence, frequency or speed control in synchronous or interconnected operation means control of one common frequency. Small deviations and transients with respect to this frequency are corrected by the inherent ability of the synchronous machines to maintain alignment.

A single generator can also be operated in a system. However, in this case there is no other generator to which it can be aligned. The operation is similar to that of a d.c. generator. Speed control and angular position are purely a matter of the balance between prime-mover torque and load torque acting on the inertia of the shaft. Following a load change, there will be a change in frequency. The speed governor, which is the frequency controller, has an important bearing on the frequency behaviour.

In contrast, a change in loading of a single generator in interconnected operation does not necesarily affect the frequency because it is maintained by the other generators. Hence it is important to know the environment in which a speed governor has to function. Modern governing systems are being designed to cope with a wide range of system conditions.

40.6.3 Role of the speed governor

The speed governor can best be understood by considering a steam engine with a generator serving a local load in isolation, the turbine being controlled by a fly-ball governor. The basic function is realised by a proportional control, whereby the valve position is the actuating signal. Speed or frequency is the input from the system, i.e. the controlled quantity which is compared with a reference signal (also called the 'set-point'). The difference between the measured signal and the set-point is converted to the valve position, which involves a gain. This gain has a two-fold meaning. First, it is a pure signal gain relating a deviation, based on nominal quantities, and a deviation of the valve position which is also referred to as a nominal position. Second, the gain means also a power amplification, i.e. the conversion of a weak electrical signal to a strong mechanical torque. A schematic diagram of the basic arrangement is given in *Figure 40.31*, where the functional relations are shown.

Assuming that the stability of such a system is guaranteed, transient and steady-state conditions can easily be calculated. The basic functions can best be understood by considering steady-state or quasi-steady-state operating points. If the need arises to deliver more power to the generator, the valve of the turbine has to be opened. This is realised by forming a difference between the set-point and the measured speed signal either by raising the set-point or by a drop in frequency. Changes in the governing system will take place until a balance is reached. Steady-state conditions are found from the relations

$$\Delta fK = M_e; \quad \Delta f = f_s - f$$

where f is the frequency, Δf is the frequency deviation, f_s is the set-point, M_e is the electrical or load torque, and K is the gain of the speed governor in units of torque per unit of frequency.

The gain determines the amount of frequency deviation. The higher K is, the smaller Δf will be. However, the gain cannot be chosen to be arbitrarily high. There are problems of stability and certain restrictions due to the allocation of changes of power. The relation $\Delta fK = M_e$ also shows that the speed governor has a double function. It carries out both load control, i.e. the matching of prime-mover torque to the load torque, and frequency control. The relation between these two functions is given by the gain K. Its dimensions in practice are megawatts per hertz; i.e. instead of torque, it is the power which is measured at the output.

Hence, in isolated operation, frequency changes follow a load change so long as the set-point is fixed. A change in set-point will cause a rise in frequency when the load remains constant.

In interconnected operation, the frequency remains practically unchanged. Therefore, a change in loading cannot be affected by the power system. The output of the generator changes without any variation of the frequency, only when the set-point is changed. In this case the speed governor is a pure load controller.

In practice the speed governor is always ready to take on either function. It is the boundary conditions that determine its momentary role, i.e. the governor controls the output power when the frequency is imposed and it controls the

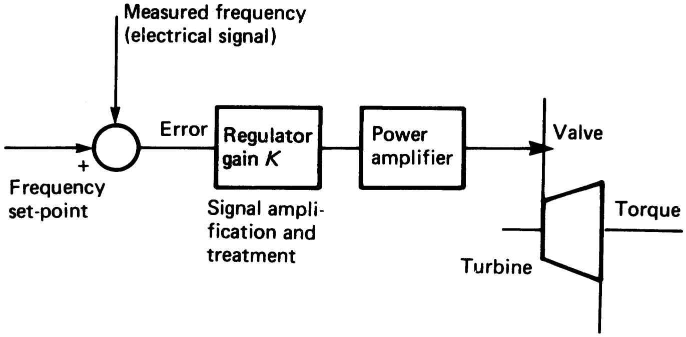


Figure 40.31 Basic arrangement of a speed governor (frequency controller: the basic functions of error detection, signal amplification and power amplification are shown)

frequency when the load is imposed. In a more complex situation, when load and frequency are dependent, the governor varies its output until an equilibrium is reached.

40.6.4 Static characteristic

Modern speed governors have amplifiers that include an integrating property. This choice is made for both reasons of principle and technical reasons, the main one of which is the control of servomotors by valves or signal converters. The proportional behaviour is still maintained by adding the output power as a measured quantity. Such a governor acting on a turbine–generator is commonly represented by a block diagram as shown in *Figure 40.32*.

In this schematic, two signals measured at the output are compared with set-points, i.e. frequency and power. All deviations are combined in one summing junction, the frequency deviation being weighted by K. The summing junction produces an error which drives the regulator as an integrator. The regulator, which is at the same time a power amplifier, drives the valve.

The steady-state behaviour of the system is described by a characteristic graph with a few parameters. The starting point is the zero value of the error e when the system (*Figure 40.32*) has reached a stationary operating point. Then

$$e = P_s - P + K(f_s - f) = 0 \quad \Leftarrow\Leftarrow$$

where P is the measured power output and P_s is the power set-point; the other parameters have already been defined. P and f are variables, represented by the axes in *Figure 40.33*. The frequency f can be expressed by $f = f_s + (P_s - P)/K$, a linear relationship between f and P. The slope is

$$df/dP = -1/K = -D$$

where D is the *droop* of the system (in hertz per megawatt). Together with the intersection $(f_s + P_s/K)$, it determines the position of the straight-line characteristic of the speed-governing system. Discussion of the role of the parameters gives an insight into the operation of the system.

40.6.4.1 No-load operation and f_s

On no load the output is zero and the speed is given by the intersection of the straight line with the ordinate. Thus f_s is used to set the no-load speed. P_s is set to zero.

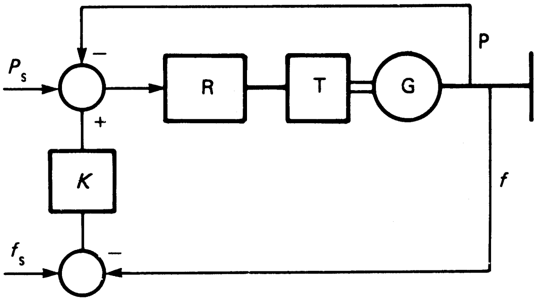


Figure 40.32 Schematic diagram of a power–frequency controlled generator: G, generator; T, turbine; R, regulator (integrating); P, measured power; P_s, set-point power; f, measured frequency; f_s, frequency set-point; K, gain

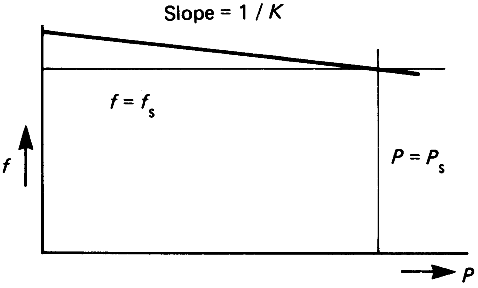


Figure 40.33 Static characteristic: the operating point is confined to the sloping straight line

40.6.4.2 Synchronous operation and P_s

When the no-load frequency has reached the system frequency and the synchronising conditions are met, the generator is first connected to the power system, then loaded by raising P_s to the desired value. The characteristic is thereby raised in parallel. The frequency remains unchanged. Should the generator breaker be opened inadvertently, the frequency will assume the value given by the intersection of the straight line with the ordinate. This involves a rise in frequency that is governed by K.

Consider a generator which has the following settings and parameters:

$$f_s = 50\,\text{Hz};\ P_s = 200\,\text{MW};\ \ K = 80\,\text{MW/Hz}$$

The generator is operating at $f = 50$ Hz, $P = 200$ MW. After disconnection of the generator, the no-load frequency will become $f = 52.5$ Hz. The droop is $D = 1/K = 0.0125$ Hz/MW. It can, however, also be expressed as a percentage:

$$D = 100(\Delta f/f_0)(P_0/P) \Leftarrow$$

where f_0 and P_0 are nominal values. In this example, the droop is 5%. Thus the droop is a determining factor for the no-load frequency after load shedding.

40.6.4.3 Isolated operation

Although the set-point f_s is maintained at 50 Hz, the frequency f will deviate from f_s because of load changes. The intersection of the characteristic with the load characteristic (e.g. vertical line) fixes the frequency.

Consider the situation given by *Figure 40.34* (not drawn to scale). The set-points are $f_s = 50$ Hz, $P_s = 160$ MW and $K = 100$ MW/ Hz. The load is 180 MW. Hence $f = 49.75$ Hz. In order to adjust the frequency to its nominal value, P_s has to be raised by 20 MW. Load shedding would result in a frequency rise according to $f_s + DP_s$.

Frequency control in isolated operation means either letting the operating point vary along the characteristic or readjusting the set-point whenever a frequency deviation arises.

40.6.5 Parallel operation of generators

Consider a two-machine system to serve a common variable load. Each set has a speed governor. The system schematic is given in *Figure 40.35*; it is assumed that the generators supply 200 MW shared in the ratio 80/120. The frequency is at the nominal value of 50 Hz.

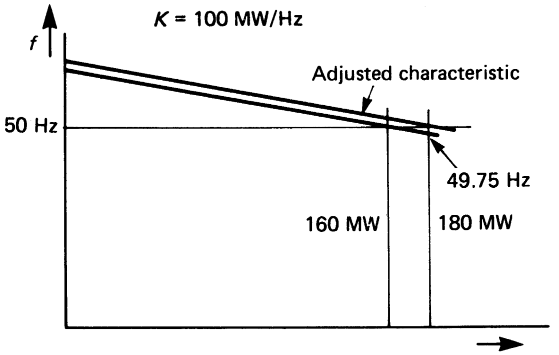


Figure 40.34 Isolated operation of one generator: load increase by 20 MW, and readjustment of characteristic in order to raise frequency to 50 Hz

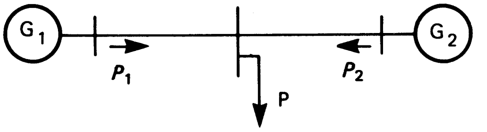


Figure 40.35 Two-machine system: G_1, G_2, generators operating at the same frequency serving P. $P_1 = 80$ MW, $P_2 = 120$ MW

How will an increase of load by 40 MW be shared? The problem is readily solved by calculation or graph. The required data for the two generators, G_1 and G_2 are:

$$G_1: f_{s1} = 50.0\,\text{Hz}; P_1 = 80\,\text{MW}; P_{s1} = 80\,\text{MW};$$
$$K_1 = 16\,\text{MW/Hz}$$
$$G_2; f_{s2} = 50.0\,\text{Hz}; P_2 = 120\,\text{MW}; P_{s2} = 120\,\text{MW};$$
$$K_2 = 32\,\text{MW/Hz}$$

The new frequency f and the power increments ΔP_1 and ΔP_2 are to be determined, with $\Delta P_1 + \Delta P_2 = 40$ MW. Then

$$P_{s1} - P_1 - \Delta P_1 + K_1(f_{s1} - f) = 0$$

$$P_{s2} - P_2 - \Delta P_2 + K_2(f_{s2} - f) = 0$$

But $f_{s1} = f_{s2} = f_s$, and f is the common frequency; hence these equations in combination give

$$(P_{s1} + P_{s2}) - (P_1 + P_2) - (\Delta P_1 + \Delta P_2) - (K_1 + K_2)(f_s - f) = 0$$

a resultant characteristic with a set-point of $P_{s1} + P_{s2}$ and a gain $K_1 + K_2$, or a droop $1/(K_1 + K_2)$ (*Figure 40.36*). The solution is $\Delta P_1 = 13.3$ MW, $\Delta P_2 = 26.7$ MW.

This principle of combining generators with equivalent generators, with a composite droop, is quite general and can be extended to multi-machine systems. Any generator can be taken as generator G_1 and the rest as generator G_2. The combined system has a composite characteristic called a power–frequency characteristic. Thus in an interconnected system, control of power and frequency is realised by adjusting the set-points of the various generators either manually or by automatic means.

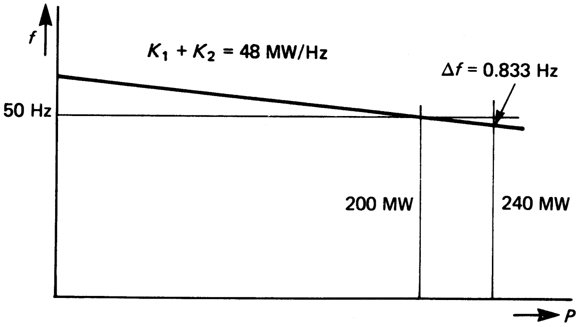


Figure 40.36 Two-machine system, combined characteristic

The logical extension of frequency control at this level (primary control) to a regional level is load–frequency control or automatic generation control.

40.6.6 Steam-turbine control system (Turbotrol)

The Turbotrol electrohydraulic control system has to fulfil the following objectives:

(1) Automatic variation of the speed set-value for speed-controlled run-up of a turbo-set with respect to critical speed and temperature conditions.
(2) Speed control during no-load operation.
(3) Frequency control when feeding house load.
(4) Accurate rapid load control in accordance with an adjustable linear frequency–load characteristic.
(5) Automatic load set-value variation for loading or load shedding, taking into consideration all measures necessary to prevent the turbo-set from being overstressed.
(6) Maintenance of the speed within admissible limits on load shedding.
(7) Co-ordinated action of the steam generator controller and turbine controller during load operation.
(8) Processing of the signals of an overriding control system, e.g. from the load dispatching system controller.

The Turbotrol system also must ensure (i) high operational safety and availability, (ii) high-response sensitivity when changing specific control variables, and (iii) ease of servicing and minimum maintenance.

40.6.6.1 Structure of control system and types of operation

The arrangement of the individual function groups of a turbine controller is shown by the block diagram in *Figure 40.37*. For clarity the numbers assigned to these groups have been inserted in the text. The control system must be designed for the following types of operation.

(1) *Normal conditions*—start-up, no load, synchronising, load operation.
(2) *Special condition*—control of the live-steam pressure by the turbo-set.
(3) *Fault conditions*—manual control with the turbine master station, turbine trips.
(4) *Testing*—overspeed, simulation.

Start-up During this phase the run-up controller (424) is in action, the turbine master station is set to ‘auto’ and the

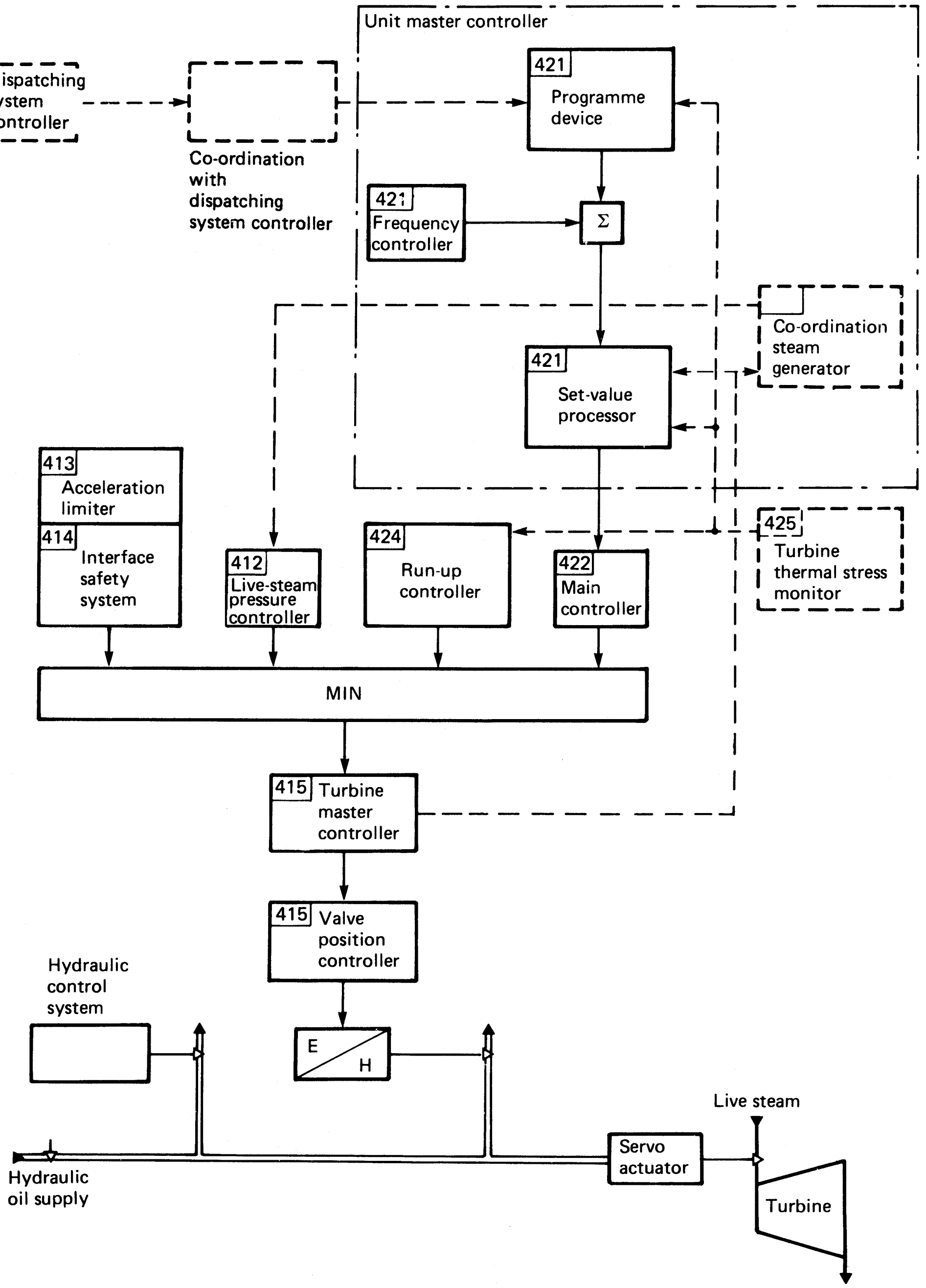

Unit master controller
Dispatching system controller
Co-ordination with dispatching system controller
421 Programme device
421 Frequency controller
Σ
421 Set-value processor
Co-ordination steam generator
413 Acceleration limiter
414 Interface safety system
412 Live-steam pressure controller
424 Run-up controller
422 Main controller
425 Turbine thermal stress monitor
MIN
415 Turbine master controller
415 Valve position controller
Hydraulic control system
E
H
Live steam
Servo actuator
Turbine
Hydraulic oil supply

Figure 40.37 The arrangement of the individual function groups of a turbine controller

valve-position controller (415) is switched on. The controller begins the run-up from any given starting speed to the nominal speed, full use being made of the admissible operating range of the turbo-set. On reaching nominal speed, the frequency controller, which is set to nominal frequency, takes over from the run-up controller. The operation arrived at is 'no load'.

No load During no-load operation the speed of the turbine is determined by the frequency set-point via the control unit (421), the main controller (422), the turbine master station and the valve-position controller (415). No-load operation at nominal speed forms the basis for both the on-line synchronisation of the generator and the overspeed test.

Synchronisation This type of operation is authorised when a signal is received from the synchroniser. The turbine speed is determined by the frequency set-value, this being varied by pulses from the synchroniser until the frequency and phase of the turbo-set coincide with those of the line, and the circuit-breaker closes.

Load operation After synchronising and closing of the generator breakers, loading can commence. The following function groups are then active: programme unit, frequency controller, set-value processor (421), main controller (422), turbine master station and valve-position controller (415).

The programme unit forms a reference variable from the predetermined target load set-value and the adjustable rate of change of load. This variable is added to the output signal of the frequency controller and then reprocessed in the set-value processor. The frequency controller compares the actual frequency value with the predetermined frequency set-value and delivers the amplified difference as an output signal. The set-value processor contains limit circuits which prevent the turbo-set from being overstressed by rapid changes of set-values. The output signal of the set-value processor is compared with the actual value of the active power in the main controller.

Thus, during load operation, continuous load–frequency control is accomplished under the direction of the load programme device, full use being made of the permissible operating range of the turbo-set.

Turbo-set control of live-steam pressure The Turbotrol system and the steam-generator control system here operate in conjunction. The live-steam pressure controller (412) intervenes when a fault occurs in the steam generation system, thereby taking over from the main controller (422). The system (412) overrides the load set-value and controls the live-steam pressure. The turbine output is then determined by the available live-steam flow.

Manual control The turbine master station is an auxiliary unit which is usually set to position ‘auto’, but which automatically changes over to ‘manual’ as soon as an appropriate monitor in the electronic control system has responded. The reference variable for the valve position existing prior to the fault is retained by the turbine master station, while faulty controllers that are not absolutely necessary for the safety of the turbo-set cease to intervene. Function groups 413 and 414 remain permanently on stand-by. Emergency operation may be carried out by manual control for as long as the fault is present.

During manual operation it is the responsibility of the operating staff to ensure that the admissible turbine limits are observed.

Tripping of the turbine Direct link-up of the turbine hydraulic safety system with the control system, and an electrohydraulic link between the Turbotrol and the safety system, ensures closure of the control valves each time the turbine is tripped. In addition, the Turbotrol is made ready for start-up, i.e. the turbine master station changes to ‘manual’ and the ‘zero’ correcting variable is dispatched to the control valves.

Overspeed test When standard tests have been completed and the necessary preparations made in the hydraulic safety circuit, the overspeed test should be carried out. This takes place semi-automatically with the aid of the run-up controller (424) and starts with the turbine in the no-load condition. Initiated manually, the run-up controller accelerates the turbo-set at a defined rate until the first overspeed monitor responds at 110% of nominal speed. A further command, likewise manually authorised, causes the run-up controller to accelerate the turbo–set up to 112% of nominal speed, at which the second overspeed monitor responds. The turbine is then tripped. A fault in the system causing the set-point value to increase to 114% of nominal speed trips the turbine automatically, i.e. independently of the overspeed monitors.

40.6.6.2 Principle of operation

Run-up controller Providing that the change-over unit is set to ‘auto’, a speed programme unit SFH (*Figure 40.38*) adjusts the speed set-value by a certain rate of change (dn/dt). Change-over to ‘auto’ can only be authorised when the speed set-value and the actual value have been matched automatically. The correcting variable of the run-up controller with proportional–integral (PI) response passes to the electrohydraulic transducer over a smallest-value selector and the valve-position controller. It then influences, via the control oil system, the position of the turbine control valves. The critical-speed ranges are passed through with the maximum admissible acceleration.

Once nominal speed is reached, the frequency controller, set to nominal frequency, takes over from the run-up controller, which then ceases to intervene since the speed set-value n_s continues to increase up to 106%.

Load controller with frequency response. Load-frequency set-value The signal for the desired load is formed by the target-load set-value P_z and the adjustable loading rate of change (dP/dt) or rate of change of load shedding ($-dP/dt$). The load set-value P_s is varied accordingly, this taking place via a transfer unit actuated by push-button.

The load component dependent on frequency is formed by a comparison of the frequency set-value f_s and the actual value f; the control deviation is gained by adjusting the frequency droop. If desired, the influence of the frequency controller can be suppressed within an adjustable range f_T.

The frequency-dependent load component and the load set-value are coupled via a summing junction. The output signal of this junction is thus representative of the load to be delivered by the turbine. All units connected beyond this output are either limiters or subsidiary controllers capable of temporary intervention. These are described individually below.

Maximum-load limiter P_{max} An analogue-value generator P_{max} transmits a value that, for the turbine or the steam generator, represents a load limit not to be exceeded under any circumstances.

Load limiter, admissible rate of change of load Neither the steam turbo-set nor the steam generator can contend with sudden large load changes. These occur particularly when control is according to a frequency–load characteristic and when large frequency fluctuations (e.g. line faults) occur. The step-change and rate-of-change limiter limits these load changes to a variable step change (±%) and the remainder to a likewise variable rate of change ($\pm dP/dt$). This limiter can be influenced by a controller in relation to the thermal stresses of the turbine, thus overriding the load set-value.

Minimum-load setting P_{min} This prevents the turbine from being erroneously tripped by the reverse-power relay. This could take place after parallel connection of the generator and the line by the synchroniser. The setting is authorised as soon as the generator breaker closes, and causes the

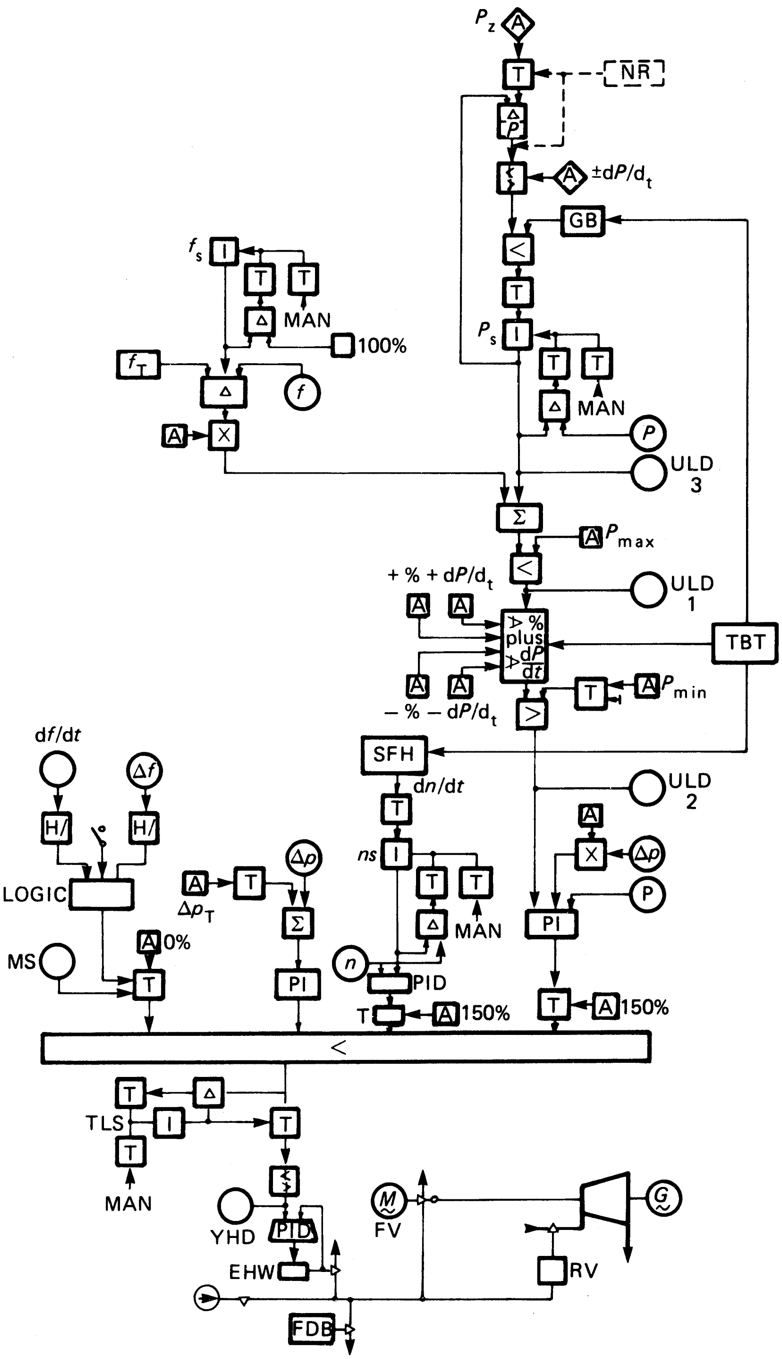

P_z
A
T
NR
ΔP
±dP/dt
GB
<
T
P_s
I
T
T
Δ
MAN
P
ULD 3
f_s
I
T
T
Δ
MAN
100%
f_T
Δ
f
A
X
Σ
A P_{max}
<
ULD 1
+ % + dP/dt
A A
> % plus > dP/dt
TBT
A A
− % − dP/dt
T
A P_{min}
>
SFH
dn/dt
ULD 2
T
ns
I
T
T
Δ
MAN
A
X
Δp
P
PI
df/dt
Δf
H/
H/
LOGIC
A
T
Δp
Δp_T
Σ
A 0%
MS
T
PI
n
PID
T
A 150%
T
A 150%
<
T
Δ
TLS
I
T
T
MAN
YHD
PID
EHW
M
FV
G
RV
FDB

Figure 40.38 Block diagram of the overall arrangement of the turbine controller

generator to deliver a fixed minimum output. As soon as the load set-value exceeds the minimum-load setting, the latter automatically ceases to intervene. It remains, however, on stand-by until the preselected target-load set-value has been attained.

'Unit load demand' signals (ULD 1, 2, 3) These signals can be employed for operation in conjunction with the steam generation control system. A signal relevant to the given application is used, e.g. as a reference variable or as a signal for corrective action.

Load control with live-steam pressure influence For the live-steam pressure controller, a load-controlled turbine represents a controlled object without inherent feedback. The pressure control can be additionally stabilised by applying the control deviation Δp, within certain limits, to the load controller of the turbine. Thus for minor pressure control fluctuations a controlled object with inherent feedback is achieved.

Valve-position controller, turbine control valves and turbine master station The output signal of the lowest-value selector for all controllers represents the reference variable for the valve-position controller and is limited to fixed minimum and maximum values.

To ensure co-ordination of the various PI action controllers (run-up, main and live-stream pressure controllers) during normal operation, the integral part of each controller continually tracks the controller having the smallest registered proportional component.

The turbine master station TLS allotted to the valve-position controller is not used during normal operation but only in the event of faults in, for example, the main controller. In this event the correcting variable received from the main controller is suppressed and the turbine master set to 'manual'. The correcting variables received from the acceleration limiter and the safety-system co-ordinator always remain on stand-by. When the turbine master is set to 'manual', any desired valve position can be commanded. It should, however, be noted that with this type of operation the operating staff are responsible for ensuring that the admissible turbine limits are observed.

Live-steam pressure controller To prevent an excessive drop in the live-steam pressure when faults occur in the steam generator, or when there are rapid positive changes in load, the live-steam pressure controller intervenes on the adjustable limit p_T being attained. This reduces the turbine load in accordance with the available live-steam flow. The limit can, if desired, be set to zero, after which the Turbotrol acts as an initial pressure controller.

Acceleration limiter When load-shedding to house load, this limiter closes the turbine control valves immediately. Simultaneously, the load set-value P_s is brought to zero within less than 1 s. As soon as the acceleration of the turbine fades out, the limiter ceases to intervene. The frequency set-value alone now determines, via the main controller, the speed of the turbine by controlled opening of the valves (operation feeding only the house load).

Safety system When the turbine trips, the pressure switch MS responds. The signal is led via the evaluator logic to the turbine controller and causes all control valves to close immediately.

40.6.6.3 Control room operation and display

Basically, the number and type of units required in a control room for the Turbotrol system vary according to the installation. The functions are actuated by means of non-arresting push-buttons. These may be adapted for use as illuminating buttons. Either solid-state output units or coupling relays should be provided for the alarm signals, depending on the type of warning system.

40.6.6.4 Operating behaviour and maintenance

The philosophy applied to safety, availability and reliability is expounded here without details of the theories of system safety and reliability.

Safety A solid-state Turbotrol control system as described is a single-channel arrangement. In assessing the safety of the entire turbo-set, however, it should be borne in mind that, parallel to the electronic control system, a completely independent hydraulic safety system can be provided with redundant monitors (speed, pressure, vacuum, etc.) in a one-out-of-two arrangement and independent actuators (main and reheat stop valves).

The safety system is then coupled to the control valves both hydraulically via a relay, and electrohydraulically via function group 414. Loss of pressure in the hydraulic fail-safe circuit causes the control valves to receive closing commands over two independent channels.

The Turbotrol system itself is designed to ensure the highest possible degree of safety whilst, with regard to the signal paths, still upholding the principles of simplicity and transparency. The most important peripheral units and transducers are actively redundant.

Two separate d.c. power sources reduce the danger of a turbine trip as a result of failure of the supply voltage. Failure of internal power-supply units simply causes clearly defined subcircuits to be disconnected, fault indication taking place simultaneously. Generally, it is still possible to carry out restricted operation with the Turbotrol. When a fault occurs, some of the internal monitoring elements set the turbine master station immediately to 'manual', so that the turbo-set can at least still be operated manually. Even during such operation, the subcircuits important for the safety of the installation (413, 414) remain active. Response by either of these, or a fault in one of them, causes an immediate trip.

The subcircuits and functions can, to a large extent, be tested during operation.

Availability Regular checks, and the modular design of the equipment, permit faults that occur to be recognised early and eliminated by replacing the appropriate module before it causes operational failure. If a breakdown should nevertheless occur, the fault may be quickly located by means of the many indication, test and simulation devices, and promptly eliminated. The availability of the installation is thus ensured.

Reliability The individual parts of the control system, its components and, finally, the complete system unit should be subjected to rigorous testing. Combined with a simple logically designed system that has the minimum number of component parts, these measures contribute to the high degree of reliability of the entire control and safety system.

Final designs of various kinds, depending on the respective importance attached to the terms safety, availability and reliability, can be achieved.

Operation with the hydraulic control system This control system serves as back-up when the Turbotrol is in operation. If a fault occurs in either the turbine master station or the electro-hydraulic transducer (415), the respective hydraulic lines must be blocked manually. Renewed turbine start-up then takes place manually via the hydraulic control system. The variable readings required for this are displayed in the control room, provided that the d.c. power supply for the Turbotrol is intact.

40.6.6.5 Fast valving

Power export is reduced by the loss of voltage in the event of short circuits in the transmission systems. Acceleration of the turbine rotors occurs in the turbine–generator units involved by a change in the balance of the mechanical-drive and the electrical-load torques. In the turbine-generator unit closest to the short-circuit point, the rotor angle reaches a critical value that depends upon the generated load and the duration of the short circuit; this causes the unit to lose synchronisation.

By briefly closing the turbine control valves ('fast valving'), the drive torque of the turbine–generator unit is rapidly reduced to control a sustained short circuit without loss of synchronism. Depending upon the load in the transmission system, it may be necessary before reconnecting the load to set a load level lower than that before the short circuit.

The subassembly 'Fast valving' (functional group 470—see *Figures 40.39* and *40.40*), takes action in the turbine control system in the event of failures in the power transmission system by brief closure of the control valves in conjunction with a reduction in the load setting. This takes place if the load on the turbine–generator unit is greater than the adjustable maximum value (approximately 60% load). Failure detection in the power transmission system and the initiation of fast valving are performed by the grid supervision system.

Both the main control and the intercept valves take part in fast valving. While the intercept valves are briefly closed (T_{FV}) via solenoid valves, the main control valves are actuated by the Turbotrol electronic control system.

The electrical position signal YHP for the main control valves is set to 0% by a minimum selection gate. The turbine master of the Turbotrol is switched to 'manual'. After the time T_{FV}, YHP is switched to a preselectable value *X*%. Simultaneously, a 'lower' instruction for tracking the turbine master is issued. The integrators of the controllers also follow the tracking of the turbine master. The actual load value then appears according to *X*%. The load set-point is tracked in the unit master.

The load set-point drops until the controlled variable YR is smaller than YFV, i.e. until the main controller is ready to take over the control task. The turbine master then automatically switches back to 'auto'.

The load target set-point has remained stationary during the entire fast-valving process so that it is immediately possible to reconnect the load by switching on the automatic loading system.

It is only possible to test the subassembly with a load of less than an adjustable minimum value (about 30%).

40.6.7 Water-turbine control system (Hydrotrol)

The Hydrotrol (*Figure 40.41*) performs, in conjunction with the electro-hydraulic transducers and the hydraulic servo-motors, the following tasks:

(1) Automatic provision of an opening for run-up of the turbine until the speed controller takes over.
(2) Speed control during no-load operation.
(3) Frequency control during isolated grid operation and when station auxiliaries are being supplied.
(4) Opening control according to an adjustable linear frequency–opening characterisic.
(5) Arresting of overspeed after load shedding.
(6) Connection and processing of signals of a higher-order control system with or without load feedback (e.g. from the water-level controller or dispatching system controller).
(7) Position control of the guide-vane apparatus in Francis turbines.

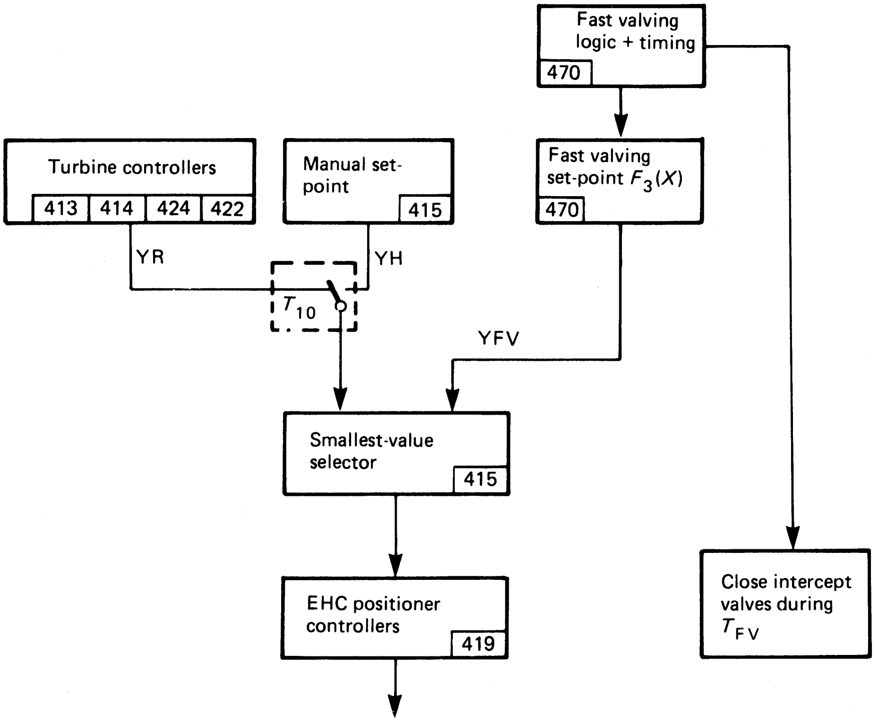

Fast valving logic + timing
470
Fast valving set-point $F_3(X)$
470
Turbine controllers
413
414
424
422
Manual set-point
415
YR
YH
T_{10}
YFV
Smallest-value selector
415
EHC positioner controllers
419
Close intercept valves during T_{FV}

Figure 40.39 Block diagram showing the action of the fast-valving function group (470)

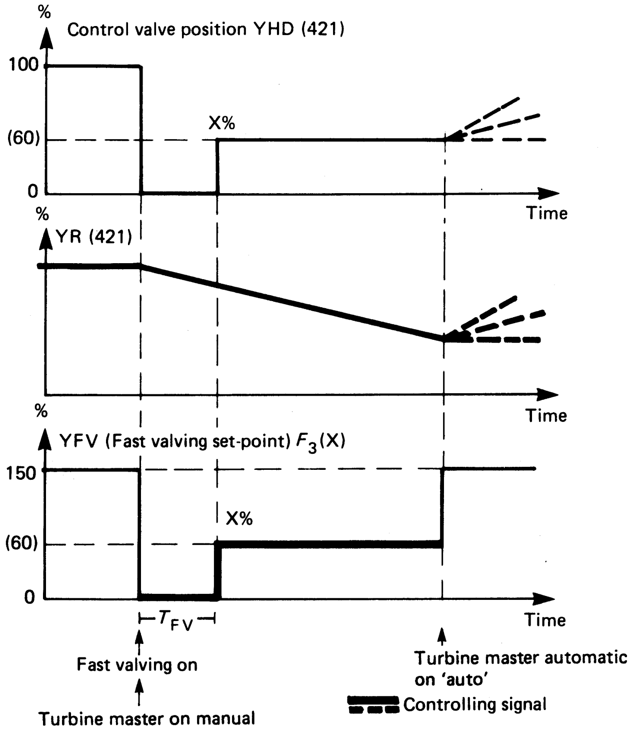


Figure 40.40 The response of the signals YHP (421), YR (421) and YFV during fast valving

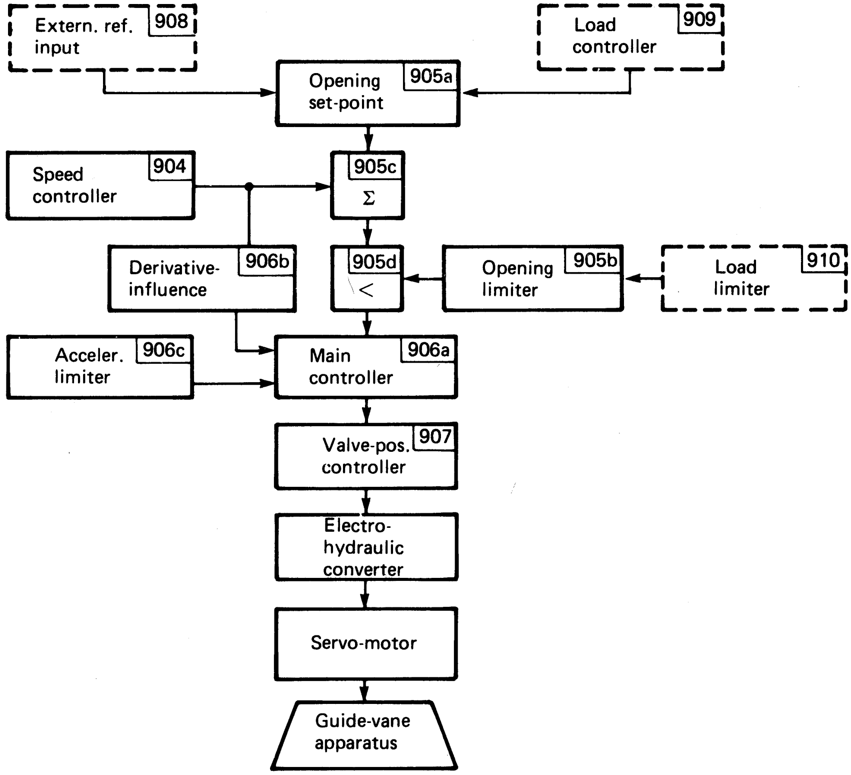

Extern. ref. input 908
Load controller 909
Opening set-point 905a
Speed controller 904
905c Σ
Derivative-influence 906b
905d <
Opening limiter 905b
Load limiter 910
Acceler. limiter 906c
Main controller 906a
Valve-pos. controller 907
Electro-hydraulic converter
Servo-motor
Guide-vane apparatus

Figure 40.41 Function groups of the Hydrotrol with valve-position controller for Francis turbines

(8) Position control of the guide vane and impeller blades (with a pre-set relationship between guide-vane and impeller-blade angles) in Kaplan turbines.
(9) Position control of the needle and deflector (with co-ordination between jet diameter and deflector position) in Pelton turbines.
(10) Redundant measurement of speed. The speed-measurement device is equipped with outputs for indicator devices, control and monitoring.

The Hydrotrol 4 ensures (i) high operating reliability and availability, (ii) high-response sensitivity in respect of changes in specific control variables, and (iii) easier servicing with low maintenance requirements.

40.6.7.1 Design and operation of the controller

Figure 40.41 shows the basic design and the way in which the function groups of the Hydrotrol are interconnected. The numbers of the functional units are referred to in the text. The controller is designed for the following methods of operation:

Normal:
- start-up (with opening limiter);
- no-load operation;
- synchronising;
- load operation with opening feedback;
- load operation with opening limiter in action;
- load operation with dispatching system controller connected;
- load operation with external reference input; shut-down.

Disturbance:
- load shedding;
- turbine tripping.

Testing:
- overspeed;
- simulated operation.

Start-up As soon as the starting command is given, the opening limiter (905b) indicates the starting opening (*Figure 40.41*). In the smallest-value selector (905d) the control deviation of the speed controller (904) is limited by the start opening and is supplied to the main controller (906a). The servo-motor opens the guide-vane equipment via the valve-position controller (907) to the value specified by the opening limiter (905b). The machine accelerates until the speed controller (904) comes into action via the derivative action (906b) below the nominal frequency. Acceleration consequently ceases and the speed controller (904) controls the turbine at nominal frequency.

No-load operation The turbine speed can be set by the frequency set-point value. The control deviation of the speed controller is formed by comparing the set-point with the actual frequency, and passes via the smallest-value selector (905d) to the main controller (906a). In this method of operation the main controller is switched to proportional–integral–differential (PID) response for reasons of stability. No-load operation at nominal speed is the starting point for synchronising the generator with the network.

Synchronising The synchronising device acts on the frequency set-point. Its pulses adjust the set-point until the frequency and phase position of the unit coincide with the network, and the generator breaker then closes.

Load operation with opening feedback After synchronising, the plant changes to load operation with the generator breaker closed. The following function units of the control system are now in operation: speed controller (904); opening set-point (905a); summing junction (905c); smallest-value selector (905d); main controller (906a); valve-position controller (907).

The turbine is loaded by run-up of the opening set-point. This is compared with the opening of the guide-vane equipment; the resultant deviation, after evaluation by the pre-set droop variable, is added to the control deviation of the speed controller. The acceleration influence (derivative action) on the main controller is blocked. In this mode of operation the main controller is switched to PI response.

Load operation with opening limiter in action As soon as the opening of the guide-vane device exceeds the value set on the opening limiter (905b), the latter begins to act via the smallest-value selector (905d) and limits the opening. The load limiter (910) automatically limits the maximum output of the turbine or generator via the opening limiter.

Load operation with applied dispatching system controller This is initiated by actuating a button on the control desk. The opening set-point is controlled by a superposed frequency–load controller (909). The load set-point can be pre-selected by a separate set-point generator or an external dispatching system controller.

Load operation with external reference input This method is pre-selected by a button on the control desk, which causes the opening set-point (905a) to approach the pre-selected reference input (908) from a water-level controller or dispatching system controller. As soon as the two signals agree, the external reference input is applied directly as an opening set-point. The opening set-point tracks the external reference input so that, when this mode of operation is switched off, the controller continues to retain the last-occurring reference input.

Shut-down After the load has been reduced by decreasing the opening set-point, the generator breaker is opened; the controller again operates in the no-load mode. The turbine can then be fully shut down by automatically switching the opening limiter to a closure command of about 10%.

Load shedding To arrest the speed rise of the turbine on load shedding as rapidly as possible, an acceleration limiter (906c) is incorporated. As soon as the acceleration exceeds a limiting value, this device, acting via the servo-motor, causes immediate closure of the guide vanes. When the overspeed vanishes, the speed controller again takes over control at nominal frequency.

Turbine tripping All the emergency-closure and rapid-closure criteria simultaneously have the effect of causing a switch-over to a closure command of -10%, as with the normal shut-down operation.

Overspeed test For this, the speed controller (904) is rendered ineffective. By increasing the speed with the opening limiter (905b), the function of the overspeed protection device can be tested.

Simulated operation This is initiated by inserting a plug in the Hydrotrol 4; it facilitates functional testing of the electronic controller with the plant shut-down, by means of incorporated simulation equipment.

40.6.7.2 *Operation of the main controller*

In order that the control deviation does not remain permanently in the stabilised state, the main controller is constructed with a PI circuit. With the circuit selected, the main controller governs for fairly small control deviations, and the servo-motor operates in its linear rapid-operating range. For large control deviations, the servo-motor can no longer follow the correcting variable (output of the main controller), since the maximum speed of the servo-motor is limited by hydraulic orifice plates. The main controller will disengage, resulting in pronounced overshoot of the servo-motor position (in the extreme case, event instability). However, electronic simulation of these hydraulic orifice plates limits the integral component of the main controller in such a way that, as soon as the servo-motor again reaches the linear range, the main controller immediately corrects itself and can take up optimised control.

Two types of stabilisation are possible in the main controller:

(1) no-load stabilisation, initiated with the generator breaker open (no-load stabilisation corresponds to a reduced proportional component, a long integration time and effective derivative action); and
(2) parallel stabilisation, initiated with the generator breaker closed (parallel stabilisation corresponds to a relatively large proportional component, a short integration time and ineffective derivative action).

If the generator is supplying an isolated network, the no-load stabilisation must be switched on. This occurs automatically when a specific frequency band is exceeded, and is signalled in the control room. It must also be possible to carry out this changeover manually.

40.6.7.3 *Operation of the valve-position controller*

The output signal from the main controller (with PID response) acts upon the valve-position controller. The latter has the task of making the servo-motors of the control devices track the main-controller output (proportional response). It represents the actual connection device between controller and hydraulic system, and serves as a command device of the electrohydraulic transducer.

There are three types of valve-position controller.

(1) For Francis turbines. In this variant, the position of the guide-vane servo-motor is regulated.
(2) For Pelton turbines. In this variant, the positions of the needle and deflector servo-motors are regulated. By simulating a function, ‘needle movement to water-jet diameter’ in the ‘deflector’ control circuit, it is possible to ensure that the deflector is positioned above the water jet. The deflector then comes into action only if rapid control movements in the ‘closed’ direction occur.
(3) For Kaplan turbines. Here the positions of the guide-vane and impeller servo-motors are regulated. Function transmitters ensure optimum tuning between the guide-vane angle and impeller-blade angle as a function of the head. The corresponding device is situated in the ‘impeller’ control circuit.

40.6.7.4 *Structure of the speed-measuring equipment*

The Hydrotrol is equipped with a redundant speed-measuring device, with outputs for indicator devices, control and monitoring. The basic design of the device is shown in *Figure 40.42*. References to the units and outputs are given in the text in parentheses.

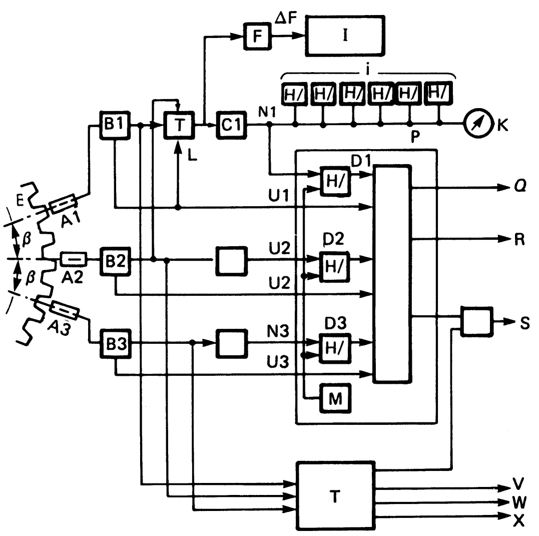


Figure 40.42 Block diagram of a speed-measuring device with speed-limiting values, overspeed protection and direction indicator: D/A, Digital/analogue

The speed is measured with three channels. The speed is picked up from a wheel (E) by ferrostat transmitters (A1, A2, A3). The generated frequency (proportional to speed) is transmitted via the pre-amplifier (B1, B2, B3) and the speed-measuring device (C1, C2, C3). The speed measurement generates an analogue voltage (N1, N2, N3) which is accurately proportional to the speed. The speed transmitters are monitored: a fault at speed transmitter A1 (such as a voltage failure, short circuit or line fault), is signalled via monitoring channel U1 and is annunciated by the alarm ‘speed transmitter faulty’ (output S). Simultaneously, the speed signal for the speed controller (I), the speed limiting-values (J) and the indicator (K) are switched over from speed transmitter A1 to transmitter A2.

The overspeed protection (P) consists of a double-channel ‘two-out-of-three logic’ (L). This triggers an emergency stop (output Q) and a quick shut-down (output R) as soon as two or three speed limiting values (D1, D2, D3) respond or faults occur at two or three speed transmitters. The direction indicator (T) serves for determining the direction of rotation in pump turbines.

The speed transmitters must be mounted at angular intervals $\beta = 360(n + \frac{2}{3})/z$ (degrees), where z is the number of gearwheel teeth and n is an integer. The outputs V and W indicate the direction of rotation, and output X indicates when the shaft is stationary.

40.6.7.5 *Testing*

Component testing Every component approved for use in electronic systems must be subjected to stringent tests, carried out in compliance with IEC publications.

Testing of the controller The controllers must be individually tested and adjusted in the laboratory on the basis of the particular turbine characteristics. Before delivery, every Hydrotrol must operate for a week in the closed-loop of an analogue turbine simulator, with all control variables and the most important control parameters continuously recorded. In this way, changes in control behaviour and premature failure, for example of semiconductor components, can be detected.

40.6.7.6 Operational reliability and availability

Reliability An electronic water-turbine Hydrotrol is, in general, of single-channel construction. The important peripheral devices and transmitters (e.g. for speed) are designed with redundancy. A failure in the supply voltage or internal supply equipment causes rapid closure and simultaneous signalling of the fault. Faults in internal monitoring devices such as those for the speed transmitters or valve-position controllers lead to an emergency closure.

Availability Regular testing and the modular construction of the equipment should permit any faults that occur to be detected at an early stage and to be rectified by replacement equipment before they cause operating faults.

40.7 Decentralised control: substation automation

40.7.1 Introduction

With the increasing complexity of power systems, the difficulties of manning substations and the requirement for shorter restoration times after power failures, more tasks are being automated in the substation. The conventional solution to automating these tasks has been by dedicated hardwired logic. Though effective, this is inflexible and difficult to test off-line. Hardwired logic is being superseded by systems based on mini- or microprocesors which have advantages over their hardwired counterparts with regard to flexibility, space requirements, test and commissioning facilities. A further and important advantage is their ability to self-monitor and detect an internal fault almost immediately, and not only when operating. However, they have the disadvantage of intolerance to voltages in the connecting cables induced by switching of h.v. apparatus or thunderstorms. Special precautions must therefore be taken to decouple the plant connections from the electronic circuits.

Several tasks can be assembled in one set of hardware, and thus a centralised system can be used. The speed of transmission and the availability of transmission channels are limiting factors in achieving reasonable response times for some substation tasks from a remote control centre; many of these tasks are therefore performed by substation equipment.

40.7.2 Hardware configuration

A simplified substation automation hardware configuration is illustrated in *Figure 40.43*, with its connections to one feeder circuit of an h.v. substation.

The central logic is a microprocessor with the automation routines stored in a non-volatile programmable read-only memory (EPROM) which can be erased and modified using special equipment. Thus, by triggering a power-failure restart routine held in the EPROM, the equipment can recover from power failures without reloading. The variable data are held in a random-access memory (RAM) which can be corrupted on power failure. It must be capable of replacement after a power failure either by local input or by down-loading from a higher level control system. Similarly, the variable data, such as limits tripping priorities, etc., must be capable of modification, e.g. from a keyboard.

The data acquisition and automation programs are executed by the microprocessor, which communicates with the input/output equipment via the system bus. The input

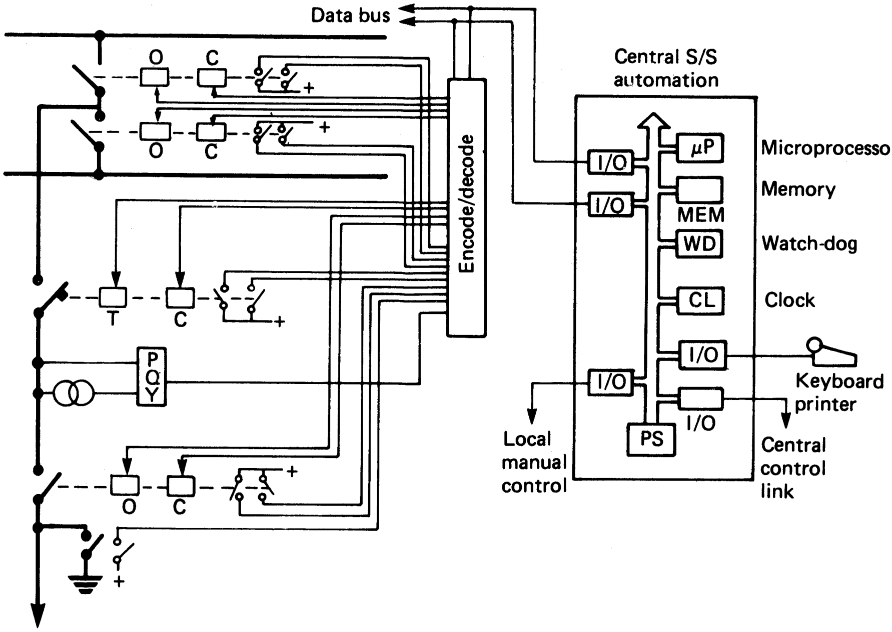


Figure 40.43 An example of a decentralised automation system: O, open; C, closed; T, trip; *P*, active power; *Q*, reactive power; *V*, voltage; PS, power supply; S/S, substation

and output modules are similar to those already described for the telecontrol remote terminal unit (TRU). The connection to the higher level control system can be via a direct link or a telecontrol TRU.

An important decision to be made in the layout of an automation system in substations is the location of the transducers and the method of transmitting the h.v.-circuit data to the central processor unit. Economically, the transducers are better located near to the measurement point to avoid long instrument–transformer connections. However, this means that light-current connections are long and vulnerable to interference and high induced voltages. A solution is to install encoding equipment at the data source and serial transmission equipment to send the messages to the central logic over a serial data bus which can be common to a number of h.v. circuits, as in the telecontrol system transmission already described. As the technique develops, this transmission can be over a fibre-optic line.

40.7.3 Software configuration

The software is modular to allow the system to be easily extended and developed. The individual application programs run independently of each other, but share data and input/output routines. For small systems the programs can run cyclically; however, for larger systems a system of priorities must be established and the resources allocated by an executive program. The program structure is illustrated in *Figure 40.44*.

40.7.4 Applications

40.7.4.1 Sequence control (switching programs)

In closing circuit-breakers on to live circuits, several criteria have to be met to ensure a successful operation; for example, in energising a feeder, the breaker must first be connected to the required bus-bar, the line isolator closed, the line–earth switch open, and the synchronising criteria satisfied. It must be ensured that no sanctions are valid that would inhibit the operation. Similarly, when energising a transformer circuit, manually or automatically, it is normal to make similar checks and circuit preselections before closing the l.v. breaker and then the h.v. breaker. In changing a circuit from one bus-bar to another, a switching sequence involving the bus coupler and perhaps the bus section switches must be performed.

These sequences can be programmed and performed by local automation equipment, provided that all switches are motorised. The sequences can be initiated by one command (e.g. 'close line 123 on to bus 3'). The responsibility from there on is taken from the operator and the instruction is performed by the local automation equipment which, in carrying it out, will respect all the interlocking requirements and inhibits.

40.7.4.2 Switching interlock supervision

Manual switch operation in a substation is required either as the normal mode of operation or as a back-up to the automatic mode. For these operations, interlocking is required to prevent incorrect switching that might endanger equipment or the stability of the power network. The conventional solution for interlocking is a device with auxiliary contacts and inter-device wiring forming a hardwired supervision logic. Interlocking can also be achieved by the local automation equipment using programmed logic accessing the shared data base. This latter solution considerably reduces the inter-device cabling as well as the number of device auxiliary switches or repeater relays used and their discrepancy supervision.

40.7.4.3 Load shedding and restoration

With a falling frequency, at certain levels of frequency it may be necessary to reduce load without waiting for remote intervention or relying on transmission channels. The load shedding can be performed automatically by local automation equipment in progressive steps, starting with the tripping of storage heating or cooling loads via ripple control signals, and progressing to more drastic action by tripping feeder circuits at the substation.

These tasks can be performed by substation equipment that scans the frequency measurand and, at various pre-set frequency thresholds, instigates load-shedding action in prearranged steps and programs. The programs can be arranged to cycle the load shedding so that each consumer

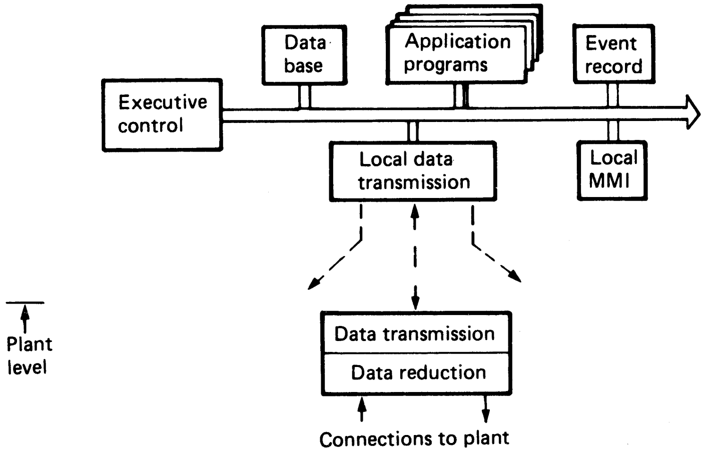


Figure 40.44 Software configuration. MMI

is tripped in turn and no particular load is singled out to be always first on the tripping list.

Once the frequency has returned to predetermined levels, supplies to loads can be progressively restored in a predetermined order until the situation is normal.

Similar actions can be initiated on a feeder overload when the current exceeds the thermal limit of the cable.

40.7.4.4 Automatic reclosing

The majority of overhead line faults are transient, caused by falling trees, birds, lightning, etc., and they clear after the breaker interrupts the fault current. To reduce the duration of the outage, the breaker can be automatically reclosed, after a time delay to allow for deionisation of the fault path. If after the first interruption the fault has not cleared, either the breaker is locked open or further reclosing attempts are made before locking the breaker open. Tripping and reclosing can be either one-phase or three-phase, depending on the selectivity of the protection and the construction of the breaker.

Conventionally, these reclosing sequences are performed by timing relays and hardwired logic requiring dedicated breaker auxiliary switches. Settings of the number of reclosures, the dead time between reclosures, and the reclaim time are adjustable. Equally, these functions can be performed by common logic using a shared database to reduce wiring and auxiliary contacts on the breaker. An additional advantage appears with microprocessor-based logic in that parameters such as dead times and one-phase or three-phase reclosing can be selected and down-loaded from the control centre, depending on the network conditions.

40.7.4.5 Event recording

Event recording in a substation can be divided into two basic categories: one that records alarms, contact closures or device operations, and another that records waveforms of currents and voltages prior to, during and immediately after a fault on the power system.

The first category detects changes of state of contacts with a time resolution of 10 ms and records the operation on a printer. Thus, in post-event analysis, the maintenance engineer has a record of the times at which the events occurred and their chronological sequence.

The second category is an 'oscillo-perturbo-graph', which continuously records, temporarily memorises and then erases the waveforms of the currents and voltages in a circuit. After the detection of a fault, the memorised pre-fault, fault and post-fault data (which can include contact operation) are permanently recorded for subsequent analysis. The conventional equipment for faulting recording is an electromechanical device that has a galvanometer logger writing the waveforms on to a rotating inked cylinder. At the end of the cylinder rotation, the marks made by the recording pen are erased and replaced with a later value. When a fault is detected, paper is brought into contact with the rotating cylinder to print the recorded parameters. This device requires periodic maintenance, but it has been effective for many years. It can be superseded by a microprocessor-based digital system but, as the scanning speed required to record the fault parameters is of the order of 1000 Hz, it is unlikely that this function can be incorporated in a central substation logic. One advantage of digital recording is that the data can be transmitted to a remote control centre for subsequent post-fault analysis.

40.7.4.6 Protection

Protection is a vast subject, the principles of which have been covered elsewhere. It suffices here to mention it as a part of substation automation. Because of the high scanning speeds required, it is unlikely that primary protection can be integrated into a central substation automation system. Thus, even though microprocessor-based protection will be available, it will be dedicated to a particular task or circuit as with the present solutions.

40.8 Decentralised control: pulse controllers for voltage control with tap-changing transformers

As a result of ever-increasing automation and rationalisation, electronic pulse controllers are becoming widely used in the field of tap-changer controls. In interconnected operation, maintenance of frequency characterises the equilibrium between generation and consumption of active power, while voltage control determines the control of reactive power in the system. The essential difference between the frequency/active-power and the voltage/reactive-power characteristics is that the frequency has the same value throughout the system while the voltage forms a system of ever-varying peaks and valleys (assuming a constant rated voltage) which in turn decides the direction and magnitude of reactive power flow.

Tap-changing transformers are the variable step functions in this range of peaks and valleys, and this introduces additional freedom at certain points in the interconnected system. Depending on their place of application, tap-changing transformers can be generator or interconnecting or consumer-load units.

The task of the pulse control unit is automatically to maintain the voltage in the system or to direct the reactive power flow. *Figure 40.45* illustrates an example of application—the control of a consumer network.

The voltage to be regulated is compared in the regulator with an adjustable reference value. Depending on the polarity and magnitude of the difference between these values, 'higher' or 'lower' impulses are given, resulting in the necessary adjustment of the tap-changer. The impulse sequence is inversely proportional to the difference signal. The integration time t_1 and the pulse duration t_2 are adjustable. As long as the difference signal is smaller than the set sensitivity, the impulses are blocked. A further adaptation of the impulse sequence is possible owing to a time factor so that a stable and quasi-steady regulation may always be obtained. *Figure 40.46* shows a typical characteristic of a final control element with a stepwise mode of operation.

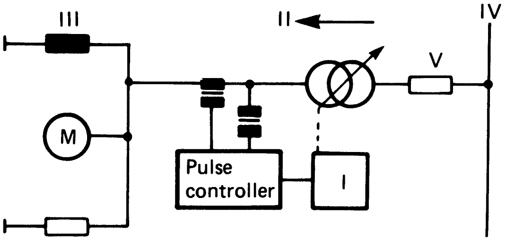


Figure 40.45 Consumer network regulated by a tapped transformer: I, tap-changer; II, reactive-power flow; III, consumer network; IV, supply network; V, line reactance

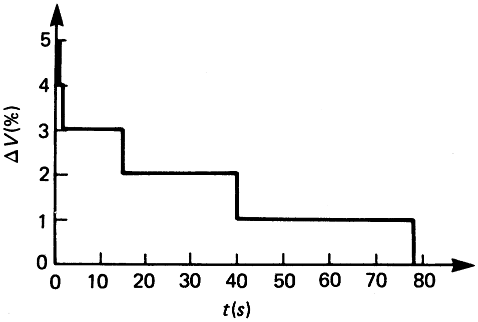


Figure 40.46 Typical characteristic of a final control element with steps of 1%. $\Delta V = 5\%$, $\alpha = 0.25$, $\epsilon = 0.8\%$, $t_1 = 50$ s, $t_2 = 0.2$ s, $t_3 = 0.5$ s

It is important that the controller does not cause the tap-changer to carry out unnecessary switching operations. If, for example, a short-time fault were to occur on the system, no purpose would be served by the tap-changer responding. The fault would be rectified before the switching operation could be completed and the controller would then immediately issue commands for the tap-changer to return to the original position. Adequate damping is therefore essential for preserving the life of the tap-changer.

If the control deviation tends to exceed the set sensitivity, an integrator comes into operation. No control command is issued before the pre-set integration time expires.

40.9 Centralised control

40.9.1 Hardware and software systems

In a control centre, the basic requirement of the operator is information about the power system, presented to him/her in a clear and unambiguous manner. He/she needs an overview of the network, usually in the form of a large wall diagram or overhead projection giving the network configuration and perhaps coarse line load levels. He/she also needs detailed diagrams of individual parts of the network showing the load of each HV circuit plus its switching configuration by isolator, circuit-breaker and earth-switch positions. These circuit diagrams are selectable on to colour visual display units (VDUs). *Figure 40.47* gives an overview of a typical control room.

In addition to the basic requirements, the operator requires other information to be calculated and automatic controls to be performed, as discussed later. These requirements can be realistically met only by an on-line, real-time computer system interfacing with the telecontrol system to provide up to date network data and to accept commands.

The availability of such a system must be of a very high order (99.5%plus); a single processor system would not be adequate. Distributed processing with dedicated multi-processor systems must therefore be considered. The classic configuration of the dual main computer system with front-end processors for the telecontrol system, and an independent wall-diagram control has given way to a distributed system. In a distributed system any of the processors may be duplicated to increase the availability of the individual critical components.

40.9.2 Hardware configuration

A typical computer hardware is shown as in *Figure 40.48, which* allows various groups of applications to be distributed on different processes. The backbone of this configuration is a Local Area Network compliant to the IEEE 802.3. The typical speed of this network is from 10 to 100 Mb/sec. The most common network protocol used on the LAN is TCP/IP corresponding to transport and IP layer of the seven-layer ISO open system model. The LAN itself could be in few segments and redundant to make it failure tolerant.

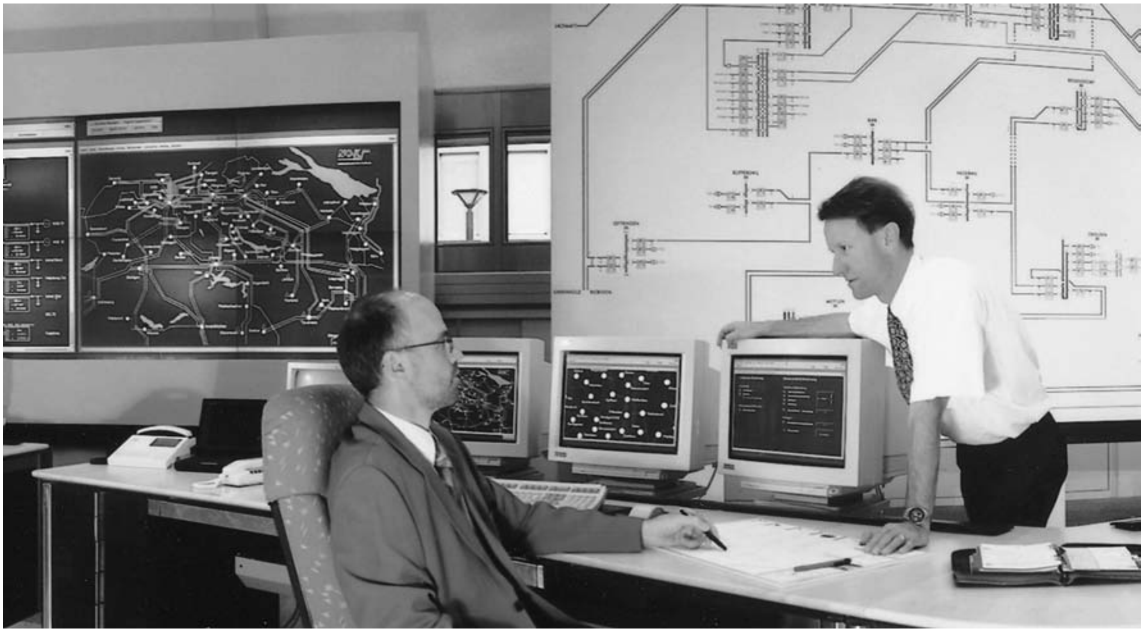


Figure 40.47 Overview of a typical control room

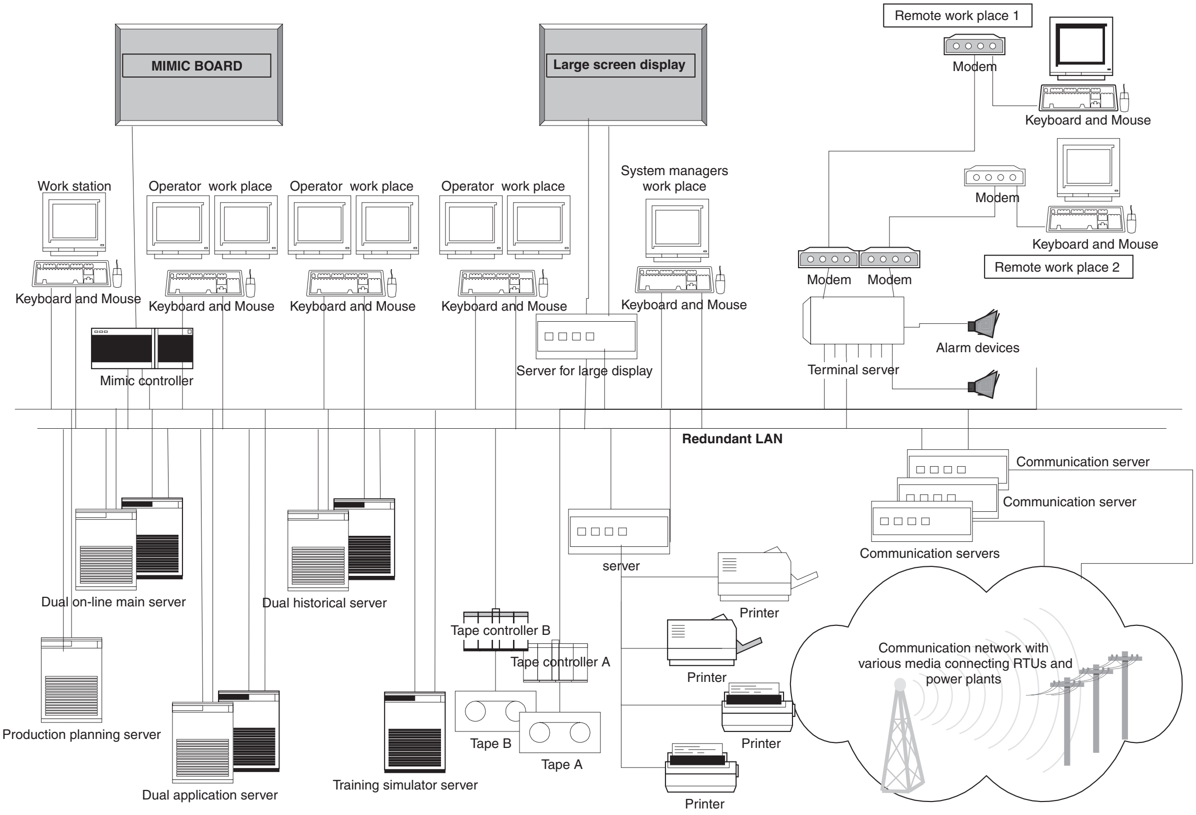

MIMIC BOARD
Large screen display
Remote work place 1
Modem
Keyboard and Mouse
Work station
Operator work place
Operator work place
Operator work place
System managers work place
Modem
Keyboard and Mouse
Remote work place 2
Keyboard and Mouse
Keyboard and Mouse
Keyboard and Mouse
Keyboard and Mouse
Keyboard and Mouse
Modem
Modem
Mimic controller
Server for large display
Terminal server
Alarm devices
Redundant LAN
Communication server
Communication server
Communication servers
server
Dual on-line main server
Dual historical server
Printer
Tape controller B
Tape controller A
Printer
Communication network with various media connecting RTUs and power plants
Production planning server
Tape B
Tape A
Printer
Dual application server
Training simulator server
Printer

Figure 40.48 Distributed control centre configuration

Individual processors can operate in on-line or redundant backup mode. The peripheral devices such as printers, loggers, hardcopy devices, acoustic alarms, Human Machine Interface devices are also connected to the Local Area Network. In case of failure of the main device the backup or standby device takes over without loss of functionality. Such a configuration can be tailored to suit individual reliability and availability requirements.

The communication servers are also used in multiple configurations for load sharing and redundancy purposes. The communication lines from the RTUs, or other control centres bring the data to these communication servers. The basic object of the communication servers is to collect data from the field devices, pre-process the data for change, limit violation, convert them to engineering units and make them available for all the users. A device supervision function keeps watch on all the devices configured in the system and in the event of the failure of the on-line or main device the watchdog may switch the function to the designated backup or standby machine. The main processor communicates with the remote terminals via the communication server systems and with the man machine interface equipment connected to the LAN or WAN.

40.9.3 Man–machine interface

The operation peripherals forming the man–machine interface equipment are connected over the LAN to various processors. These can be configured and switched either individually or as a group to main, standby or backup servers. Normally, these are switched to the main server. The main human machine interfaces devices are the keyboards and the graphic or semi-graphic VDUs, mouse or track ball type pointing device, acoustic device for alarms, wall diagram, printers, loggers, hardcopy as shown in the hardware configuration diagram (*Figure 40.48*).

The VDUs with the mouse and keyboard are the main tools of interaction. Each operator's place may have several VDUs driven by one HMI processor. Each of these VDUs may be configured to display several windows of different types of information. All windows may continuously display updated information from the network but at a time only one of the windows is active for the operator interaction. These display windows may be used to display geographic or schematic network diagrams, alphanumeric information such as reports and lists, trend curves and other type of pictures. The selection or navigation through these diagrams is done via a fixed or dynamic menu, which may popup or be pulled down depending on the context of the cursor or pointer location. The operator may pan in a large picture to locate himself in the area of interest for monitoring and control. He may zoom in to display more detailed information. A certain zoom level may be associated with the level of information to be displayed. These panning, zooming and de-cluttering techniques may be very effectively used to display or condense information on the level of information required by the user.

The displays on the VDU Windows may consist of both static and dynamic data, the later being updated as often as necessary. For example, when a station diagram is displayed, the switch-position indications are automatically updated after the information of the change of state has been received from the substation. Measurand values are also updated on the screen at the same rate as the telecontrol cycle. For trend curves, the screen is used rather like a multi-pen chart recorder, using colours to identify the curves.

The keyboards contain both functional keys for operations that are repeated frequently, and alphanumeric keys for inputting numerical data and text. The keyboards are interactive with the displays on the VDUs, which allows parameter changes and device control by identification of the object to be addressed by device or position reference input, or via the function and alphanumeric keys, or by the positioning of a cursor. The cursor movement is controlled by a mouse, trackball, joystick or direction keys. Thus a dialogue is possible between the operator and the computer system to select displays, to give commands, and to input data for limits, set-points, calculation parameters, etc. Text, for tagging or recording or sanctions, can be temporarily added to the displays, entered via the alphanumeric keyboard.

40.9.4 Wall diagram

The large wall diagram also called Mimic board is used to give the operator an overview of the power system network. The large screen displays using LCD or projection technology serves the similar purpose. On the mimic board however, few details need to be given, as these are available on the VDU screens, whereas large screen or projectors can display the same level of information as on the VDUs with all the full-graphic functions. A typical wall diagram can be used as a backup to the VDU system and, in addition to showing the network topology, it can also display the loading of the network with line load indicators. In their simplest form, the line load indicators give direction of power flow and load level in quartiles, by lamps or light emitting diodes. The wall diagram can also be used to display some alarms if, for example, a substation has an alarm state current or if it has been blocked from the telecontrol scan.

The drive unit for the wall diagram can be the main computer or a separate unit deriving the data direct from the communication servers. In the latter case, the wall-diagram system can serve as a back-up control system in the event of the main computer being unavailable. If a separate console is added, commands can be sent and a limited number of measurands can be displayed under emergency conditions. The large screen can be used as a full-fledged human machine interface. Wall maps or large screens are very useful when more than one operator are interested in monitoring a larger area of the network.

40.9.5 Hard copy

Events and alarms are recorded for subsequent analysis on event and alarm printers. Printers and hard copying devices connected to the LAN may also be primary and backup devices. If any of the printer fails, the print message will be automatically directed to the alternative unit. The printers which operate at a lower speed are suitable for small amounts of data; however, output from a study program producing large quantities of data would be sent

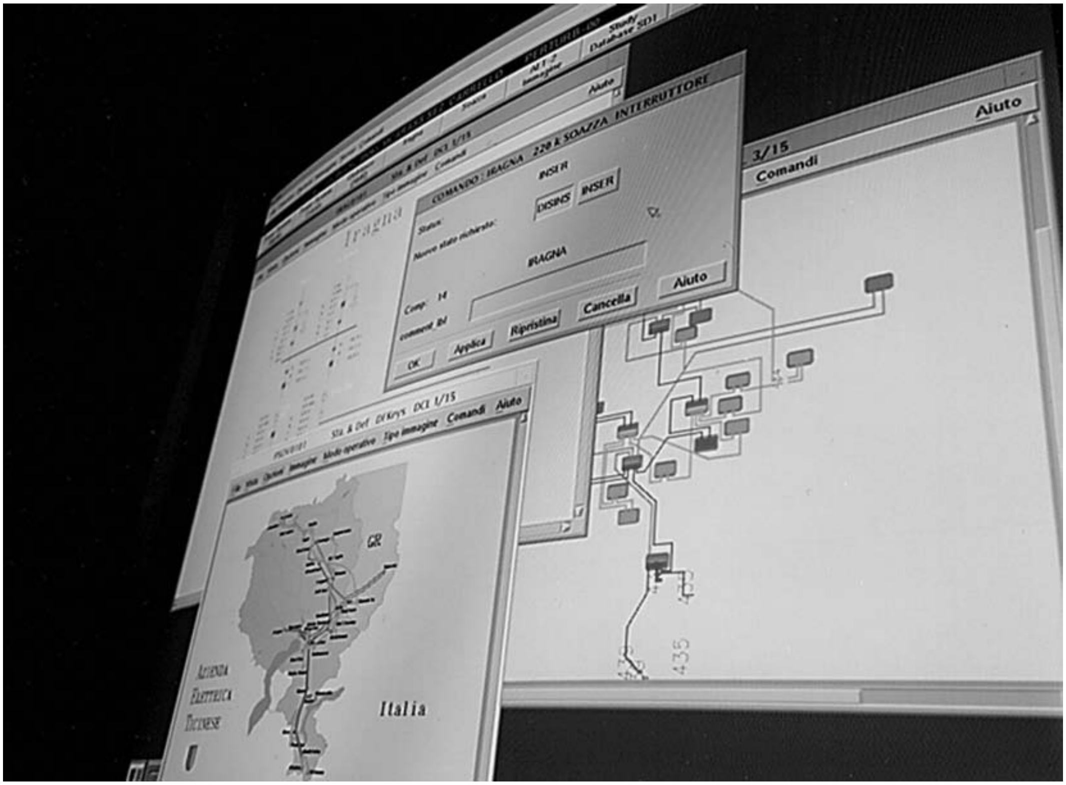


Figure 40.49 Layered and zoomed full graphic display windows

to high-speed printers. Modern laser printers, which can print graphics in addition to the alphanumeric characters, are used as printers in the control room. Typically event and alarm logs are printed on the slower printer on endless paper.

Other hardcopy devices are available that can copy either alphanumeric or graphic data displayed on the VDU screen, in colour to make a permanent record of a display existing at a particular moment. Such devices temporarily freeze the VDU picture and take a snapshot of the screen and then come back to continuous update mode. It is typical to have a VDU hardcopy in A4 or A3 format. For distribution networks hardcopies up to size A0 are plotted on large plotters.

40.9.6 Software configuration

The software system of a computer based control system consists of many individual tasks which fall into three basic categories of real-time, extended real-time and batch or background processing. The programs of the real-time group, such as data acquisition, human machine interface, automatic generator control, etc., have to respond to external events within a given time. The response time of extended real-time programs, such as state estimation or economic dispatch control, is not so critical, though these programs work with real-time data. Extended real-time programs, and batch programs, have no stringent response time definition.

In a centralised network control system, there may be several hundred programs and subroutines all competing for the limited hardware resources; they must therefore share the computer time and store. Consequently, an overall co-ordination system is required to allocate the processor time and the memory space and access. A real-time multi-programming operating system and a database management system ensure that all programs can share the computer system *resources and* all data, without mutual exclusions and inter-program or data corruption.

The heart of an operating system is the real-time executive, which allocates the computer *resources*, in order of priority, to the programs requiring them at a particular moment. When *real-time or* extended real-time programs are required to run, they need high priority on the system *resources including* main memory space and processor time. The *executive allocates* the *resource in* order of priority by assigning partitions of main memory to a program and the appropriate input/output handler routines. When the higher priority programs do not require the *resources, the* batch (off-line) programs can run in turn. Programs can be interrupted at checkpoints in their sequence and be delayed while a program of higher priority is executed.

The priority control is *triggered* by hardware and software interrupts which can be externally or internally generated. The interrupts are assigned *levels of* priority. Several interrupts can be assigned to the same *level*, *in* which case they are *queued* to run when the resources are available to that priority level.

40.9.7 Memory management

The main memory must be shared *between many* programs which cannot all reside simultaneously in the main memory. Thus they must be loaded into a partition of the main memory, from the mass memory (usually disc) when they are required to run. Once a program has been *executed*, the partition can be liberated for another program. The memory management system governs the allocation of the main memory, which may also *require rearrangement of* the space already allocated to obtain a contiguous area for a partition large enough for a program that has been called to run. The operating systems of modern computers take over the memory management tasks and the memory allocation is done dynamically for multi-user and multi-process computing environment.

Programs that run frequently and need fast *response times* are normally resident in the main memory, as this avoids loading time from disc, which may be of the order of hundred milliseconds. It is obviously desirable to *keep the* disc access to a minimum, but it *becomes an* unavoidable overhead for programs that overflow their partition allocation.

40.9.8 Input/output control

The operating system incorporates *peripheral device drive* routines, which are available for all application programs and can be called to read data from, or write data to, the *peripheral devices*. The driver conducts the communication with the device and reports malfunction or non-availability of the device. The operating system temporarily allocates the driver to an application program requesting it.

40.9.9 Scheduling

Some programs (e.g. automatic generation control (AGC)) must run periodically, others must run at certain times of the day (e.g. daily log), and yet others after a certain delay (e.g. time outs). All timers for this program scheduling are available in the operating system, which also keeps track of the date. The computer time can be periodically *corrected by reference to* a ‘standard’ time unit external to the computer system. Typically this could be an input signal from a global positioning system (GPS) or a radio. The GPS provides a cost effective and accurate measurement of time. Using such a signal the time may be synchronised within a few microseconds of each other.

40.9.10 Error recovery

An important part of the operating system lies in the *detection* of errors in hardware and software, and in taking subsequent action to *ensure the* security of the system. The errors may be *device errors* which, if the device is redundant, do not jeopardise the satisfactory running of the system. In the case of serious errors from which the system cannot *recover* (e.g. disc drive or a *recurrent parity* failure), the errors must initiate a switch-over to the standby machine.

At the detection of a power failure, the *executive arranges* that the contents of the volatile registers are stored in the non-volatile memory before the system halts. When the power supply is restored, the system can restart automatically under control of the executive, which restores the system to its former state. However, where program operation (e.g. data acquisition or AGC) is affected by an extended power failure, reinitialising is necessary.

40.9.11 Program development

Throughout the life of the control system, power network extensions will have to be added to the database, and program development will be required without interference with the on-line systems. The operating system must permit and assist this work; hence it must include language compilers (for languages such as FORTRAN, Coral, Pascal and Assembler, C, C++)—as well as editing and debugging facilities. The present trend is to provide application-programming interfaces, which can be used by the programs in higher language, and the program developer does not have to take care of the physical layout of memory or location of the programs.

40.9.12 Inter processor communication

If the control centre is a part of a hierarchical network control system and must be capable of exchanging data with its neighbour, a method of processor to processor communication is necessary. The communication system should have minimal effect on the computer system; thus it should have direct memory access to perform cyclic redundancy encoding and checking (CRC) with separate hardware logic and buffers. The security of the code should have a minimum hamming distance of 4 and automatic repetition of messages if errors are detected. Such types of mechanism are used by master station to RTU communication. In a hierarchical control centre the communication may also be based on protocols developed for such communication. Inter centre communication protocol (ICCP) developed by EPRI and adopted by IEC are a step in standardising this communication, irrespective of the supplier of the control centre.

40.9.13 Database

The many programs and subroutines that make up the control centre software require access to data that, in many cases, are common to several programs. For example, the display update routines and the state estimation program both require the same measurands supplied by the data acquisition program. To avoid the necessity of passing data, created or acquired by one program, to many different programs that may require them, is more convenient, secure and economical to have the data stored once only. If the central data storage is organised and manipulated by a data management system, the data storage becomes independent of any application program. Data reorganisation then necessitates no modifications to the programs using the data, an essential requirement if developments and extensions are to take place economically and with the minimum interference to the working system.

The data base management system (DBMS) organises and keeps track of all system data and makes them available to any authorised program, through standard access routines. Thus, in a multi-programming environment, a certain discipline is imposed on the data users, which prevents accidental corruption of the system data. The management system also guarantees consistency of the data: for example, when an interdependent set of data are being modified, access by other programs must be prevented until modification is complete. To access the data, the user program holds the data access references and, when calling for the data, presents these references to the access routines. The access routines refer to a 'schema' that contains a unique definition of each data item in the database. Thus, as long as the references do not change, even though the data have changed their physical position, the user program is unaffected by database extensions or reorganisation.

In present day control centres the core functionality, which is time critical, is handled by a fast and robust database, which is in most cases a proprietary database. The slower information and large amount of data is stored and managed by a commercially available Data Base Management system, which may be relational or object oriented in their approach. The advantage of commercial DBMS is their openness and ease of access for all the authorised users. Such databases may be physically distributed over several processors and can be accessed over local or wide area network. All the redundancy, security levels are configurable and the application layers are published and well established. Each developer may use these application-programming interfaces to the data.

40.9.14 System software structure

The power system network is continuously expanding as new lines and substations are commissioned. The software structure must allow the control system to expand in step with the network to include not only additional data but also the new network control facilities demanded as the complexity of the network increases. The software structure must therefore be flexible so that modifications or additions have minimal effect on the rest of the system.

The ideal method of achieving this goal lies in the modularity of all application programs, whose execution is organised by the operating system and by co-operating with other programs via the database (see *Figure 40.50*). However, with well established co-operating programs, in the interest of response time this is not always necessary or even desirable, thus, in certain cases, it is advantageous to exchange data directly between programs.

40.10 System operation

The various controls and the incidence of a vast number of disturbances make the operation of a system a very complex task. It is manageable only by the allocation of appropriate subtasks to different hierarchical levels; this breaks down the problems, which can then be handled effectively by man or machine.

So long as changes in load and topology, and fault incidence, appear as foreseen in the planning stage, the power system is able to maintain stable conditions through the action of decentralised control. This is true for protective devices and voltage, load and frequency control. However, whenever topological changes take place or become necessary, limits are reached or potential risks appear, control actions that cannot be handled at the decentralised level are imminent. This is where centralised control has to take on its important role. To what degree these control actions can be executed by automatic means is still an open question. The only centralised automatic control system is that for load frequency control (automatic generator control) together with economic dispatching. Switching operations, and start-up and disconnection of generators, are still subject to manual intervention. However, extensive support for the human operation is provided by the computer based control system. Primarily, it provides the real time information about incidents in the power system; next, it forecasts load development and the effects of possible faults, facilitating efficient decision making. Beyond that there is a host of control functions, which can be conceived as a black box, having a set of inputs and a set of outputs. They can be called upon by manual intervention or by another function. The output is available after the lapse of a certain response time. The mode of using these functions is either repetitive or event driven.

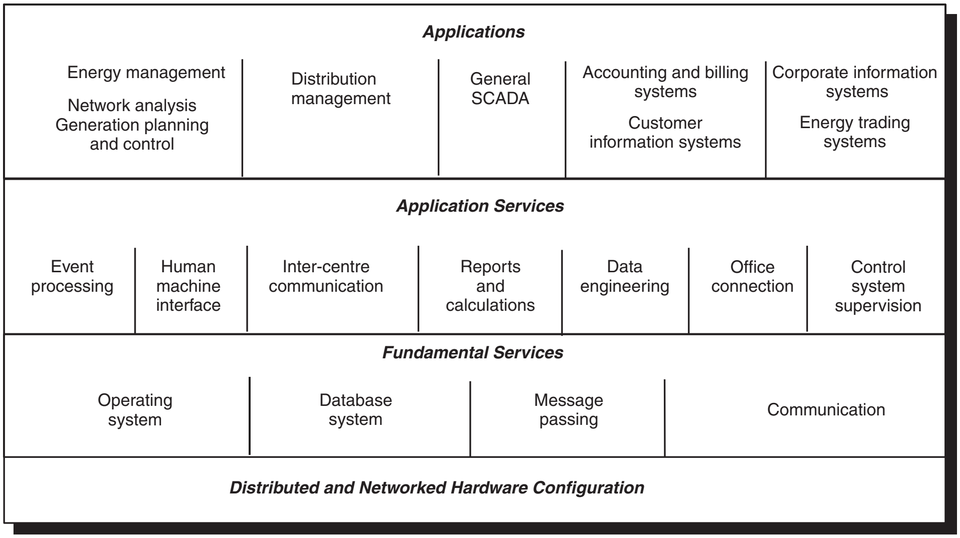

Applications
Energy management
Network analysis
Generation planning and control
Distribution management
General SCADA
Accounting and billing systems
Customer information systems
Corporate information systems
Energy trading systems
Application Services
Event processing
Human machine interface
Inter-centre communication
Reports and calculations
Data engineering
Office connection
Control system supervision
Fundamental Services
Operating system
Database system
Message passing
Communication
Distributed and Networked Hardware Configuration

Figure 40.50 Software configuration for control centre

As seen, most of these system control functions are implemented in terms of digital programs, i.e. by software. However, some functions are realised by hardware.

In order to illustrate typical software necessary for the control of a power system, the structure of the framework of functions is given in *Figure 40.51*. It comprises centralised control only, and is typical for the control centre of a utility. It shows distinct categories of functions that are characterised by: data acquisition (level A); dynamic control (e.g. frequency) (B); security assessment and optimisation (C); and adaptation (D). A common link is established by the databases (real-time, operation planning). Commands generated by the functions or originating from the human operator are executed via the telecontrol system).

40.11 System control in liberalised electricity markets

In most countries the electrical industry has been undergoing through drastic and dramatic changes. The institution of state-owned, protected, privileged infrastructure electrical industry has been abolished in almost all the industrialised countries. The electrical utilities have been structured in separate generation, transmission and distribution segments bringing in a transparency in the utility business. Depending on the market structure, the individual segments have also to compete in liberalised markets. The third party access gives equal rights of transmission system to each market participant. The control of transmission network becomes more complex due to third party access. In most of the countries the transmission system control is entrusted to a transmission system operator (TSO) or independent system operator (ISO). TSO/ISO may or may not own the transmission network but are responsible for system security and reliability. In addition to the system integrity the TSO has the main activity of enabling power transfer among various participants in the liberalised electricity market.

Such an operational scenario treats all the other services as auxiliary services. These services are to be organised, supervised and provided by TSO/ISO. The TSO has to control the frequency, voltage, stability and loading of the transmission network. Under certain circumstances the TSO might have the job of system restoration or start-up. In order to fulfil these responsibilities the TSO must procure services such as automatic generation control, governor control, reactive power, operating and standing reserves, black-start capabilities, emergency control actions, and adjustment of inadvertent tie-line exchanges. In some markets the TSO may be responsible for compensation of grid losses and providing means to solve the transmission constraints.

In a vertically integrated electrical utility the cost of generation, transmission and distribution was lumped together and passed on to end-user. In the liberalised energy markets each of these segments must have their own transparent cost structure. Depending on the market model in the country the transmission tariffs may vary. The transmission costing may vary from a simple postage stamp type model over MW-Mile method to full fledged load-flow based models. In most of the cases the TSO is responsible for setting up a fair and transparent tariff structure.

The congestion management, in classical vertically integrated electrical utility, was done by doing adjustment in generation or load. The increase of costs was absorbed in the overall cost of the operations. In the open access and liberalised market such a method is no more acceptable. There are organisational and analytical factors associated with the congestion management problem. The transmission constraints or congestion management problem becomes more complex when the wheeling of power is over the national boundaries. In such a case several TSOs in different countries may have to co-ordinate the transactions. *Figure. 40.52* shows interaction of TSO applications in a liberalised energy market.

40.12 Distribution automation and demand side management

The field of distribution systems, where the customer base is very heterogeneous and scattered over a wide area, open competition is forcing new ways to search for improved

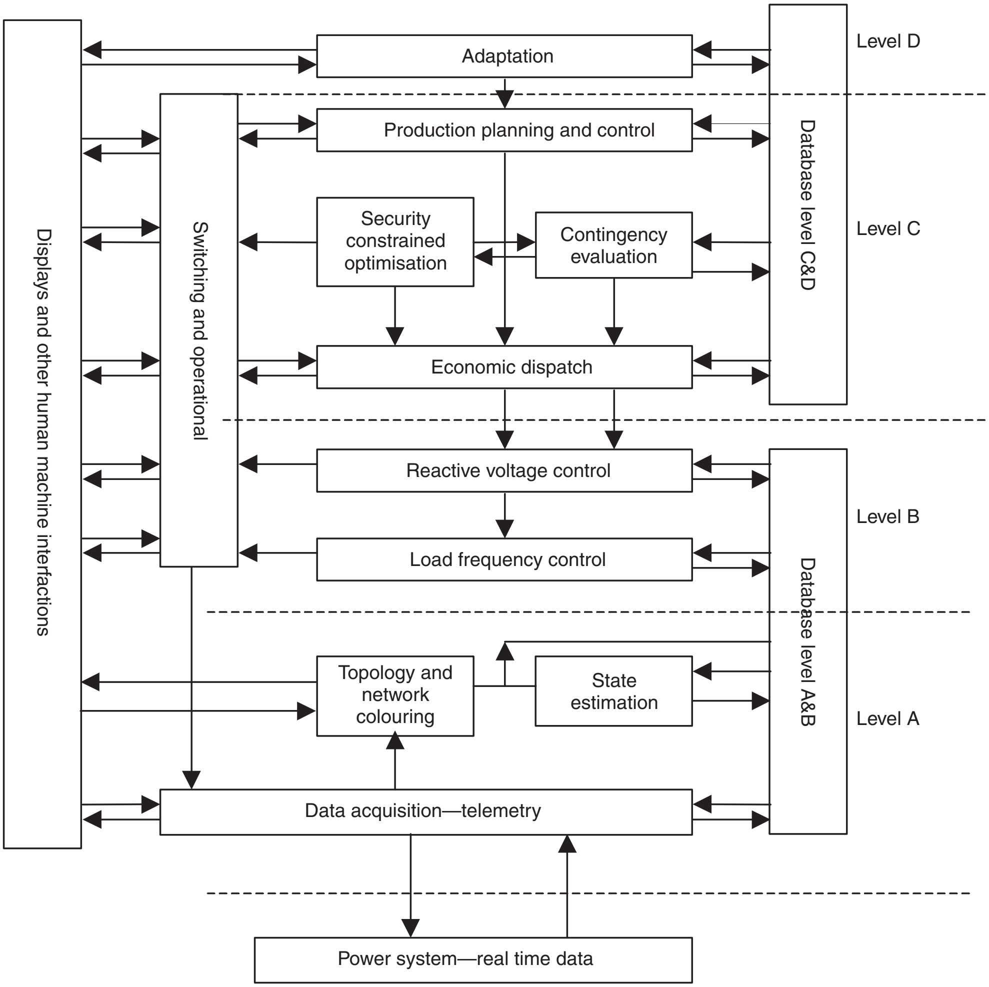

Displays and other human machine interfactions
Switching and operational
Adaptation
Production planning and control
Security constrained optimisation
Contingency evaluation
Economic dispatch
Reactive voltage control
Load frequency control
Topology and network colouring
State estimation
Data acquisition—telemetry
Power system—real time data
Database level C&D
Database level A&B
Level D
Level C
Level B
Level A

Figure 40.51 Framework of functions: software structure for a control centre

efficiency and better customer service, while rethinking the network operation. Distribution automation and demand side management (DA/DSM) provide solution to the efficient, reliable operation of the distribution utilities.

Distribution automation and demand side management has to be seen as a part of a holistic picture where the solution starts from network planning/refurbishment. The medium voltage breakers, substations, switchgears and along with the feeder automation, SCADA, automatic meter reading systems and load management systems are the other system components in the solution scenario. These are communicating with each other via communication systems specifically designed with requirements of distribution systems.

A distribution utility faces various challenges in this new liberalised energy market: The end-customer is free to choose its supplier in a non-franchised market. A combination of satisfactory services and competitive pricing is a must to survive. In open competition, efficiency and cost savings are critical for survival. Only well managed companies will have a chance in market driven scenario. Mandatory requirements imposed by the market regulator on energy distributor may make it obligatory to monitor power and prove the quality of the electrical energy delivered. The political guidelines of the region or country for energy delivery must be adhered to.

For distribution systems planning and engineering are performed to optimise the refurbishment of the network or extension of the medium voltage networks. These studies and consulting services are necessary to optimise the investment, which is required for these activities without deterioration in the customer service. Substations, transformers and power lines to reach a desired level of power quality must be designed to make the best use of the existing resources. An integrated planning, design, installation, commissioning and maintenance of these vital components in a medium voltage system must be undertaken to get full benefit of DA-DSM activities. Distribution switchgears are to be designed for integrated communication and control and fulfil the needs of flexible control and operation.

Automatic meter reading (AMR) provides systems to collect meter values read over different communication medium and different generations of meters. AMR improves flexibility in meter reading periods and selective services to the end customer. Load profile provided by the AMR systems helps to plan energy consumption and distribution. Feeder

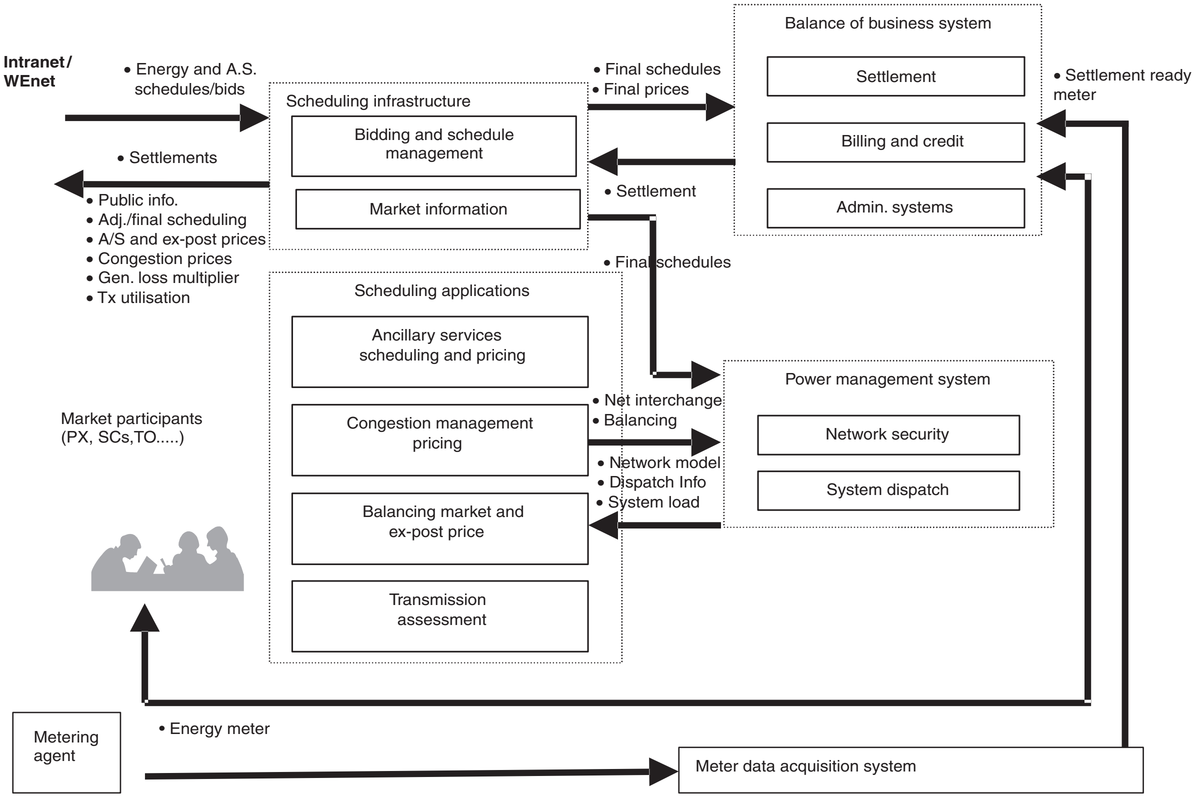

Intranet/
WEnet
• Energy and A.S. schedules/bids
• Settlements
• Public info.
• Adj./final scheduling
• A/S and ex-post prices
• Congestion prices
• Gen. loss multiplier
• Tx utilisation
Scheduling infrastructure
Bidding and schedule management
Market information
• Final schedules
• Final prices
• Settlement
• Final schedules
Balance of business system
Settlement
Billing and credit
Admin. systems
• Settlement ready meter
Scheduling applications
Ancillary services scheduling and pricing
Congestion management pricing
Balancing market and ex-post price
Transmission assessment
• Net interchange
• Balancing
• Network model
• Dispatch Info
• System load
Power management system
Network security
System dispatch
Market participants
(PX, SCs,TO.....)
Metering agent
• Energy meter
Meter data acquisition system

Figure 40.52 TSO applications in liberalised energy markets

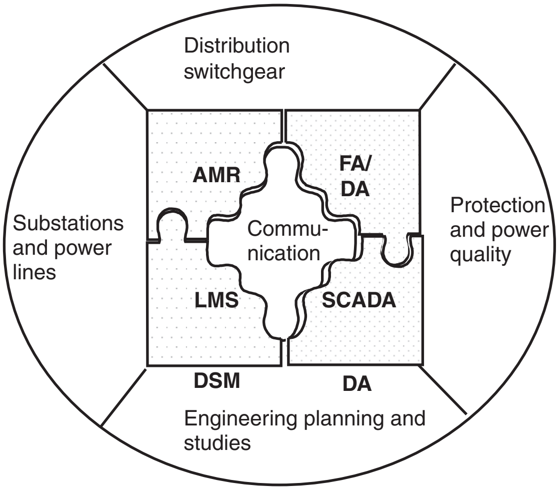

Distribution switchgear
Substations and power lines
Protection and power quality
AMR
FA/ DA
Commu- nication
LMS
SCADA
DSM
DA
Engineering planning and studies

Figure 40.53 DA/DSM for distribution utilities

automation/distribution automation gives the possibility to integrate local automation with the co-ordinated customer services. The fault detection, location, isolation and restoration keeps the customer minutes lost to minimum. Automation further reduces the cost of operation by reducing the manpower required for system operation.

Load management systems actively support in the reduction of peak demand and improving the utilisation of available assets. Classical ripple control systems are now augmented by two-way demand side management systems to tailor each customer load as and when required. Supervisory control and data acquisition (SCADA) as used in the transmission system control centres must be extended to monitor and supervise distribution network. SCADA for distribution system is typically based on fewer measured values and gives possibility to keep an overview of the system in order to achieve a high level of reliability and safety. The advanced distribution management applications, on this platform empower distribution utilities to simulate, plan and optimise their operation.

Power quality is a major issue for a significant number of end-customers who need a very high level of availability and distortion free voltage supply. Various solutions based on powerful static components, advanced relaying and reclosure schemes, which help in maintaining a purer waveform by reducing dips and flicker and interruptible power supply.

Communication systems and gateways for distribution system are crucial for any DA/DSM system. The DA/DSM systems should have possibility to mix and match different physical mediums for data communication up to the customer premises. Depending on topography and performance requirement communication systems can be designed using heterogeneous media mixing DLC, radio, microwave, GSM, PSTN, PLC giving an optimum cost effective coverage.

The individual solutions for the distribution utilities must consider the requirements from each of these components to achieve the best results for distribution systems in an open energy market.

40.13 Reliability considerations for system control

40.13.1 Introduction

The overall objective of power system operation, namely the provision of power to the customer at all times (i.e. with a high availability), places certain obvious requirements on the reliability and performance of power system components and their controls. Accidental events in the outside world (e.g. the atmosphere) that influence the system, failures of system components and imperfections of the control system including the operator can all affect the performance of the overall system and the final objective, and must therefore be taken account of. The positive or negative contributions of the generation, transmission and distribution systems should also be distinguished. For these subsystems, statistics on the failure rates, unavailabilities, etc., are given, and the question arises as to what degree system control in the widest sense can improve the performance of the system, i.e. the availability of power and energy.

In a more detailed analysis of an overall system with the objective as outlined above, a closer look has to be taken at the following:

(1) the meaning of availability in a power system, particularly at the consumer's end;
(2) the availability of components and subsystems;
(3) the specification of reliability characteristics for the control functions;
(4) the method of reliability analysis and of availability evaluation; and
(5) availability optimisation considering certain cost constraints.

It should be implicitly understood that improvements in system performance can be achieved via both the heavy equipment and the control system. Since certain similarities exist in the procedures for both domains, the emphasis here is placed on the control side, where hardware and software have to be considered. In addition to hardware considerations, which are very often put in the foreground, we consider the treatment of data and manual interventions.

The main line of the given approach is the establishment of a functional relationship between the supply of power at the consumer's end and the characteristics of the control functions. Bearing this in mind, one can evaluate controls at all levels from protection to the system control centre.

40.13.2 Availability and reliability in the power system

Before working out various details, we discuss the meaning of the terms 'availability' and 'reliability' as applied to a power system. How these concepts are understood depends largely on the object, the location, etc., to which they are applied.

Here three points of view are considered; namely those of (i) the consumer, (ii) the utility, and (iii) the system specialist. Quite different considerations and requirements apply to the three domains.

40.13.2.1 The consumer

At the consumer's end, i.e. at a single load point or at a supply point to a l.v. system, an availability in terms of up times and down times can be defined:

$$A = \frac{\mathrm{MUT}}{\mathrm{MUT} + \mathrm{MDT}} = 1 - \frac{\mathrm{MDT}}{\mathrm{MUT} + \mathrm{MDT}} = 1 - \bar{A}$$

where MUT is the mean up-time, MDT is the mean down-time, A is the availability, and $\bar{A}$ is the unavailability. A represents an average fraction (per unit) of the time during which power and energy could have been delivered to the load point. In developed systems this figure reaches values of up to 0.9995–0.9997, so that the average down-time per year is as low as 2–4 h. This applies to single supply points and customers. The figure may vary from point to point and cannot be transferred to a complete voltage level or to the transmission system.

40.13.2.2 The utility

Single outages at supply points and unavailabilities at the consumer's end are undesirable and should be kept to a minimum that is determined by economic considerations. Technically, however, the disconnection of single consumers has no detrimental effects on the overall system. There is simply a certain reduction of the load, which is balanced by the various control mechanisms. The utility must be interested to reduce these outages in accordance with its legal obligations, but its real concern lies in the continuous operation of the transmission and generating system. There, any disturbance that might endanger the overall system is to be avoided. There are concepts that measure the amount of non-served energy accumulated over a given period, e.g. 1 year, in terms of the maximum load multiplied by 'system minutes', where the system minutes constitute a measure of the unavailability of the system. However, this figure in minutes is of secondary importance to the utility as long as no complete breakdowns of the transmission system occur. This consideration is further supported by the fact that many utilities employ load shedding in order to avoid emergency situations. Load shedding also causes outages, so that consumers are not served, but it is done for the benefit of the integrity of the overall system.

So, what finally counts for the utility is the availability of the transmission system measured in terms of up and down times. Many utilities have remarkable records, i.e. availabilities of 100% over tens of years. However, there have also been catastrophic black-outs lasting for hours in utilities all over the world; these have received considerable attention and have motivated significant research efforts.

40.13.2.3 The control specialist

Here, we consider the reliability aspects of control levels in a power system from a technical point of view. These aspects include protection and decentralised and centralised control. Clearly, improving the availability of control functions will improve the availability of the supply of power and of the system itself. However, the control specialist differentiates between flat improvement of a characteristic of control functions and augmentation of a parameter which might have significant effects on the system. Detailed investigation reveals that the power system is quite tolerant, as it can maintain its function in the absence of certain control functions, at least over a certain period. Hence, the control

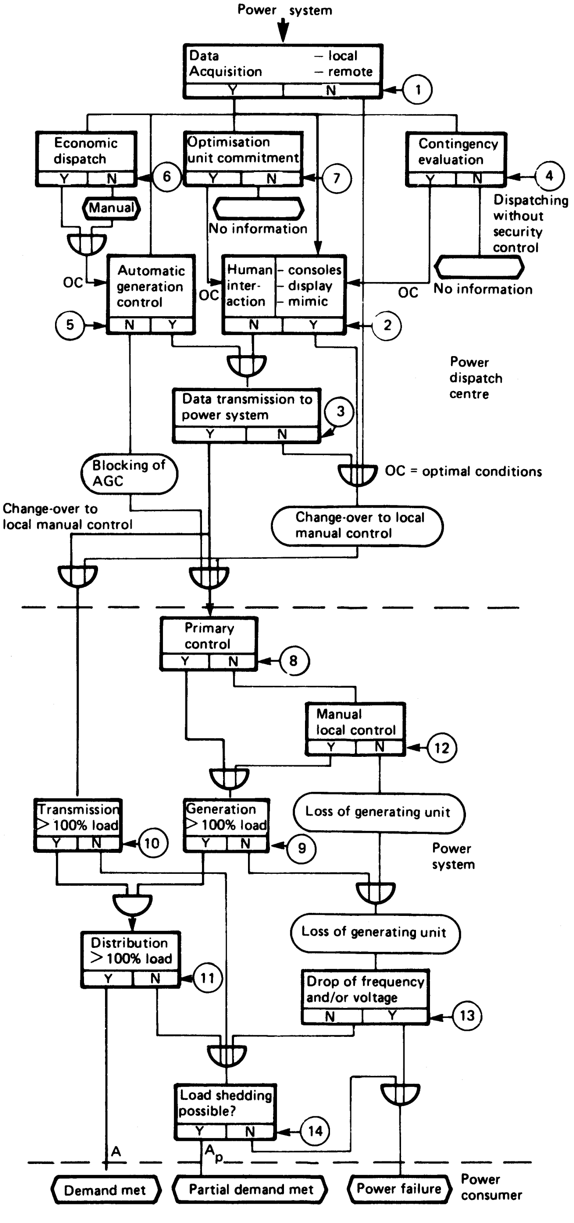

Power system
Data Acquisition – local – remote
Y N
1
Economic dispatch
Y N
6
Optimisation unit commitment
Y N
7
Contingency evaluation
Y N
4
Manual
No information
Dispatching without security control
OC
Automatic generation control
N Y
5
OC
Human inter-action
– consoles – display – mimic
N Y
2
OC
No information
Power dispatch centre
Data transmission to power system
Y N
3
Blocking of AGC
OC = optimal conditions
Change-over to local manual control
Change-over to local manual control
Primary control
Y N
8
Manual local control
Y N
12
Transmission > 100% load
Y N
10
Generation > 100% load
Y N
9
Loss of generating unit
Power system
Loss of generating unit
Distribution > 100% load
Y N
11
Drop of frequency and/or voltage
N Y
13
Load shedding possible?
Y N
14
A
A_p
Demand met
Partial demand met
Power failure
Power consumer

Figure 40.54 A CCC for the power system, its primary control and its system control centre

specialist is interested in identifying those elements and controls that contribute most significantly to the goal of system availability. There are, however, other elements which tolerate a lower availability or a reduced performance, and the control specialist must certainly take a broad view, i.e. he must also consider events and failures with low probability. In the end, it is the concerted effort of all types of control over a long period which constitutes success, i.e. the high performance of the power system under acceptable economic conditions.

To illustrate the interaction between disturbances and control actions on various levels, a cause-and-consequence chart (CCC) for a power system with its primary controls and its system control centre is shown in *Figure 40.54*, and explained in *Table 40.1*. The CCC is the basis for any reliability analysis of a controlled power system. It reveals the causal relations between events and faults on the one hand and effects on the system on the other.

Table 40.1 Details of the CCC in *Figure 40.54*

Circle	*Cause*	*Consequences*
1	Loss of transducers, telemetry system, front-end equipment	Loss of information on system state
2	Failure of computer or peripherals	Loss of information on status of substations, line loading, alarms
3	As for 2	No updating of set-points in power-stations
4	As for 2, loss of application program	No contingency evaluation, no security check
5	As for 4	No automatic generation control
6	As for 4	No economic dispatch control
7	As for 4	No operating planning unit commitment
8	Loss of auxiliary equipment in speed governor	No telemetered change in set-point
9	Loss of generation	Not sufficient reserve
10	Loss of transmission	Not sufficient transmission capacity
11	Loss of distribution	Not sufficient distribution capacity
12	Loss of generation	Reduction of reserve
13	Load demand not met	Load shedding
14	Load shedding not possible	Power failure

40.13.3 System security

The concept of system security is often discussed together with reliability considerations. It is primarily a deterministic concept which gives an answer as to whether the system can survive a given set of contingencies. It uses the model of system states as explained in Section 40.2 and load-flow techniques to check contingencies. In the basic concept, nothing is said about the duration of the various states (probabilities). However, the concept is amenable to extension to include this, and various approaches to the reliability analysis of the transmission and generation system have followed this direction.

Basically, a secure system is understood to be one that can withstand a number of outages, mostly single outages. This leads to the idea of $n-1$ security, in which, out of *n* components, the disconnection of any component alone does not endanger the operation of the system

40.13.4 Functions

To assess the various control functions and their contributions to the availability of the power system, it is necessary to know their structure and framework as given in *Figure 40.51*, though with much additional detail and an allocation to the various hardware components. On the basis of such a structure and its functional relations, the possible contribution of the control system could be assessed by a simulation. In principle the following relations must hold.

The contribution of a *perfect* control system to the availability A is given by

$$A = A' + \Delta A_0$$

where A' is the availability of the uncontrolled power system and ΔA_0 the possible contribution of the control system. A *real* control system, however, has itself got a finite availability A_c; hence

$$A = A' + A_c \Delta A_0$$

The availability A_c of the control system can be derived from the various functions f_j and their individual functional availabilities A_{cj} under the assumption that certain stationary probabilities g_j are known which give a rate at which the function f_j are called upon. The probabilities g_j must add up to unity:

$$\sum_j g_j = 1$$

then

$$A_c = \sum_j g_j A_{cj}$$

A_{cj} is a functional availability that applies directly to a so-called 'real-time' (RT) function. Its contribution comes from its on-line operation. It is called upon whenever an event bearing some risk for the power system arises. In contrast, so-called 'preventive' (PR) functions condition the system for the event in advance. Their operation does not coincide with the appearance of the event, and the control function is not needed in the event. Thus, there are less stringent requirements for the actual availability A_{aj} of such a function. For a more detailed discussion of this subject, see references 1 and 2.

40.13.5 Impact of system control

At this point, the question arises as to the possible contribution of system control to the availability of the power system. The answer would be of great value for the design of control systems, telecontrol equipment, regulators and control centres. However, a complete assessment for a real system seems impossible. With the help of a mathematical model, though, a limited answer can be given, at least in terms of the relative merits of the various functions. The treatment of such a model requires a Monte-Carlo simulation of the system behaviour over sufficiently long time periods.

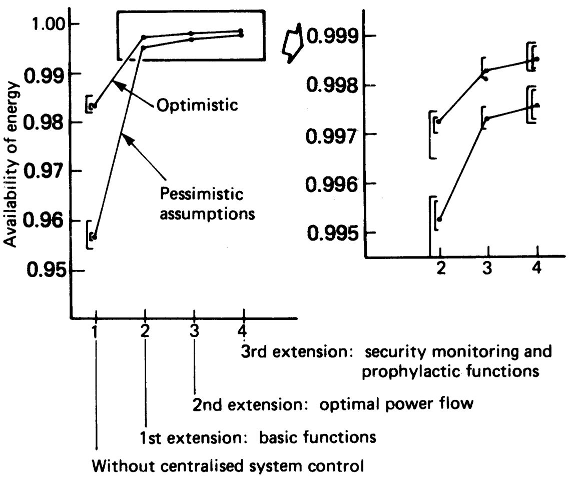


Figure 40.55 Effectiveness of different functions. +, Mean value (five simulation runs); [, range of results assuming small uncertainty within the failure data;], range of results assuming large uncertainty in the failure data

Such simulations prove that the performance of the power system can actually be improved, although the degree of improvement will depend upon the inherent performance, the loading, redundancy, etc. (for details see reference 3). It can also be shown that the distinction between RT functions and PR functions is well justified. It turns out that repair plays an important role for the PR functions, which is further supported by the fact that the requirements for control change with the daily load cycle. Stress situations appear two or three times per day. In between, the system can be conditioned for a possible event. A PR function need not be available at a specified time but can be delayed. Thus, repair is possible.

In order to give an idea of such a result, an example from reference 3 is presented in *Figure 40.55*. The figure and its enlarged section show the increase in the availability of energy as a function of the categories of control functions. Clearly, the most significant improvement can be achieved by the basic functions of system control like SCADA. Further improvements are harder to realise, but some gain is still possible. The remaining unavailability is due to the particular structure of the test system. It is an isolated system with a peak load of about 4000 MW. The system was assumed to be heavily loaded. Thus, control had a limited effect because of the lack of reserves in the system.

40.13.6 Conclusions

As far as the performance of a power system as expressed by its availability is concerned, three domains have to be considered: (i) the generation system and its reserves; (ii) the transmission system; and (iii) the control system. Assuming that the first two systems are fixed, it has been shown that system control will improve the performance. Quite detailed studies are necessary in order to evaluate the contributions of the various control functions. In such a treatment the final question concerns the configuration of the computer system.

Over the years one basic set-up has evolved which has not changed very much and seems to prove its validity as time goes on. It is the multi-computer concept on three levels having several front-end computers, a double system on

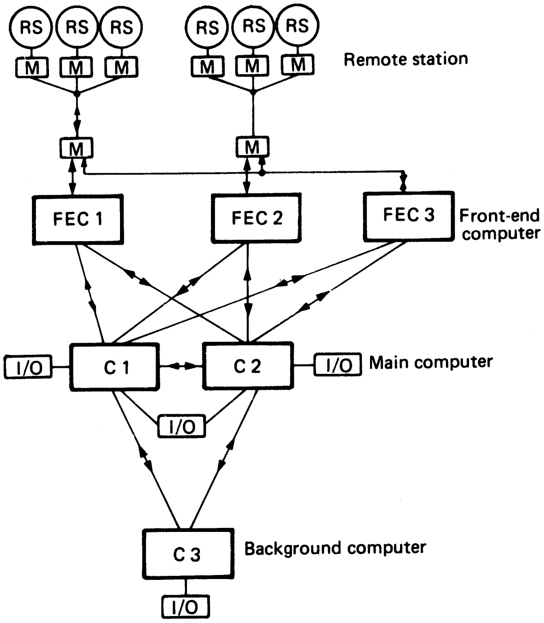


Figure 40.56 Hierarchical and redundant multi-computer system with three levels: I/O, input/output; FEC, front-end computer; C, main processor; M, modem; RS, remote station

the main level and a background computer. The configuration is shown in *Figure 40.56*.

References

1 FREY, H., GLAVITSCH, H. and WAHL, H., Availability of power as affected by the characteristics of the system control center, Part I: Specification and evaluation, *Proc. IFAC Symposium on Automatic Control and Protection of Electric Power Systems, Melbourne*, pp. 21–25 (February 1977)

2 FREY, H., GLAVITSCH, H. and WAHL, H., Availability of power as affected by the characteristics of the system control centre, Part II: Realization and conclusions, *Proc. IFAC Symposium on Automatic Control and Protection of Electric Power Systems, Melbourne*, pp. 21–25 (February 1977)

3 GLAVITSCH, H. and KAISER, W., Assessment of reliability parameters of the power system and their dependence on control functions, *CIGRE SC32 Study Committee Meeting, Rio de Janeiro, Paper No. 81 SC05*, pp. 21–24 (September 1981)

Bibliography

Introduction

DYLIACCO, T. E., The adaptive reliability control system, *IEEE Trans. PAS*, **86**(5) (1967)

GLAVITSCH, H., Power system control and protection—the interaction of new and existing concepts in providing economic and reliable supply, keynote address given to the *IFAC Symposium on Automatic Control and Protection of Electric Power Systems*, Melbourne, Australia, 21–25 (February 1977)

Objectives and requirements

Load Dispatching Systems (Special issue), *Brown Boveri Rev*., **66** (March 1979)

System description

KOLLER, H. and FRÜHAUF, K. PRIMO database management system, *Brown Boveri Rev*., **66**, 204 (March 1979)

Data acquisition and telemetering

FUNK G. and SODER, G., Indactic 13 and 33 Telecontrol Systems based on ED 1000 modules, *Brown Boveri Rev*., **61**, 393–398 (August 1974)

Decentralised control: excitation systems and control characteristics of synchronous machines

BONANOMI, P., GÜTH, G., BLASER, F. and GLAVITSCH, H., *Concept of a practical adaptive regulator for excitation control, Paper No. 79 453-2, IEEE Summer Power Meeting* (1979)

PENEDER, F., Modern excitation equipment for power station generators, *Brown Boveri Rev*., **67**, pp. 173–179 (March 1980)

PENEDER, F. and BERTSCHI, R., Slip stabilization equipment, transfer functions and the relevant experience, *Brown Boveri Rev*., **65**, 724–730 (November 1978)

PENEDER, F. and BERTSCHI R., Slip stabilization, *Brown Boveri Rev*., **61**, 448–454 (September/October 1974)

Decentralised control: electronic turbine controllers

COHN N., *Control of Generation and Power Flow on Interconnected Systems*, Wiley, New York (1966)

GLAVITSCH, H. and STOFFEL J., Automatic generation control, *J. Elec. Power Energy Syst*., **21**, (1) (January 1980)

ANDRES, W., SCHAIBLE, W. and SCHATZMANN, G., Turbotrol 4 electronic control system for steam turbines based on PC 200 equipment, *Brown Boveri Rev*., **62**, 377 (September 1975)

MÜHLEMANN, M., The electronic water turbine controller Hydrotrol 4, *Brown Boveri Rev*., **66**, (October 1979)

Decentralised control: substation automation

LEUZINGER, J. and BAUMANN, R., BECOS 10—a software package for out-stations and small dispatching centres, *Brown Boveri Rev*., **66**, 175–180 (March 1979)

JERABEK, A. and RISCHEL H., 'BECOS 20—a software system for regional dispatching centres, *Brown Boveri Rev*., **66**, 181–187 (March 1979)

Centralised control

REICHERT, K., Application software for power system operation, *Brown Boveri Rev*., **66**, 197–203 (March 1979)

FROST, R., HUYNEN, M. and STAHL, U., BECOS 30 software system for large dispatching centres, *Brown Boveri Rev*., **66**, 188–196 (March 1979)

REICHERT K., Systems engineering for power system operation, *Brown Boveri Rev*., **66**, 225–233 (March 1979)

SCHAFFER, G., Power application software for the operation of power supply systems, *Brown Boveri Rev*., **70**, 28–35 (January/February 1983)

System operation

MILLER, R., *Power System Operation*, McGraw-Hill, New York (1970)

41

Reactive Power Plant and FACTS Controllers

A Gavrilović OBE
Formerly of ALSTOM T&D Ltd, Power Electronic Systems

W P Williams BSc, CEng, FIEE, SMIEE
Formerly of ALSTOM T&D Ltd, Power Electronic Systems

H L Thanawala PhD, BE(Elect), BE(Mech)
Formerly of ALSTOM T&D Ltd, Power Electronic Systems

D J Young BA
ALSTOM T&D Ltd, Power Electronic Systems

Contents

41.1 Introduction 41/3
41.2 Basic concepts 41/3
  41.2.1 Characteristics of system components 41/3

41.3 Variations of voltage with load 41/5
  41.3.1 The flow of power and vars 41/6
  41.3.2 Voltage instability 41/6
  41.3.3 Var balancing for steady state conditions 41/6
  41.3.4 Power transmission over long distances 41/6
  41.3.5 Transmission characteristics 41/6
  41.3.6 Transient instability 41/9
  41.3.7 Var balancing for dynamic conditions 41/9

41.4 The management of vars 41/10
  41.4.1 Objectives 41/10
  41.4.2 Tools for var management 41/11
  41.4.3 Var management by the supply authorities 41/11
  41.4.4 Var management by consumers 41/11

41.5 The development of FACTS controllers 41/11
  41.5.1 Synchronous compensators 41/12
  41.5.2 Sign convention for vars and reactive current 41/12
  41.5.3 Basic features of static compensation 41/13
  41.5.4 Harmonic-compensated self-saturated reactor 41/13
  41.5.5 Thyristor switches 41/14
  41.5.6 Convertor-based FACTS controllers 41/15

41.6 Shunt compensation 41/19
  41.6.1 General 41/19
  41.6.2 Mechanically switched reactors and capacitors 41/21
  41.6.3 Synchronous compensators 41/24
  41.6.4 Static var compensators 41/25
  41.6.5 STATCOM 41/26
  41.6.6 STATCOM applications 41/29

41.7 Series compensation 41/30
  41.7.1 Series capacitor compensation 41/30
  41.7.2 Controllable series compensation 41/31
  41.7.3 Buffering of loads from system disturbances 41/33

41.8 Controllers with shunt and series components 41/34
  41.8.1 Quadrature booster transformer (QBT) 41/34
  41.8.2 Unified power flow controller (UPFC) 41/35

41.9 Special aspects of var compensation 41/35
  41.9.1 Lamp flicker compensation 41/36
  41.9.2 Phase balancing 41/37
  41.9.3 Switched resistors 41/38
  41.9.4 Harmonic currents and harmonic filter design 41/38

41.10 Future prospects 41/39
  41.10.1 Recent progress 41/39
  41.10.2 Further advances 41/39
  41.10.3 Future applications of FACTS controllers 41/39

41.1 Introduction

When designing a transmission or distribution system or studying its operation, the engineer must take into account not only the power requirement of the loads, but also the fact that they consume reactive power and, equally important, that the networks include inductive and capacitive elements which themselves absorb or generate reactive power. In contrast to power, which is generated and consumed in a controlled manner only at specific points in the networks (ignoring losses), reactive power is generated and absorbed throughout the network in significant quantities which vary with the system loading and configuration.

Growth in the demand for energy, the development of competition for the production and supply of this energy, together with increasing opposition to the construction of new overhead power lines, make it increasingly important to improve the utilisation of existing assets. However, there are constraints on the operation of power systems that often result in under-utilisation of their assets.

For a power system to operate efficiently and securely, the importance of the correct and co-ordinated provision and control of reactive power cannot be overemphasised. It is necessary to examine reactive power requirements under both steady-state and dynamic conditions. Although it has been normal in the past to consider these requirements separately, it is preferable that they should be dealt with in a well-co-ordinated way.

'FACTS' is an acronym for 'Flexible AC Transmission Systems', which is defined as 'alternating current transmission systems incorporating power electronic-based and other static controllers to enhance controllability and increase power transfer capability'. Many of the reactive power plant items described in this chapter are covered by this definition. FACTS equipment that provides control of one or more of the parameters of an electrical network is often referred to as a FACTS controller.

The International Electrotechnical Commission, IEC, has defined the unit for reactive power as the 'var' (volt-ampere reactive) and uses the name 'vars' for reactive power. For brevity, the IEC terminology will generally be used within this chapter.

41.2 Basic concepts

In electrical power systems, electrical energy is produced by generators, transferred by means of transformers into a transmission system by which it is conveyed to distribution systems and supplied, again via transformers, to the users of the energy. *Figure 41.1* shows a system where different components and reactive power devices have been combined to illustrate the main features of var supply and demand in such a power system. The characteristics of individual components and the methods of network analysis are covered in other chapters. The following sections outline the characteristics and parameters relevant to the management and supply of vars, network voltage control and reactive compensation.

41.2.1 Characteristics of system components

41.2.1.1 Generators

The main generator parameters relevant to the var balance in a system are the transient reactance and the short-circuit

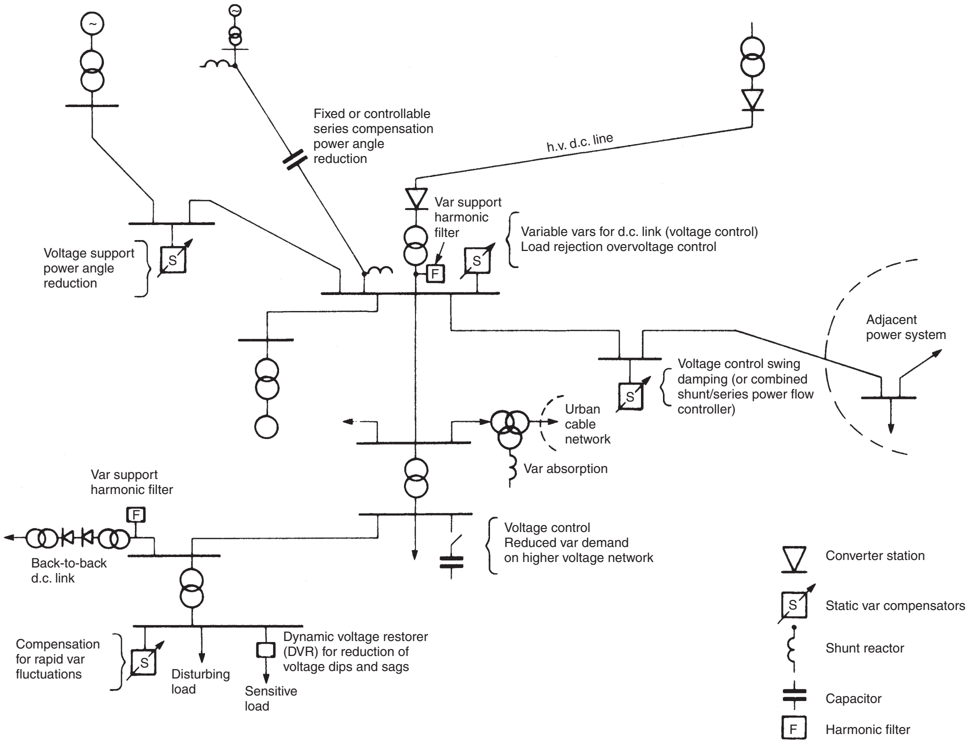

Fixed or controllable
series compensation
power angle
reduction
h.v. d.c. line
Var support
harmonic
filter
Variable vars for d.c. link (voltage control)
Load rejection overvoltage control
Voltage support
power angle
reduction
S
F
S
Adjacent
power system
Voltage control swing
damping (or combined
shunt/series power flow
controller)
S
Urban
cable
network
Var absorption
Var support
harmonic filter
F
Back-to-back
d.c. link
Voltage control
Reduced var demand
on higher voltage network
Compensation
for rapid var
fluctuations
S
Disturbing
load
Dynamic voltage restorer
(DVR) for reduction of
voltage dips and sags
Sensitive
load
Converter station
S
Static var compensators
Shunt reactor
Capacitor
F
Harmonic filter

Figure 41.1 Applications of FACTS controllers in a flexible transmission system

ratio (SCR, which is approximately the reciprocal of the synchronous reactance). Where transient stability is to be examined, or where var control is required to be used for the damping of system oscillations, the inertia of the turbine–generator combination also needs to be known.

To obtain economies from standardisation, many manufacturers of turbogenerators will offer a machine from a standard range of frame sizes; different combinations of megawatts, rated power factor and short-circuit ratio are possible with each design of the range. The characteristics and capabilities of each generator must be allowed for in considering the overall var balance for a particular system.

The frame size of a generator depends upon its megavolt-ampere (MVA) rating; hence, the cheapest design for a given megawatt output would be one rated to operate at unity power factor (pf). However, unity pf operation may not be acceptable for stability and optimum operation of the network. Operation of a generator at a lagging power factor, i.e. when generating vars, requires a higher level of field excitation than at unity pf; this reduces the generator load angle for a given power output. To prevent a generator losing synchronism during angular swings following a system disturbance, it is desirable to operate it in steady state with an adequate margin of load angle from its stability limit. A typical cylindrical-rotor turbogenerator will have an SCR in the range 0.45–0.6; this would require operation at around 0.9–0.95 lagging pf, as shown in the capability chart in *Figure 41.2*, if the generator is feeding into a reasonably strong system. Under these conditions the generator will be supplying a significant amount of vars to the system, amounting to 0.48–0.33 Mvar/MW of output.

When it is necessary for a generator to absorb vars (i.e. to operate at a leading pf), as occurs during light load conditions for generating stations supplying remote load centres through long high-voltage lines, the field excitation must be lower than that at unity pf. The generator must then be designed to maintain stability, which requires that positive excitation is needed under all conditions. This gives a design with an SCR of, say, 1.0–1.5. Such high values are common in salient-pole low-speed generators in hydroelectric stations, which account for most such applications. Such a high value is unusual in round-rotor turbogenerators and would require an unusually large machine for the MW output. It may be more appropriate to provide var absorption equipment within the transmission network in these circumstances.

The proportion of the system var requirements supplied by generators varies from system to system. In many systems, the high-voltage transmission network has a surplus of vars, i.e. its shunt capacitance generates more vars than are absorbed by the I^2X losses in its series reactance—even under maximum-demand conditions. This allows the generators to operate at high power factors, even if the system loads themselves have low power factors. The operating pf may, in practice, be dictated more by stability than load requirements and it is quite usual for generators to be operated at a lagging pf higher than their rated value.

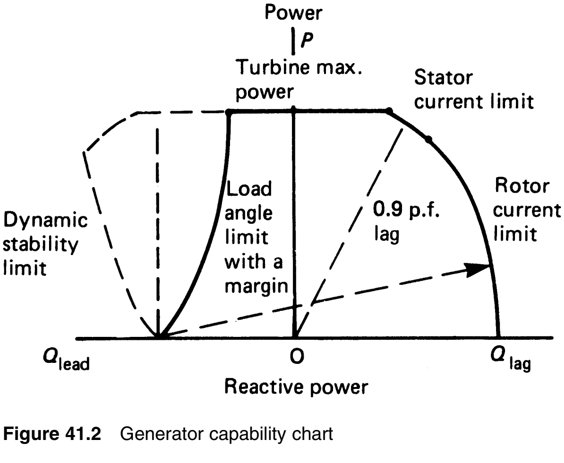


Figure 41.2 Generator capability chart

41.2.1.2 *Transformers*

Although generators are now becoming available which are suitable for connection directly to system voltages of 145 kV and higher, in most cases it is necessary to couple the generator to the transmission system by means of a transformer. Similarly, transformers are used to reduce the voltage level at the points of connection of the distribution companies and again for the supply of individual customers or groups of customers. They are also used between the different voltage levels used by the transmission company, or companies.

The transformer leakage reactances have great significance for the operation of the transmission system. Although the power loss and magnetising current of a transformer can be neglected when considering var flows, the var absorption in its leakage reactance (given by I^2X) is important. During temporary network overvoltages, transformer cores may become saturated, resulting in abnormally high magnetising currents. Although the increased var consumption will assist in reducing the overvoltages, harmonic currents will be generated. In flowing through the network, these currents may excite undesired resonances which themselves can cause overvoltages.

On-load tap-changing of transformers is a useful function in the co-ordination of var control between various voltage levels within a system and for balancing var flows within a single voltage level. Two distinct practices occur with transformers used to interconnect different transmission voltage levels in a single system; a number of authorities, with a view to using transformers having the highest reliability, install fixed ratio transformers, whereas others use on-load tap-changers having a typical total range of 20–30%. For the interconnection of transmission networks to sub-transmission or distribution networks, the use of tap-changing transformers is almost universal.

41.2.1.3 *Transmission lines and cables*

Transmission lines and cables absorb vars in their series inductance. They also have an inherent capability to generate vars by their shunt capacitance, which causes a reactive ‘charging current’ to flow into the line. Both the series inductance and the shunt capacitance are distributed along the length of the line.

From the analytical point of view, cables are indistinguishable from overhead lines except that the ratio between shunt capacitance and series reactance is considerably higher. The use of cables in high-voltage networks is increasing because of the difficulties of building new overhead lines for electrical power transmission into urban areas and, to a more limited extent, for a.c. underwater links; the use of cables is still very small compared with overhead lines because of the cost disadvantages. Where cables are used in any quantity, their high shunt capacitance reduces the need for additional var generation at peak demand periods but increases the need for var absorption during low loading conditions.

When there is no load current flowing through a line or cable, it is found that the voltage at the receiving end of the line is higher than that at the sending end; this is caused by the flow of the capacitive charging current through the series inductance and is known as the Ferranti Effect. The magnitude of the voltage rise increases very rapidly with increasing length of the line, from only about 3% for 200 km to about 50% for 800 km. The magnitude of such a voltage rise within a transmission system must be limited (generally to either 5 or 10% above rated voltage), in order to avoid exceeding the safe limits of operation for the insulation of the line or cable itself as well as the safe limits for the various equipment connected to the transmission system. Very long lines are normally split into sections 200 to 300 km long; voltage or var control measures can then be applied at the intervening substations, if required.

As the power flowing through the line is increased, the var absorption in the series reactance also increases. There is a critical current at which the magnitude of 'series vars' absorbed in the line is just enough to balance the 'shunt vars' generated by the shunt capacitance of the line. With this level of power flow, the line voltage will have the same value all along the line (neglecting resistive losses). When there is a further increase in power flowing through the line, the vars absorbed will outweigh the vars generated and the voltage at the receiving end of the line will start to fall very rapidly with increasing load and may reach the point of complete collapse.

The surge impedance of a transmission line having series inductance L and shunt capacitance C, is a resistance equal to $\sqrt{(L/C)}$. If a load that has this value of resistance is connected to the end of the line, the power that flows into the load is called the surge-impedance load or SIL. This is the load, mentioned above, at which the var absorption in the series reactance equals the var generation in the shunt capacitance, and for which the voltage along the line is constant. For a given surge impedance, the SIL increases in proportion to the square of the line voltage.

The value of surge impedance for a 132 kV line is about 400 Ω. For higher voltage systems, multiple conductors are normally used, in part to reduce corona effects, and the series inductance is also somewhat lower, giving a surge impedance of about 250 to 300 Ω for a 400 kV line. One effect of choosing an increased transmission voltage is to reduce transmission losses at a given power flow, but another effect is to enable an improved power transmission capability per right of way. In selecting a transmission voltage, therefore, a balance must be struck between equipment costs, operating costs and future capacity.

The SIL of typical cables usually exceeds their rating, so that cables generate more vars than they absorb and compensation is often required. This imposes a limit to the uninterrupted length of cable that can be used; for example, for submarine cables the limit is typically 50–75 km before the capacitive charging current reaches the current rating of the cable.

Cables that use cross-linked polyethylene (XLPE) insulation are now available for transmission voltages of 400–500 kV; for a 400 kV, 1100 MVA XLPE cable design, the SIL is about 2500 MW, which is still much higher than the rating. Nevertheless, these cables generate rather less charging current than the earlier types and allow uninterrupted lengths of up to about 100 km. Where longer distances are required, another solution becomes necessary, for example, h.v.d.c. or a gas-insulated transmission line (GIL).

Most countries have two or three transmission voltage levels within their systems, the highest normally being in the 400–500 kV range, with a few countries in North and South America and the USSR operating or installing networks in the 750 kV range. Although the use of u.h.v at 1000–1500 kV is technically feasible, the expense is high and its commercial application seems likely to be limited for the foreseeable future.

41.2.1.4 Loads

In general, loads are inductive and their power factor can vary widely, dependent on *type*, from industrial through commercial to domestic loads, and on *location* (for example, commercial and domestic loads in hot climates have an increasing proportion of motors for air conditioning). For most studies, load representation as constant impedances (or constant P and Q) is adequate but, when studying dynamic conditions involving wide variations in voltage, a more detailed representation that allows for changing power factors will be required. Power-factor-correction (pfc) capacitor banks located in load and distribution networks should also be taken into account in any studies.

Air-conditioning and many industrial loads use induction motors; when these comprise a substantial proportion of the total load on a network they display a typical voltage stability challenge. A short interruption due to a fault causes each motor to slow down. When the fault is removed and the voltage re-established the motor tries to regain full speed but, because of the increased slip, there is an increased var demand. The effect of many such loads is to depress the system voltage so much that recovery from a fault is slow and, in extreme cases, voltage collapse stability may be threatened and all load motors may stall.

Loads that can produce network disturbances or distortions require special attention. These are, in the main, found in the metal and mining industries and in a.c. traction systems; the large and rapid fluctuations in the load currents of arc furnaces and large thyristor drives, particularly their fluctuating var demands, can cause annoyance to other customers. These loads often also cause problems due to the generation of harmonics.

41.3 Variations of voltage with load

For each part of a power system it is necessary to define a nominal operating voltage and also a maximum operating voltage, which is usually 5% or 10% higher than the nominal voltage; these voltages are used in establishing the rated voltages of circuits and components. The rated MVA capacity of each circuit or component is then determined on the basis of its rated current and rated voltage. When an item is required to operate at its rated capacity but at less than its rated voltage, the current will exceed its rated value; this overload condition may only be acceptable for limited periods when the ambient temperature is less than the defined maximum value.

It is therefore usual to try to operate a system in steady state so that the voltage is constrained, throughout the system, to be within a very few per cent above or below its nominal value. Wider deviations of voltage are inevitable during and following faults and other disturbances. The most severe transient and switching overvoltages are usually limited by means of surge arresters; slower, dynamic or long-term, overvoltages and undervoltages may require var compensation equipment to reduce their magnitudes and durations to acceptable levels.

41.3.1 The flow of power and vars

Electricity supply networks have predominantly inductive series impedances. *Figure 41.3*(*a*) shows an elementary circuit in which a load is supplied from a generator via inductance, X. The generator output voltage, V_S, is constant; the voltage at the load is V_R. When there is no load, the voltages V_S and V_R are equal in magnitude and phase. When a load is connected it will draw a current, I, through X.

Figure 41.3(*b*) shows the vector diagram for a resistive load; I is in phase with V_R and $\phi = 0$. If the value of the load resistance is R, the power is $P = V_R \cdot I = I^2 \cdot R = V_R^2/R$. The voltage drop through the reactance X, is $\Delta V = I \cdot X$ and is in quadrature with V_R. It causes a phase angle difference, δ, between V_S and V_R. Also, V_R is smaller than V_S because of the vars ($Q = I^2 X$) consumed in X.

Figure 41.3(*c*) shows the vector diagram for a purely inductive load, with I lagging by $\phi = 90°$ behind V_R. The voltage drop through X is in now in phase with V_R, which therefore remains in phase with V_S but is appreciably smaller. *Figure 41.3*(*d*) is the general case, with a lagging power factor load that causes differences both of phase angle and magnitude between V_S and V_R.

These examples illustrate that, when power flows through inductive impedance from one point to another, it is accompanied by a difference in the phase angles of the voltage at those points. Var flow between two points is accompanied by a difference in the voltage magnitudes at those points. It is important to bear in mind that these features may be expressed in the converse way. Thus, when a difference in phase angle is enforced between the voltage vectors at two points in a network, power will be transmitted from one point to the other; by enforcing a difference between the voltage magnitudes at the two points; var flow will take place. Put simply, differences in phase angle control the flow of power and differences in voltage magnitude control the flow of vars. This is a highly simplified, but crucial, description of the principles of a.c. power transmission; reference should be made to Chapters 3 and 35 for the theoretical background.

41.3.2 Voltage instability

Figure 41.3(*e*) plots the relationships between power, vars and current supplied by the generator, together with the load voltage, V_R for the purely resistive load illustrated by *Figure 41.3*(*b*). As the magnitude of load resistance, R, is reduced, the current, power and vars each increase at different rates, but V_R decreases. When R is equal to X, $P = Q$, $V_R = V_S/\sqrt{2}$ and $\delta = 45°$. For this load resistance, the power has its maximum value, given by $P_{max} = V_R^2/R = V_S^2/2X$. Any further reduction of resistance will cause a further reduction of V_R and also of P whereas Q continues to increase. *Figure 41.3*(*f*) plots the relationship between voltage V_R and power P and shows how important it is, for the stable operation of a system, that voltages at load points should not be allowed to drop below certain limits. If the load voltage falls too far, the condition can become progressively worse, leading to a complete collapse of voltage. Such voltage instability is more prone to occur in systems with large inductive impedances and can be exacerbated by loads that tend to consume constant power and vars irrespective of the magnitude of their supply voltage. Loads of this kind include those that are supplied by transformers with on-load tap-changers, when the tap-changers have an automatic control which attempts to maintain a constant secondary voltage.

Figure 41.3(*g*) shows a family of voltage-load curves for different power factor loads, including the unity pf case of *Figure 41.3*(*f*). When the load current has a lagging component, i.e. is at less than unity pf, the maximum power capability will be smaller and the voltage for a given load will be reduced compared with *Figure 41.3*(*f*). For a particular power factor curve and for any power less than the maximum there are two operating points; points A, A_1, etc., are stable (i.e. dV_R/dP is negative) and points D, D_1, etc., are unstable. The upper, stable, values represent possible system operating conditions; however, if the load is increased, or the load pf is decreased, the operating point moves towards the 'nose' of the curve, accompanied by a progressive reduction in voltage, which tends towards a complete collapse of the system.

Figure 41.3(*g*) includes the interesting result for a load with a leading pf. This increases both the load voltage and the maximum load that may be transferred.

The phenomenon of voltage instability has been the cause of, or a contributory factor in, a number of major system failures. The adequate provision and the correct control of the network var resources can play a vital role in preventing voltage instability.

41.3.3 Var balancing for steady state conditions

It is clear from the previous sections that there are significant benefits in local balancing of load vars by power factor correction. In addition to the improvement in load voltage conditions, the magnitude of the load current is reduced, which reduces transmission losses and releases circuit capacity for increased power transfer. Capacitors draw a reactive current that leads the voltage across their terminals and shunt capacitor banks provide a relatively economic means of improving the power factors of loads connected to a power supply system especially if the load is constant, or almost constant. When there is a slow variation of the load, shunt capacitors may be switched into or out of service to provide an approximate balance of the load vars. When the load varies quickly, it can no longer be regarded as 'steady state' and a compensating system that can be quickly controlled to counteract the load fluctuations must be used, as will be described later.

It is also clear that, if the vars absorbed in the supply inductance were to be reduced by series var balancing, the effective inductance of the supply would be reduced. The adverse effects of both active and reactive currents would then be lessened and the stability margin could be increased.

41.3.4 Power transmission over long distances

The use of reactive compensation to improve transient stability and so increase the amount of power that can be transmitted over given transmission lines is discussed next. In contrast to the example of a simple load discussed above, it is usual for machines to be present and to provide voltage sources within the systems at each end of the line.

41.3.5 Transmission characteristics

A number of performance features are associated with long distance transmission.

At light loads which are only a small fraction of SIL, or at energisation of the line from one end only, the distributed inductance and capacitance of the line can cause a large overvoltage due to the Ferranti effect; this must be countered by either (a) reducing the vars from the line capacitance

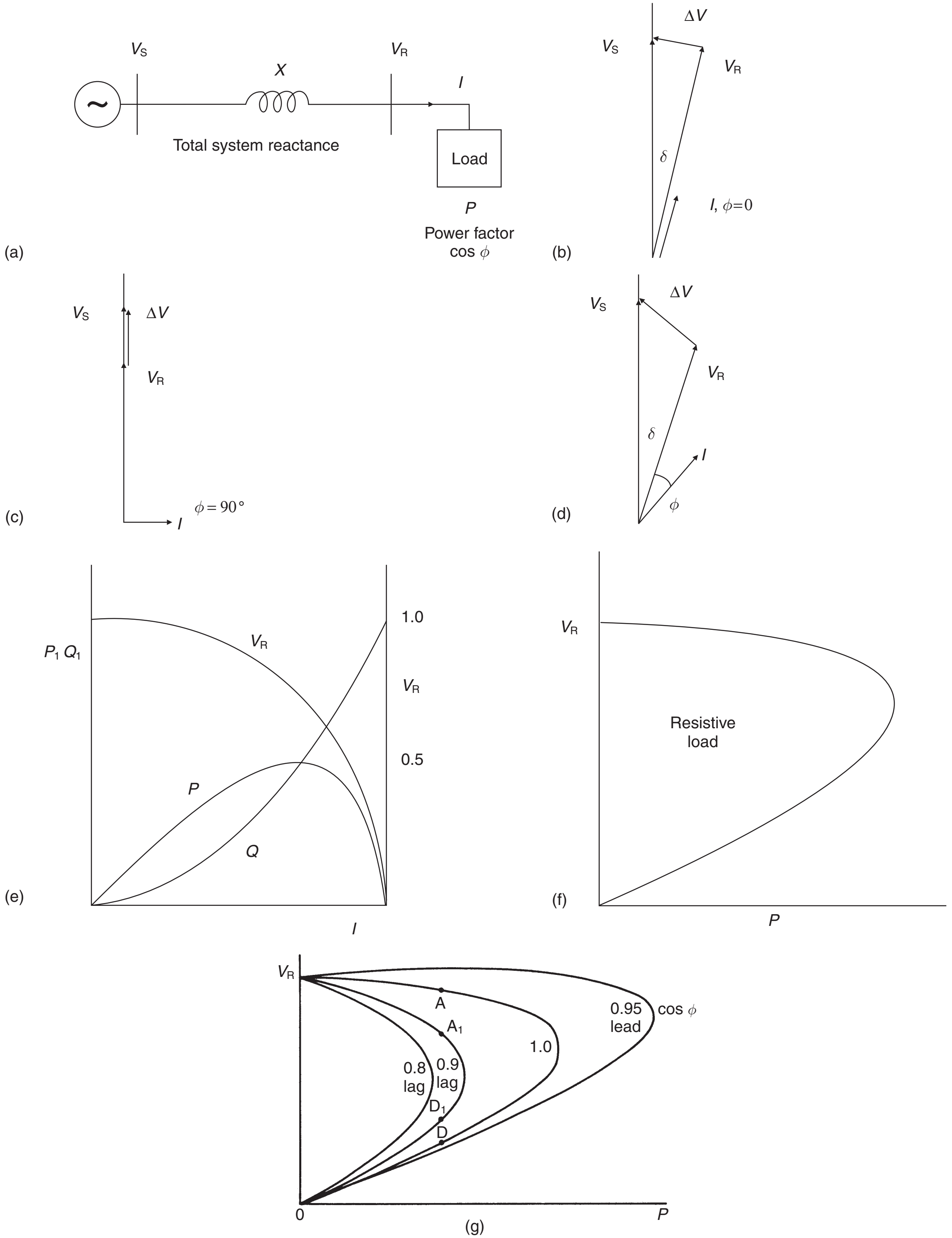

V_S
X
V_R
I
Total system reactance
Load
P
Power factor
cos ϕ
(a)
ΔV
V_S
V_R
δ
I, $\phi = 0$
(b)
V_S
ΔV
V_R
$\phi = 90°$
I
(c)
ΔV
V_S
V_R
δ
I
ϕ
(d)
P_1 Q_1
V_R
1.0
V_R
0.5
P
Q
I
(e)
V_R
Resistive load
P
(f)
V_R
A
A_1
0.95 lead
cos ϕ
1.0
0.8 lag
0.9 lag
D_1
D
0
P
(g)

Figure 41.3 Transmission line voltage/load characteristics

by shunt reactive compensation, or (b) cancelling part of the line inductance by series reactive compensation.

The maximum power that can be transmitted down a long uncompensated line is limited by its series inductance and by the line surge impedance. Increase in transmitted power can be achieved by series or shunt reactive compensation.

Figure 41.4(*a*) represents a long transmission line, with a total series inductive impedance X; to simplify this illustration, the effect of the line shunt capacitance is ignored. In *Figure 41.4*(*b*) the voltages V_S and V_R at the two ends of the line are assumed to be maintained constant and equal for all values of current, I. As the current increases, so does the angle, δ, between the voltages. The voltage at the mid-point of the line will be

$$V_M = V_S \cos \delta/2$$

and, with $V_S = V_R$, will be in phase with the line current.

The power flowing through the line will be

$$P = V_M \cdot I = (V_S V_R/X) \cdot \sin \delta = (V_S^2/X) \cdot \sin \delta$$

As shown in *Figure 41.4*(*c*) the transferred power rises to a maximum (given by $P_{max} = V_S^2/X$) as the angle δ reaches 90°. The power decreases as δ increases beyond 90°. It may be noted that control of the load voltage has doubled the power obtainable in the simple case illustrated in Section 41.3.2 where V_R is not controlled.

The shunt capacitance of a practical line will have the effect of increasing the voltage at all intermediate points along the line, including the mid-point, and consequently will give a corresponding increase in the maximum power. However, the voltage rise must not be allowed to exceed the maximum operating voltage of the system under no load or light load conditions.

Owing to the inertia of the rotating machines connected within the networks at either end of transmission lines, any system operating near its steady-state stability limit (corresponding to the phase angle of 90°) is bound to become unstable following a major disturbance, due to the ensuing increase of phase angles. It is therefore necessary to design a line to be able to transmit more than pre-fault power up to the point of maximum angular swing. Consequently, sufficient var generation must be available to compensate for the increased var consumption by the line current at increased phase angles. Thus, for transient stability to be secured, some excess var generation must be available. This can be achieved either by operating the line sufficiently below its surge impedance power before the fault or by adding var generation transiently for the period in which the line phase angle will be abnormally increased. However, without adequately fast voltage control along the line, either solution can lead to dangerous overvoltages, particularly under the condition of a severe 'back-swing' (i.e. a transient phase condition when the power transmitted is much less than the pre-fault level) or during a load rejection situation consequent to loss of stability.

Then the total vars Q to be absorbed from the line when operating at a voltage V and power P approximates to $Q = Q_0\,[V^2 - (P/V)^2]$ where Q_0 is the vars generated by the line shunt capacitance, C, at rated voltage V_1. Here P, Q, and Q_0 are expressed in per-unit of the SIL, P_S, and V is expressed in per-unit of V_1.

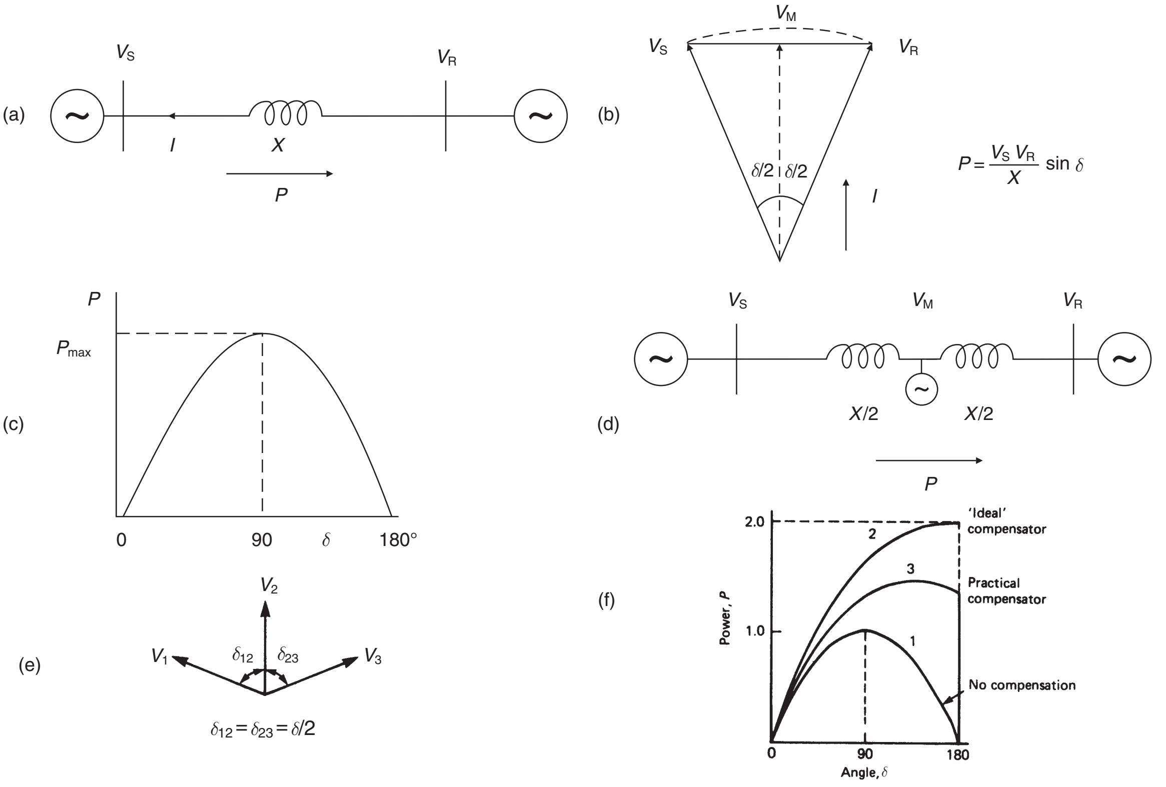


Figure 41.4 Power–angle features of a transmission line with and without shunt var compensation

For $V = 1$ p.u., Q is equal to Q_0 at $P = 0$, and falls to zero for $P = 1$, i.e. for a power equal to P_S. Power can be increased above P_S, if Q can be made negative, i.e. if var generation is added.

41.3.6 Transient instability

There has been a tendency to increase the distance over which power is transmitted for the following typical requirements:

(a) the supply of large amounts of power to conurbations and industrial centres from economic, but distant, energy sources—e.g. hydroelectric or mine-mouth thermal stations;
(b) the supply of moderate amounts of power to areas remote from the power grid of the central region—i.e. by radial sub-transmission lines, having loads tapped at one or several locations along each line.

Transient instability can occur in a transmission system following a sudden disturbance such as a short-circuit. The power output of a generator located near a fault may be greatly reduced, thereby allowing its speed to increase, while the output of a remote generator may be hardly affected by the event. As the acceleration of the two generators will differ, the phase angle between their internal emfs will change. In a simple system the two generators will drop out of synchronism if the fault persists long enough to allow the phase angle between their emfs to swing too far above 90°. The loss of synchronism takes place usually within 1 s of the disturbance. It is clear that the larger the series inductive reactance of the transmission lines and system transformers in the network interconnecting the generators, the greater attention must be paid to the possibility of phase-angle instability. In a well-interconnected system that does not have long distances between generators, transient and dynamic instability is rarely a difficult problem because fast clearance times can be obtained by using modern relays and circuit breakers.

On the other hand, slow swings or oscillations may develop between two parts of the system (or perhaps between one remote generator and the rest of the system), possibly initiated by minor disturbances, and these may build up to unacceptable levels after perhaps 20 or 30 seconds. Modern automatic controllers for generator excitation, with auxiliary stabilising signals, can exert some damping on such oscillations. Shunt or series dynamic compensation may be used alternatively or in addition to generator excitation control.

41.3.7 Var balancing for dynamic conditions

Figure 41.4(*d*) represents the same transmission line as *Figure 41.4*(*a*) but with the assumption that the mid-point is connected to a virtual infinite bus-bar, which will always keep the magnitude of V_M constant and equal to the voltages at the ends of the line. In this context, the 'infinite bus-bar' is not required to supply active power, (i.e. it will not control the phase angle of V_M) but must behave as an 'ideal' FACTS controller to supply the vars required to control the voltage at the mid-point.

Figure 41.4(*e*) shows that this additional voltage support effectively divides the line into two equal sections, each with a reactance of $X/2$ and operating at an angle, $\delta/2$. Each line section now has an independent stability limit which is reached when $\delta/2 = 90°$. Thus the theoretical total angle between V_S and V_R can now be 180°, i.e. double the conventional critical value of 90°, with a resultant increase in power transfer capability.

Without any voltage support at the mid-point the power will be as in curve 1 of *Figure 41.4*(*f*) (repeated from *Figure 41.4*(*c*)).

With voltage support at the mid-point, the condition will be as in curves 2 and 3. If the vars were to be supplied instantaneously, the power transfer would be doubled (curve 2),

i.e. $P_{\max} = (V_S^2)/(X/2)$

In practice it is not possible for the theoretical maximum to be achieved, owing to losses and the response delay of a practical controller (curve 3). However, it is possible to provide voltage support at several places along the line and so make up, to some extent, for the lack of the 'ideal controller'.

Figure 41.5 illustrates an example of conditions to be expected on a very long high-voltage transmission line with voltage support, provided for example by SVCs (static var compensators) at two intermediate substations, *Figure 41.5*(*a*).

For each of the three sections there are two conditions that can be responsible for limiting the maximum power that is reached.

Assume that the controllers can hold the voltage rigidly constant irrespective of load. In this case the stability limit will theoretically be reached if the phase angle between two adjacent voltage controlled bus-bars reaches 90°. In practice, even with controllers, the resistance of the line and the controller response delays limit the total angle at maximum power to less than the theoretical maximum of 3 × 90°, e.g. to 175° in *Figure 41.5*(*b*). *Figure 41.5*(*c*) illustrates the phase angles of the line at this steady-state power limit.

The second limiting condition might be that any one of the controllers reaches the limit of its var generation

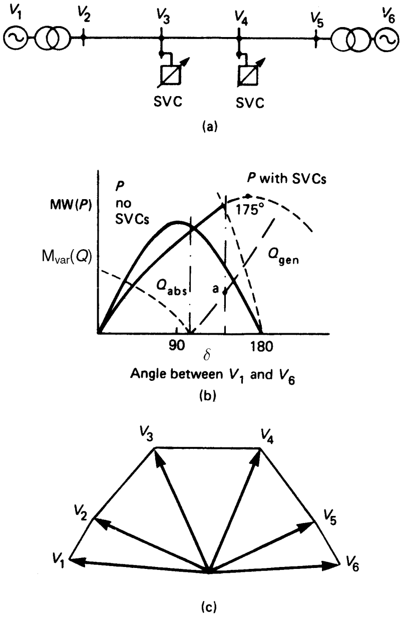


Figure 41.5 Long transmission line with multiple static var compensators

capacity; the required vars from the controller are illustrated in *Figure 41.5*(*b*) by the curve *Q* (generation or absorption). When the power *P* and angle δ reach the condition of this limit, as for point 'a' in *Figure 41.5*(*b*), the controller ceases to be effective in controlling the line voltage at its point of connection and the stability limit is reached even if the total angle has not reached the expected limit (e.g. 175°).

The action of a shunt controller in improving the transmission capacity of a line can be thought of in terms of its effect on the natural surge impedance of the line; the shunt *C* is controlled so that the effective surge impedance matches the load being transferred. For heavy loads the shunt *C* is increased and the surge impedance is decreased, thereby increasing the value of SIL.

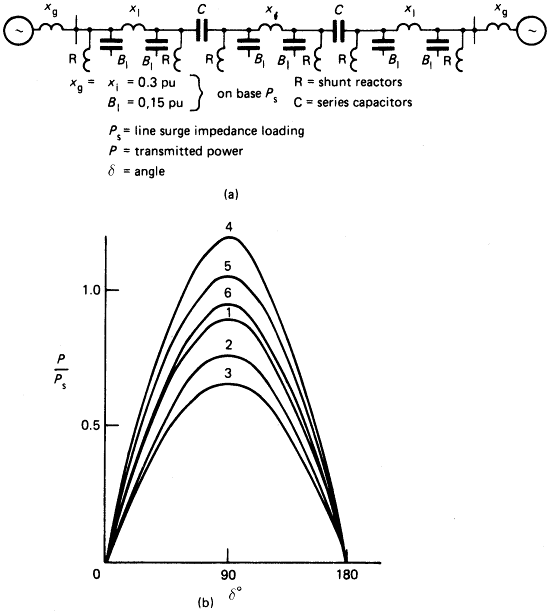
An alternative way of reducing the surge impedance in order to provide extra margin for a long transmission line is to reduce the series inductive reactance; this can be done by inserting a controller with a negative inductive reactance (usually in the form of capacitive reactance). Series compensation inserted in one or two locations along a transmission line has the following important characteristics.

(1) By reducing the effective series inductive reactance of a transmission line, series compensation allows a higher power transfer capability. This enhancement may provide a higher power over a given distance or enable a given power to be transmitted over a greater distance; alternatively, series compensation may be used to reduce the steady-state phase angle for a given power and distance.
(2) By increasing the (capacitive) vars generated as the load current increases, the series compensation improves the var balance in a line and hence reduces its voltage regulation, thereby reducing shunt var generation requirements at the line terminals.

Even when a long h.v. or e.h.v. line is series compensated, shunt reactors are normally needed to absorb part of the shunt vars of the line. Typically between 40% and 100% of line shunt capacitive power may be compensated at light load to prevent overvoltages on the line. To restrict insulation stresses caused by overvoltages following sudden load rejection, a substantial part of the shunt reactive compensation is usually left permanently connected; clearly, this reduces the maximum power limit of the line.

Figure 41.6(*a*) shows an illustrative theoretical example of the power–angle characteristic of a line including its terminal impedances; the line is represented as three π-sections of about 200 km length each, with shunt reactors (*R*) at each line section terminal and series capacitors (*C*) at two intermediate locations. The curves in *Figure 41.6*(*b*) were calculated neglecting power losses, and the values are given in per-unit terms on the base of the surge impedance load (P_S).

As the degree of shunt compensation is increased, the line requires a greater amount of series compensation to maintain the power–angle characteristics of the line. A comparison of curves 1 and 6 and the table in *Figure 41.6*(*c*) shows that with 100% shunt-reactor compensation the line requires almost 50% series compensation to reach the same maximum stable power limit as for the case without any shunt or series compensation. The maximum allowable limit of series capacitor compensation and choice of its location are governed not only by economic considerations, but also by various practical aspects, such as subsynchronous resonance (SSR, described later) and by the series capacitor short-circuit overcurrent protection needs. The latter requires spark gaps or non-linear resistors. The line distance protection relaying must be co-ordinated with these features.



Curve	Compensation of line shunt capacitance (%)	Compensation of line series reactance (%)	$\frac{P_{max}}{P_s}$
1	0	0	0.91
2	50	0	0.77
3	100	0	0.66
4	0	50	1.20
5	50	50	1.06
6	100	50	0.95

(c)

Figure 41.6 Transmission line with series and shunt compensation

41.4 The management of vars

41.4.1 Objectives

Traditionally, a single organisation was responsible for all aspects of generation, transmission and distribution but, increasingly, these functions are being separated and allocated to organisations that are independent of each other. This subdivision of responsibilities usually results in competition between generating companies to supply power to the transmission network and often between distribution companies to supply power to customers, but the responsibility for transmission sometimes rests with a single organisation. The transmission company must then ensure that an adequate supply of energy is called up from the generators and that satisfactory levels of voltage exist throughout the transmission system in order to satisfy the moment by moment demands of the distribution companies; this requires the effective management of vars at the points where the energy is received and delivered, as well as within the transmission system itself. Adequate provision of vars at all times can be crucial for the security of a system.

It is necessary to balance various factors to arrive at an optimum solution for the management of vars. The objectives in such management must be both economic and technical.

(1) to optimise the capital investment in, and location of, plant for the generation and absorption of vars;
(2) to minimise system losses;

(3) to optimise plant utilisation, so postponing the need for system reinforcement;
(4) to obtain adequate system security;
(5) to maintain adequate quality of supply; and
(6) to control system overvoltages.

41.4.2 Tools for var management

To attain the above objectives of efficient var management, system designers, operators and consumers can use some or all of the following types of plant or procedures: generators; on-load tap-changing transformers; switching of transmission circuits; various types of FACTS controller. Later sections provide information on the development and application of these controllers, which include shunt reactors and capacitor banks, switched or unswitched; series compensation; synchronous compensators; static var compensators; STATCOM; other FACTS controllers.

41.4.3 Var management by the supply authorities

The benefit of good var management by the supply authority can be significant. Three areas of management can be identified.

The first is in the system planning and design stage, where a choice must be made on the rating and type of any reactive compensation equipment to be used, where it should be connected in the network, and when is the optimum time for installation. The characteristics of any power system, particularly the proportion of generation connected at the different transmission voltage levels and the ratio between the system maximum and minimum loading, greatly influence the decisions to be made.

Most transmission networks, even at peak demand, operate below their surge impedance load and hence generate more vars than they absorb, i.e. they have a positive var balance. However, under emergency conditions following a loss of one or more lines, the positive balance may turn into a deficit which under extreme conditions can give problems of voltage instability as well as possible loss of synchronism due to the network's low power-transfer capability.

A var planning study involves the balancing of the vars in the complete network. Then, after allocating an amount of var production to generating plant, it is necessary to decide on the best method to deal with the resultant surplus or deficit. For a large system this can require a protracted and major design exercise, and over the last two decades much effort has gone into the development of mathematical and computational techniques to aid system designers in this task.

The second area of var management is in system operation, where the facilities provided must be used in the most efficient manner to maximise system operational security and minimise system losses. Again much work has been done on the development of both off-line and on-line computer programs. Various studies have shown that, in general, minimising network losses also gives maximum security and minimum unnecessary circulation of vars in the network.

The third area is the commercial one of tariff policy, in which financial incentives are adopted to persuade consumers to maintain a power factor close to unity.

41.4.4 Var management by consumers

The induction motor is by far the most common driving machine and, in consequence, the natural power factor of most industrial loads is in the 0.7–0.8 lagging region in the absence of reactive compensation. Such a load requires a var supply of between 75 and 100% of its power, increasing the loading on the plant (e.g. transformers and cables) and causing significant losses and voltage drops. By generating the required vars close to the loads, these disadvantages can be largely overcome and for this reason supply authorities impose tariffs on industrial consumers to encourage them to install power-factor correction plant within their systems. This can often be achieved in a relatively inexpensive way by shunt capacitors, either direct on the motor terminals or as larger banks on supply voltage bus-bars. Often a combination is the optimum solution.

Fixed capacitors or switched banks may not always be adequate where the var requirements fluctuate widely and rapidly and cause greater voltage variations at a point of common coupling with other consumers than the supply authority allows. Such loads are found, for example, in the steel industry where, with thyristor driven rolling mills, the sudden impact of the billet meeting the rolls produces a rapid increase in low-power-factor current; and with electric arc furnaces, where frequent and random short-circuits occur between electrodes, producing short bursts of high current, again at a very low power factor.

Most supply authorities have regulations covering allowable limits for voltage fluctuations, particularly in respect of lamp flicker; as the human eye is most sensitive to fluctuation frequencies in the region of 8–10 Hz, these limits vary with frequency. Typically they might allow only 0.25% voltage dips at 10 Hz but 1–2% at 1 Hz. To compensate for such conditions requires a fast and continuously variable reactive source such as some types of static var compensator.

Thyristor power conversion equipment (which is increasingly being used in large ratings) and electric arc furnaces generate harmonic currents that flow into the supply network. These can affect other plants, causing losses, possible overheating or interference with electronic and telecommunication circuits. Supply authorities must therefore limit the harmonic distortion on their networks and many authorities stipulate maximum permissible values. To avoid exceeding these, it may be necessary to prevent a proportion of the harmonic currents generated by the consumer from penetrating the supply system. It is common to bypass these currents into harmonic filters. It is often possible to convert capacitor banks for power-factor correction into harmonic filters, giving significant cost savings. In any case the harmonic filters will generate vars at the power supply frequency and these must be accounted for in the system var balance.

41.5 The development of FACTS controllers

Earlier sections have pointed to the important part that reactors and capacitors play in improving the steady state characteristics and capacity of a.c. power systems. Although the application of reactors and capacitors can often be enhanced for slowly varying conditions by means of mechanical switching, faster changes of compensating vars are needed to provide dynamic improvements. At one time, the synchronous compensator was the only tool available for dynamic compensation; it acted as the benchmark for the early development of static FACTS controllers, which now provide a wide range of functions and characteristics for both shunt and series application. The following sections provide a simple outline of the components that have been developed to provide variable var output for FACTS controllers.

41.5.1 Synchronous compensators

All synchronous machines can give continuously variable var compensation and, in industrial systems where large synchronous motors are installed, they are frequently used in this way to provide power factor correction of the local load in addition to their main driving duty. Self-driven synchronous machines not connected to any mechanical load (synchronous compensators) can be used anywhere in a system either to generate or to absorb vars on a balanced three-phase basis. The synchronous compensator directly generates a voltage at its terminals. The magnitude of this voltage is controlled by a high-speed excitation system, which is often arranged to provide control of the voltage at another point, usually at the high voltage side of the compensator transformer. When overexcited, the machine generates vars and operates in a stable condition; when underexcited, it absorbs vars but its stability decreases as the excitation falls towards zero. Because of this tendency towards unstable operation, the var absorption rating of a synchronous compensator is typically only 50% of its var generation rating. Most machines operate with an automatic excitation control that needs to have a rapid response to assist the machine in maintaining stability through system disturbances or to follow rapid reactive load changes. The control time constant is normally in the range from 5–10 cycles.

Figure 41.7(*a*) illustrates the voltage–current characteristics of a synchronous compensator. The rated excitation voltage corresponds to the rated voltage of the system. If the system voltage falls below its rated value to V_d and the field current remains unchanged, the var generation will initially follow the sub-transient characteristic to point B but it will finally settle to point A on the synchronous reactance characteristic. In practical applications of synchronous compensators, the machine excitation will usually be controlled either to maintain a nominally constant Mvar output or to maintain a nominally constant ‘target voltage’ on the system, subject to a defined ‘droop’, or slope characteristic. Thus, for 5% droop, and starting from the float condition with zero output, a reduction of 5% of terminal voltage will provoke an increase in excitation sufficient to generate rated leading current. Conversely, a voltage rise of about 2.5% will cause the synchronous compensator to absorb its rated lagging current. When there is a sudden disturbance, the output will initially follow the sub-transient characteristic as before, but after a few cycles the fast excitation control will cause the output to settle at the appropriate point C on the controlled characteristic. As shown in *Figure 41.7*(*b*) the slope of this characteristic is equivalent to a reactive impedance, X_m, such that a voltage change, ΔV_m, at the controlled bus-bar will cause a reactive current change ΔI_m in the synchronous compensator, i.e.

$$\Delta I_\mathrm{m} = \Delta V_\mathrm{m}/X_\mathrm{m}$$

Synchronous compensators have several limiting or disadvantageous features; costs are relatively high and losses significant; they need good foundations, buildings, complex auxiliary systems and regular maintenance and refurbishment that detracts from availability. These drawbacks prompted the development of alternative types of equipment, using static components, with the synchronous compensator acting as the initial benchmark for dynamic response.

41.5.2 Sign convention for vars and reactive current

There are standard sign conventions for power and vars when current flows from a system bus-bar into a load. Power is positive for the component of current that is in phase with the voltage; vars are positive for a load operating at lagging power factor, i.e. when the current flowing into the load lags the voltage on the bus-bar. By this convention, when treated as loads as in *Figure 41.7*, shunt reactors will draw positive vars from the system, shunt capacitors will draw negative vars and generators will draw negative power. However, a reversed sign convention has been adopted to define the ratings of var compensation equipment for the following reason.

Generators supply both the power and the vars to satisfy the demands of the system and its loads. The sign convention for generators is that both power and vars are positive when current is flowing from the generator into the system bus-bar to supply a lagging power factor load. A synchronous compensator behaves in the same way as a generator operating at zero power and therefore it uses the same sign convention for vars. Thus, vars are positive when the synchronous compensator is over-excited, acting as a shunt capacitor and supplying an inductive (positive var) load and negative when it is under-excited, supplying a capacitive (negative var) load.

The same sign convention has become the norm for defining and specifying the var duties and output ratings of shunt static var plant and is thus consistent with that for rotating plant, i.e. shunt capacitive vars are designated as positive and shunt inductive vars are designated as negative. Note, however, that as shown in *Figure 41.7*, which treats the

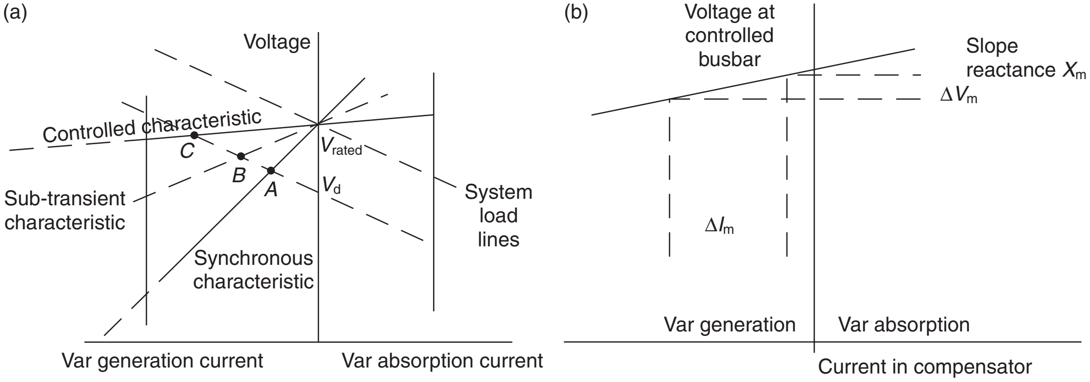


Figure 41.7 Voltage–current characteristics of a synchronous compensator

compensating plant as a load, the corresponding currents have opposite signs to those for vars.

41.5.3 Basic features of static compensation

The main elements of the traditional type of static var compensator (SVC) are a capacitor to generate vars and an inductor to absorb vars. To provide operation in both the generation and absorption modes, both elements must be used; at least one of them must be rapidly variable, *Figure 41.8*(*a*), in order to replicate the dynamic characteristic of a synchronous compensator shown in *Figure 41.7*(*b*). A capacitor element cannot give a stepless variation: therefore, if a smooth continuous variation is needed over the whole range A–A′ in *Figure 41.8*(*b*), a capacitor bank having a rating to give at least the current I_C must be connected and the variable element must be inductive, with a rating to give a current at least equal to $I_C + I_L$.

Although a SVC does not generate a real voltage, its dynamic voltage–current (*V–I*) characteristic enables it to be represented in studies of the system behaviour as if it were a synchronous compensator. The reference parameters for the SVC's characteristic are a target voltage, V_{ref}, equivalent to the excitation voltage of a synchronous compensator, and an equivalent slope reactance, *X*. The SVC inductive current, *I*, (or capacitive current if *I* is negative) at an actual system voltage, *V*, is given by the function

$$I = (V - V_{ref})/X$$

The equivalent machine can be controlled to generate and absorb vars over a defined range of system voltage variation, but has no inertia. At the limits of the linear range, the SVC representation is changed to that of a simple shunt capacitor or reactor, as appropriate.

In some system applications the SVC must be capable of operating under different system conditions at the extremes of its range (i.e. at A and A′), but it need not be capable of swinging in a continuous manner from one extreme to the other. Its 'dynamic' or 'swing' range, which is the range over which it can give a very fast response, may be only a proportion of its total range. In such a design the size of the variable element, which is the most expensive item in per unit of rating, can be reduced to deal only with the required swing; mechanically switchable L and/or C elements, *Figure 41.8*(*c*), can then be controlled automatically to adjust the position of the dynamic range within the total range. The overall cost of the SVC is then reduced. Sometimes the total range of a SVC is extended by using its control system to control the switching of shunt capacitors or reactors connected elsewhere in the substation.

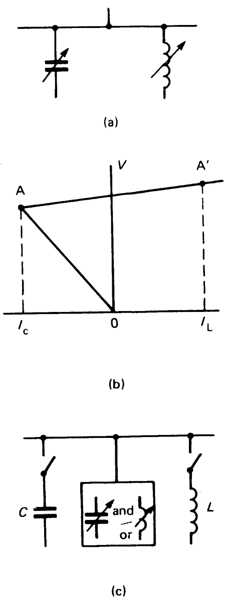


Figure 41.8 Basic static var compensator (SVC)

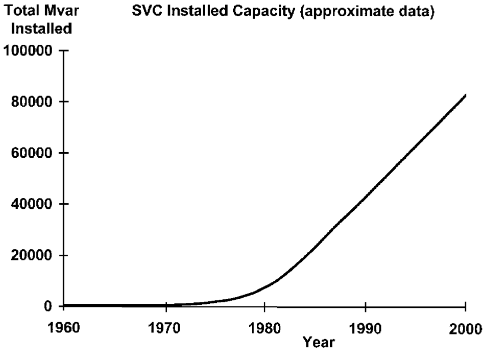


Figure 41.9 SVC installed capacity (data approximate only)

The first commercial application of a SVC was in 1964. It used a self-saturated reactor to reduce the lamp flicker caused by an electric arc furnace. The first static var compensator for a transmission system was installed in 1967, followed by another in 1969: both used saturated reactors (SRs) as the continuously variable inductive component, the first was a d.c.-controlled transductor and the second a self-saturated reactor with harmonic-compensation and slope correction.

During the next decade many SVCs were commissioned, particularly on industrial systems where SVCs were installed to reduce voltage disturbances and to suppress disturbing lamp flicker caused by the var fluctuations of loads such as arc furnaces, large converter-fed drives in rolling mills and, in some cases, particle accelerators. The first thyristor-controlled reactor (TCR) for a transmission system entered service in 1978 and this coincided with a rapid growth in such applications of both thyristor and SR types of compensator, *Figure 41.9*. This growth resulted mainly from two factors. First, the rapid increase in the cost of power made utilities more conscious of the need to minimise transmission losses. Second, the growing opposition in industrialised countries to the construction of new transmission lines encouraged the maximisation of power transfer down existing rights of way. The 1990s saw the advent of a completely new type of static controller, the 'STATCOM', which uses voltage sourced converter technology.

41.5.4 Harmonic-compensated self-saturated reactor

The saturated reactor (SR) is the original type of SVC; it gives an inherent variation of reactive current with voltage and its response does not depend on an imposed control system. The variable element comprises an iron-cored inductor whose core becomes saturated during each half cycle of the supply voltage and this core saturation results in a r.m.s. voltage–current characteristic as shown in *Figure 41.10*. This voltage–current characteristic resembles the controlled characteristic of a synchronous compensator. A saturated reactor can only absorb vars from the system

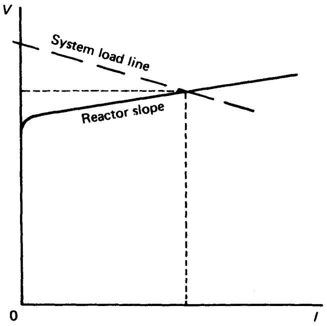


Figure 41.10 Saturated reactor characteristic

but not generate vars; its saturation voltage is the 'excitation voltage' of an equivalent machine and its slope reactance is the 'droop'. The operating current of the saturated reactor automatically adjusts itself to the point at which its characteristic intersects the system load line. If the terminal voltage falls below the saturation voltage, the SR current becomes negligibly small. Shunt capacitor banks are used to enable a SR type of SVC to generate vars.

An iron-cored inductor with a simple winding is rarely used on a power network, as it generates large magnitudes of odd harmonic currents. However, by interconnecting sections of all three-phase windings it is possible to cancel out most of these harmonics in normal operation. The twin-tripler and treble-tripler harmonic compensated SRs (invented and developed by Dr E. Friedlander of the former General Electric Company, England) use this principle and have, respectively, six and nine active iron-cored limbs with inter-star connected windings and ratings up to 170 MVA for a treble-tripler saturated reactor, *Figure 41.11*. In addition to reducing the harmonic content of the reactor current to negligible proportions the designs give a voltage–current characteristic with a sharp knee point and a slope that remains linear to within about 1%, for currents from about 10% to more than 300% of rated load. The slope reactance is normally within the range 8–15% on the basis of nominal full-load rating.

41.5.5 Thyristor switches

In the 1960s, the demand for thyristors in industrial applications led to the very rapid development of devices with high withstand voltages and also capable of operating at high currents. Thyristor pairs with inverse-parallel connection *Figure 41.12*(*a*) are suitable for use as static switches, capable of extremely fast and frequent switching operation without the time delays, wear and tear and maintenance requirements of mechanical switchgear. A.c. thyristor switches were at the heart of most of the SVCs supplied during the 1980s and 1990s.

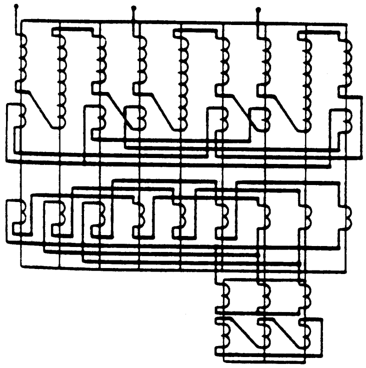


Figure 41.11 Circuit of the treble-tripler reactor

For a practical application, thyristors are assembled in a modular way into an 'a.c. thyristor valve'. An a.c. valve is made up of inverse-parallel connected thyristors, which are themselves connected in series (and occasionally in parallel) to obtain the necessary rating, with voltages up to 40 kV and currents up to about 6 kA. The cost of a thyristor valve is influenced not only by the ratings of the thyristors, but also by the number of voltage levels in series. Thyristors are available with voltage ratings up to about 8 kV (peak), although derating factors of 2 or more need to be applied in a series string of thyristors to determine the actual rms operating voltage of each thyristor.

Differences in the thyristor leakage currents, stray capacitances and stored charge all tend to cause an uneven voltage distribution between series-connected thyristors. A resistor–capacitor a.c. grading network is used in parallel with a d.c. grading resistor across each thyristor level, *Figure 41.12*(*b*). These grading networks reduce the voltage unbalance, and the *R–C* network also serves to damp the transient overshoots in the reverse voltage that occur when thyristors come out of conduction (twice per cycle per valve). Individual thyristor levels within a module are protected against overvoltage by breakover diodes or similar devices, which provide a gate pulse for forward firing of the thyristors above a pre-set overvoltage level. Special care must be taken in the design of the gating electronics and the signal transmission circuits to ensure coherent (simultaneous) firing of all the series-connected thyristors.

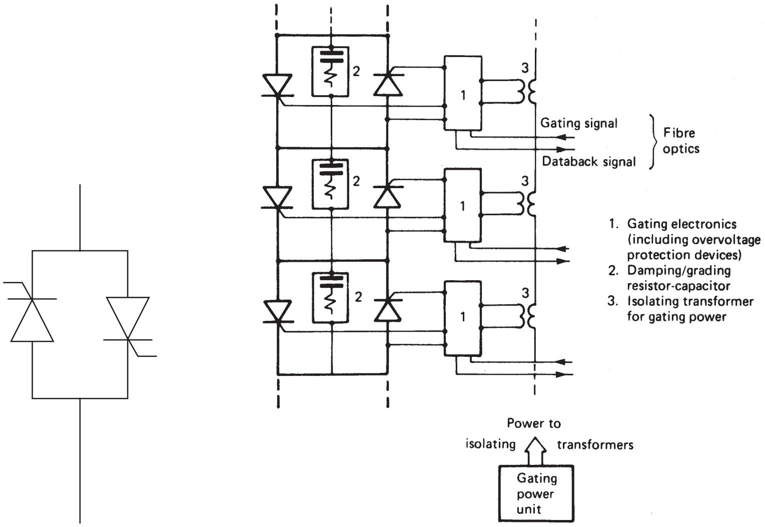
Most thyristor valves are indoor installations, with air insulation and water cooling. Only pure demineralised water having very high resistivity is allowed to flow through the thyristor heat sinks, which are at high voltage with respect to ground. The water cooling plant is mounted separately at ground potential and the pipework from ground to the thyristor stacks includes some sections made of insulating material. Air cooling is much simpler than water cooling, but it limits utilisation of the current capability of the thyristors and is rarely used for SVC applications. *Figure 41.13* shows one type of thyristor valve used in a SVC installation in a transmission network.

The thyristor-controlled reactor (TCR) comprises an a.c. thyristor valve connected in series with a linear reactor. One phase of a TCR is shown in *Figure 41.14*. Variation of the current in the TCR is obtained by control of the thyristor conduction duration in each half-cycle, from a 90° firing angle delay (as measured from the applied voltage zero) for full conduction to 180° delay for no conduction.

A.c. thyristor valves consisting of inverse-parallel connected thyristors, very similar to those used for the TCR, are used to enable rapid and frequent switching of blocks of capacitors, *Figure 41.15*. The thyristor-switched capacitor (TSC) gives directly the effect of a variable capacitance, although for reasons given below the variation is in steps.

Once gated, the thyristor valve will conduct for a half-period and unless again gated the current will then cease. A TSC, therefore, can only give a variation of capacitance between two states, on and off, for a discrete period of an integral number of half-periods. The kind of point-on-wave control used to give variable output with a TCR is not a practicable option for a TSC.

At the end of each half cycle of conduction, the capacitor current through the thyristor reaches zero; unless re-gating


2
3
1
Gating signal
Databack signal
Fibre optics
1. Gating electronics (including overvoltage protection devices)
2. Damping/grading resistor-capacitor
3. Isolating transformer for gating power
Power to isolating transformers
Gating power unit

(a) (b)

Figure 41.12 (a) Basic thyristor switch; (b) typical thyristor valve circuit

occurs the capacitor will remain charged at peak voltage while the supply voltage peaks in the opposite polarity after a half-period. This imposes a doubled voltage stress on the non-conducting thyristor, so that a TSC valve usually needs approximately twice the number of thyristors in series compared with a TCR valve of equivalent voltage. Ideally, gating is arranged to take place when the voltage across the thyristor is zero and the supply voltage is at its peak (i.e. $\mathrm{d}v/\mathrm{d}t = 0$); this results in a transient-free energisation. Except when continuous conduction is achieved by repetitive re-gating, such ideal switching conditions are rarely achieved, owing to mismatch between the capacitor voltage (which normally slowly loses its charge from its previous energisation) and the system voltage (which may have changed its value during the same interval).

The TSC control system is normally arranged to choose the gating instant that gives the minimum transient. Nevertheless, quite large transients occur during initial energisation of an uncharged capacitor and current limiting reactors must normally be installed to reduce the magnitude of the inrush current to a value within the thyristor capability. With the exception of these switching transients the TSC does not produce harmonics.

It is possible to arrange for a very rapid discharge of the capacitors after each period of conduction; this can reduce the voltage stress on the non-conducting thyristor and provide more consistent conditions for the next energisation. With a delta-connected three-phase TSC, current zero is reached sequentially in the three phases. By connecting a suitably designed three-phase transformer across the capacitors, most of the energy stored in the capacitors can be commutated back into the system, and into discharge resistors, by transformer action. Core saturation effects supplement this transformer action.

41.5.6 Convertor-based FACTS controllers

41.5.6.1 Basic concepts

As described in Section 41.5.3, the behaviour of conventional SVCs is often represented by a model of an equivalent, but inertialess, machine incorporating an equivalent ‘excitation voltage’ and an equivalent ‘slope reactance’. However, there is a type of circuit, sometimes called a ‘synchronous voltage source’, in which a real alternating voltage, V_e, is produced from a d.c. source C_s by the process of inversion in a solid-state d.c. to a.c. converter; the a.c. voltage source is connected to a system bus-bar through a coupling reactance, X_e, *Figure 41.16*. The converter and its coupling reactance behave as an idealised rotating machine with an extremely fast response and, again, no physical inertia although there may be significant stored energy in some cases. The class of SVCs which uses this principle has been assigned the name STATCOM (a shortened form of STATic COMpensator).

The basic behaviour of a STATCOM is very similar to that of a synchronous compensator. If the voltage generated by the converter is less than the voltage of the system busbar to which it is connected, the STATCOM will act as an inductive load, drawing Mvar from the supply system. Conversely, a STATCOM will act as a shunt capacitor, generating Mvar into the supply system, when its generated voltage is higher than the system voltage. The rated inductive current and the rated capacitive current are usually equal, giving the STATCOM an almost symmetrical lagging and leading Mvar rating. This is very advantageous for some applications, but for those that require an asymmetrical MVar rating, it is usually economical to reduce the total STATCOM rating and to bias its output using conventional fixed or thyristor switched reactive elements.

The losses of the converter are normally supplied from the system, in the same way as for a machine, and not from the source of direct voltage or current. Nevertheless a STATCOM can exchange real power with the a.c. system if the d.c. source is arranged to supply or absorb the power that it is desired to exchange on the a.c. side.

41.5.6.2 Voltage-sourced converters

Various types of inverter circuit and source have been suggested and examined. D.c. voltage-sourced converters (VSCs) have received the most attention in the practical

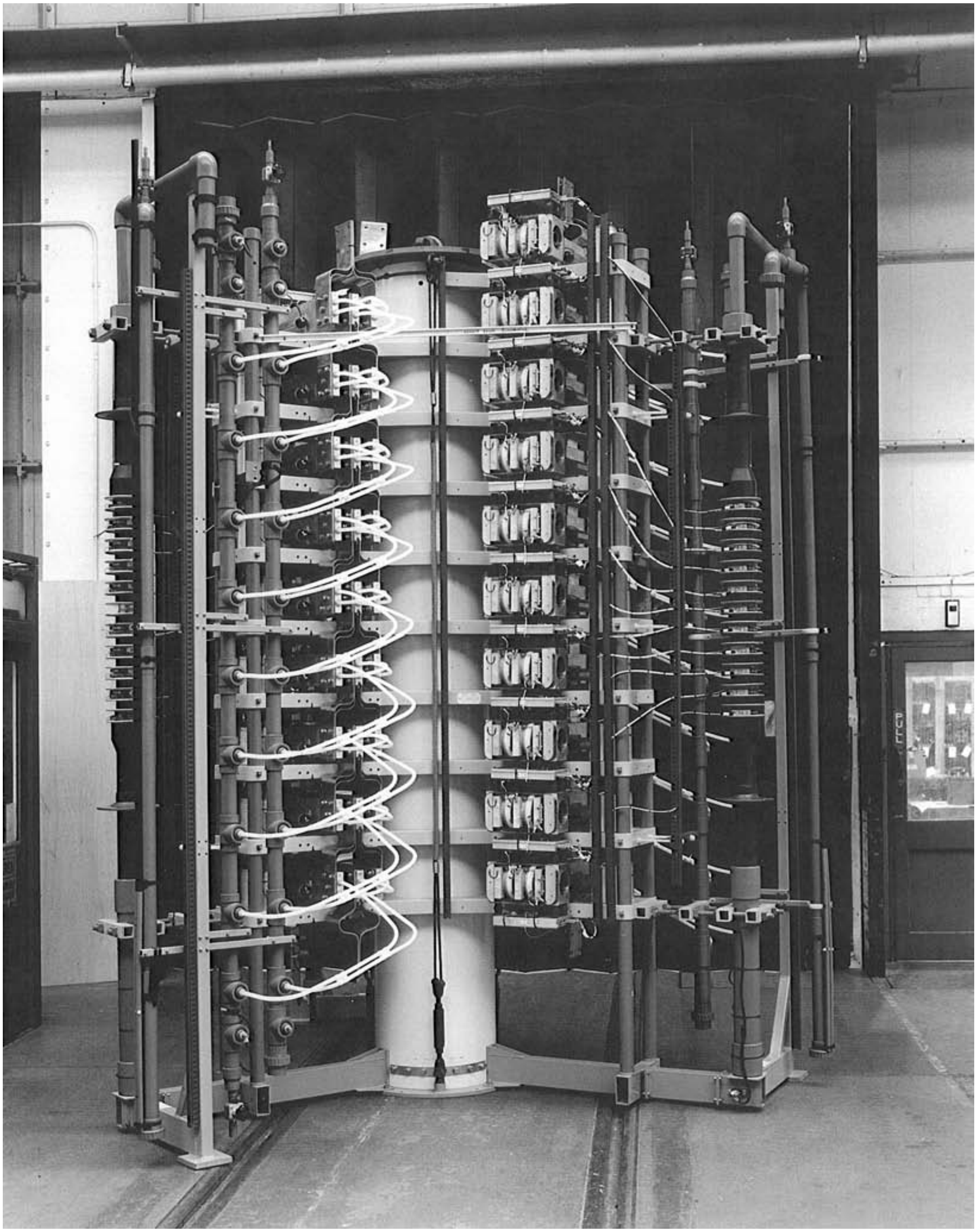


Figure 41.13 Thyristor valve for an SVC

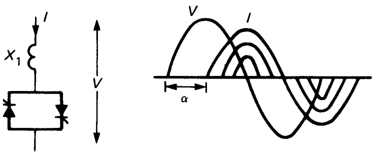


Figure 41.14 Basic TCR operation

realisation of the STATCOM principle and can also be used in FACTS controllers connected in series with a transmission line or load.

The inverter may use either conventional thyristors with forced commutation, or devices that have been designed to be turned off as well as turned on, such as gate turn-off (GTO) thyristors; these devices have been used for many years in drives for traction and industrial applications. Other devices, for example the Integrated Gate

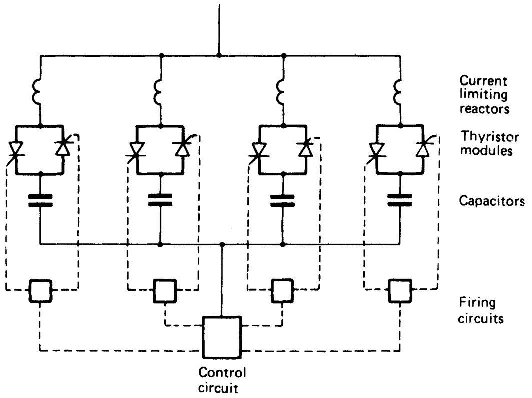


Figure 41.15 One phase of a thyristor-switched capacitor compensator

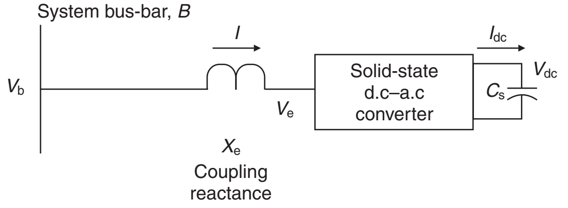


Figure 41.16 Synchronous voltage source

Commutated Thyristor (IGCT) or the Insulated Gate Bipolar Transistor (IGBT), require less energy for the switching process and are available with ratings that can readily be used in some FACTS controllers.

Figure 41.17 illustrates a basic single-phase inverter employing diodes and GTOs connected in inverse-parallel; the converter produces a square voltage waveform at the output terminal as it switches the direct voltage source on and off. This basic two level converter can be developed to produce various multi-level converters, such as the five level converter of *Figure 41.18*, for which the harmonic content is reduced and the voltage approximates more closely to a sine wave.

The direct voltage source can be a battery, whose output voltage is effectively constant and which has the capability to supply or absorb energy. For FACTS controllers a simple d.c. capacitor is usually used as the source; the d.c. terminal voltage can be raised or lowered by controlling the inverter in such a way as to increase or decrease the capacitor's stored energy.

41.5.6.3 Three-phase converters

Three single-phase converters, for example of the type illustrated in *Figure 41.17*, can be connected, one per phase, to form a three-phase converter. The three single-phase converters can be controlled in a co-ordinated way to generate a balanced three-phase set of voltages. There is also freedom, under system fault conditions—or for compensating unbalanced loads—to control each phase independently, in order to assist in balancing the system while avoiding converter overload. When using independent capacitors for the d.c. voltage sources, the voltage in each phase can easily be changed independently of the other two phases if necessary.

When the primary objective of the STATCOM is to respond to balanced load conditions, it is convenient to use the same d.c. source voltage for all three phases; the three converters can take the form of a Graetz bridge as shown in *Figure 41.19*(*a*). The three phase voltages are shown in *Figure 41.19*(*b*), with respect to the apparent neutral of the three-phase system. It will also be seen that the line-to-line voltages reach twice the peak magnitude of the phase voltages but for a shorter duration (120°⇐conduction angle rather than 180°;⇐due to the commutating action of the individual phases). This inherently eliminates third harmonic components from the line-to-line voltages under balanced system conditions, but during fault and unbalanced conditions, particular care is needed in the control and protection of the converters because of the use of a common d.c. source voltage.

41.5.6.4 Improving the voltage waveform

Clearly the basic square waveform is too distorted with low order harmonic components to be of practical application but can be improved by means of multi-pulse operation similar to that used for large rectifiers. Identical three-phase bridges (as in *Figure 41.19*(*a*)) can be connected to transformers that have phase-displaced outputs. Star- and delta-connected windings have a relative 30°⇐phase shift and a converter bridge connected to each transformer would give 12-pulse operation and eliminate 5th and 7th harmonics from the line currents for balanced operating conditions. The transformers may be operated in parallel, in which case 5th and 7th harmonic currents will circulate between the converters, limited by the transformer reactance. Alternatively the transformers can be connected in series so that their respective 5th and 7th harmonic voltages

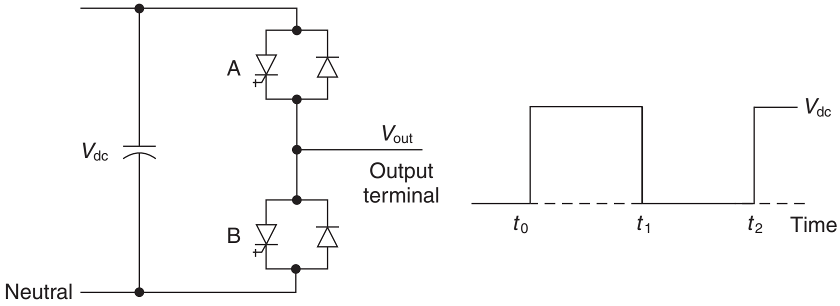


Figure 41.17 Basic single-phase voltage-sourced converter

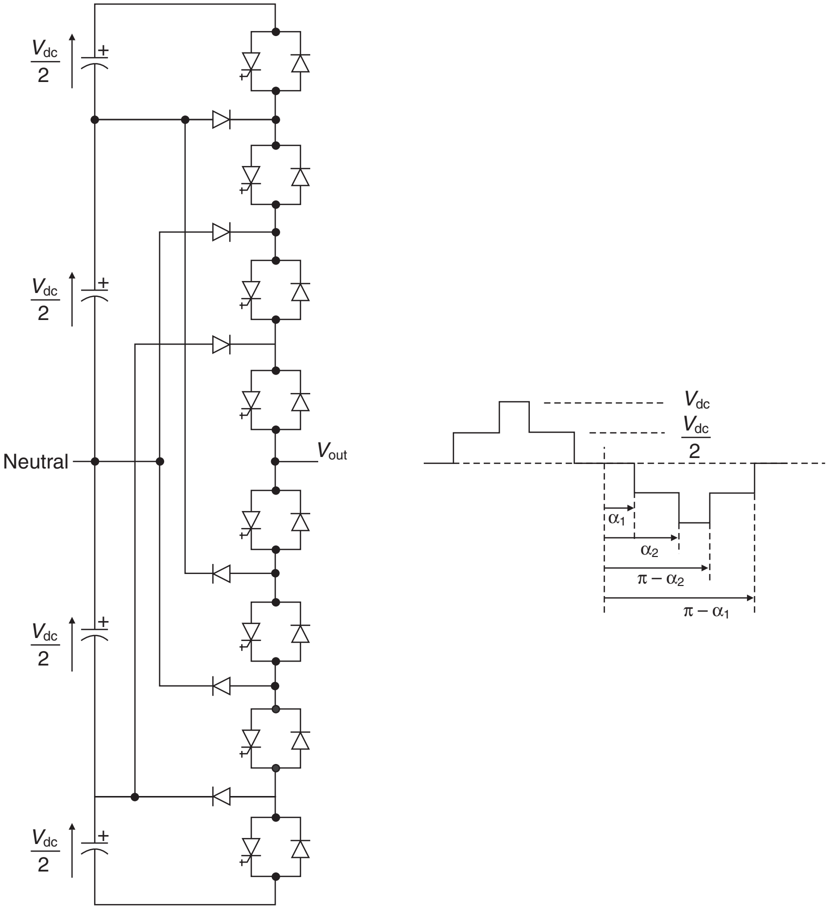


Figure 41.18 Single-phase multi-level voltage-sourced converter

will cancel out and so prevent any flow of the corresponding currents.

With four transformers having relative phase displacements of 15°, 24-pulse operation can be obtained; with eight transformers having 7.5° phase shift, 48-pulse operation results. It is usual to employ a simple transformer to provide the step-down duty from the high voltage system to an appropriate intermediate voltage; several different designs of phase-shifting transformers are then required to supply the individual converter bridges.

A simplified arrangement is illustrated in *Figure 41.20*. Two star–star and two star–delta transformers are connected with their primary windings in series and each secondary winding is connected to a 6-pulse bridge converter. All four of the converters utilise the same d.c. voltage source but they are operated effectively with relative phase-displacements of 15° to produce a quasi 24-pulse configuration. The 12-pulse harmonics which are characteristic for each pair of converters are not perfectly cancelled, but the residual magnitude is acceptably small. For practical applications of this kind of STATCOM, the Graetz bridges must use several GTOs in series in each leg of the bridge to obtain an adequate rating. The GTOs in each leg must be arranged to turn on and off at precisely the same instant (within microseconds) to ensure good voltage sharing between the individual GTOs.

41.5.6.5 Pulse width modulation

Pulse width modulation (PWM) is an alternative method of harmonic control. The power electronic switches are repetitively turned on and blocked several times during each half cycle. The sequential switching instants are selected in a co-ordinated manner, to satisfy simultaneous requirements, i.e. to develop the desired fundamental voltage and to eliminate selected low order harmonics. *Figure 41.21* illustrates how two poles of a converter (such as that shown in *Figure 41.17*) can be controlled, each with 5 on/off actions per cycle, to eliminate both 5th and 7th harmonics together. Extra, correctly timed, switchings can eliminate additional higher harmonic components. IGBTs and IGCTs require much lower switching energy than GTOs and are better suited to applications that use PWM techniques. Multi-module arrangements are necessary to obtain high ratings.

41.5.6.6 Multi-level converters

By combining sets of GTOs and their diodes in series, additional steps can be added to the output voltage waveform. Co-ordinated control of the switching instants of the series levels increases the controllability of the fundamental and harmonic content of the waveform. One important form of multi-level converter is the ‘chain circuit’ in which several

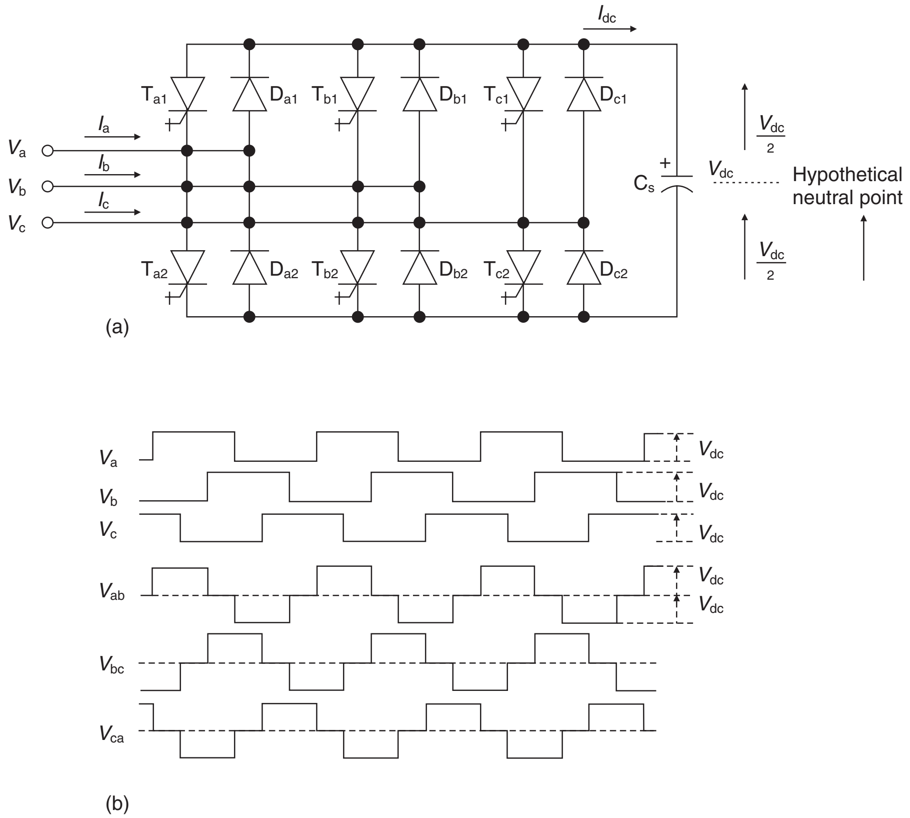


Figure 41.19 Simple three-phase voltage-sourced converter circuit and voltage waveforms

converter bridges, each with its own source capacitor, are connected in series. By appropriate switching of the GTO and diode pairs, the d.c. sources are connected into the circuit with either positive or negative polarity or are bypassed to give zero voltage contribution.

This is illustrated in *Figure 41.22* that shows that a single-phase chain with only three links can produce seven output levels. Each link in the chain produces its own rectangular block of voltage, with an amplitude equal to its own source voltage (normally all these source voltages are controlled to have the same value). Each voltage block therefore makes a contribution to the total fundamental voltage and to the total spectrum of harmonics. The more links there are in the chain, the more closely the overall waveshape approximates to a sinusoid. This stepped waveform appears broadly similar to the current waveform of a conventional multi-pulse a.c. to d.c. rectifier. However, the rectifier waveform consists of a series of pulses of varying height but equi-distant in time, whereas for the chain circuit, the pulses are of equal height and the interval between switching instants varies. In principle there is no limit to the number of converter bridges (or ‘links’) that can be connected in series in a chain to generate, directly, a high value of alternating voltage with a small content of harmonic distortion.

41.5.6.7 DC source voltage ripple

The d.c. source should preferably be strong enough for its voltage to remain effectively constant at the chosen level, under steady state conditions. In practice, especially for capacitor voltage sources, a compromise is necessary to allow the capacitor to charge and discharge to some extent between each switching operation, i.e. a constant average voltage is maintained but with a super-imposed ripple voltage of a few per cent, *Figure 41.23*. This ripple must be taken into account in selecting switching instants and in evaluating the overall harmonic behaviour of the converter system.

41.6 Shunt compensation

41.6.1 General

The load carried by a power system is continuously changing. Major variations in total load occur quite slowly, over a period of tens of minutes or longer in a daily cycle see *Figure 41.24*, which itself changes with a weekly and yearly variation. Normally any very rapid changes of load are only a small proportion of the total and do not affect the overall pattern.

For a system operating under normal conditions the overall var compensation requirements also change relatively slowly. In a well-proportioned system the var flows are not sensitive to small load changes and it is generally possible to use fixed shunt reactors or capacitor banks to balance the vars for average load conditions. Often, however, it is advantageous to arrange that some or all of these shunt compensation elements are switchable, so as to give better control and to reduce unnecessary losses. These discretely

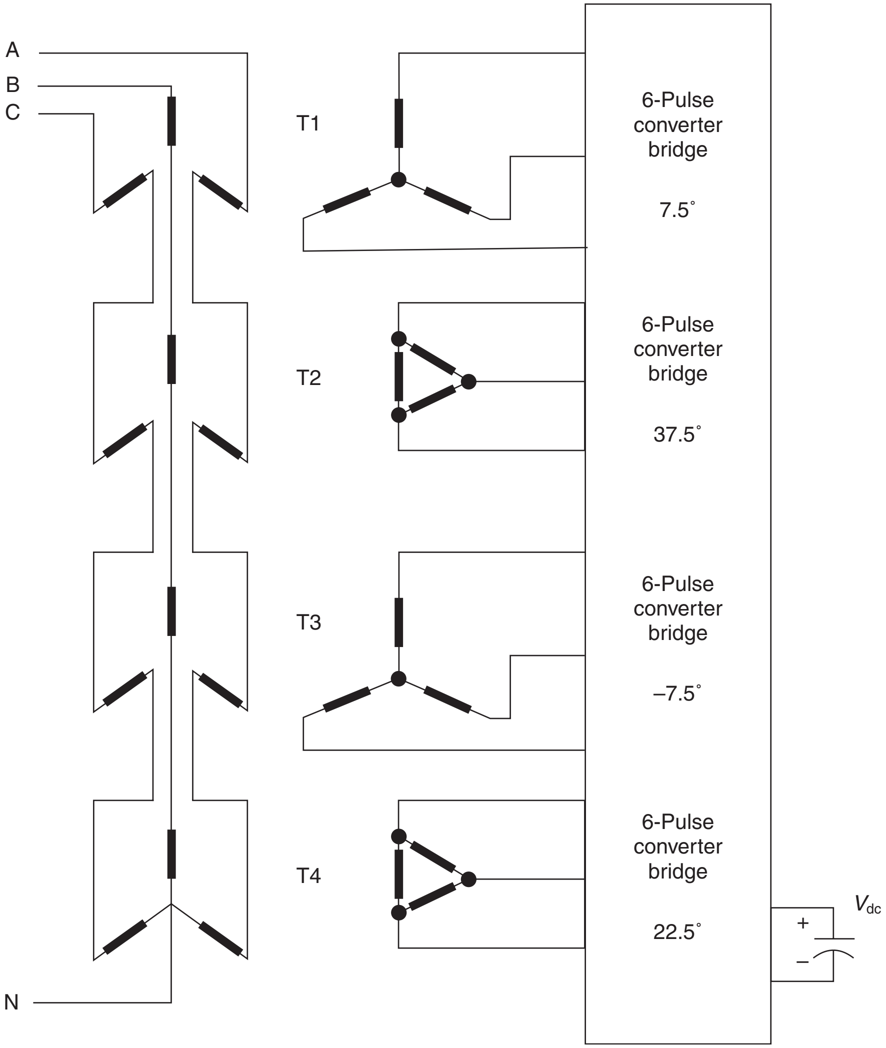

A
B
C
N
T1
T2
T3
T4
6-Pulse converter bridge
7.5°
6-Pulse converter bridge
37.5°
6-Pulse converter bridge
–7.5°
6-Pulse converter bridge
22.5°
+
–
V_{dc}

Figure 41.20 Multi-pulse operation using phase displacement transformers

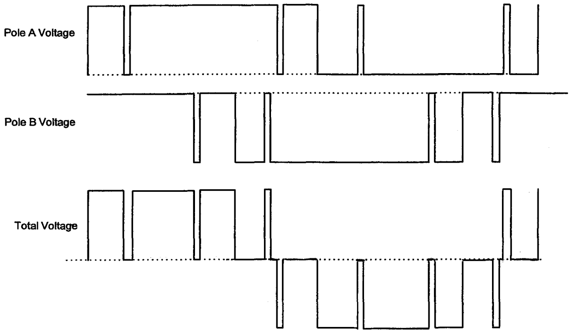


Figure 41.21 Pulse-width-modulated (PWM) converter voltage waveforms

variable devices are the simplest and the most common var compensators.

Many power systems now take advantage of the benefits offered by continuously variable var compensators. These can be used for particular categories of application such as:

(1) the compensation of rapidly varying loads that could otherwise cause unacceptable voltage disturbances on the power network;
(2) providing emergency voltage support in response to a circuit fault that increases the load current through each of the transmission lines that remains in service;
(3) giving continuous voltage control at a weak point of a network, where voltage variations due to normal load changes could otherwise become excessive and pose a threat of voltage instability;
(4) giving almost instantaneous overvoltage control in the event of a major load rejection on a long-distance transmission system;

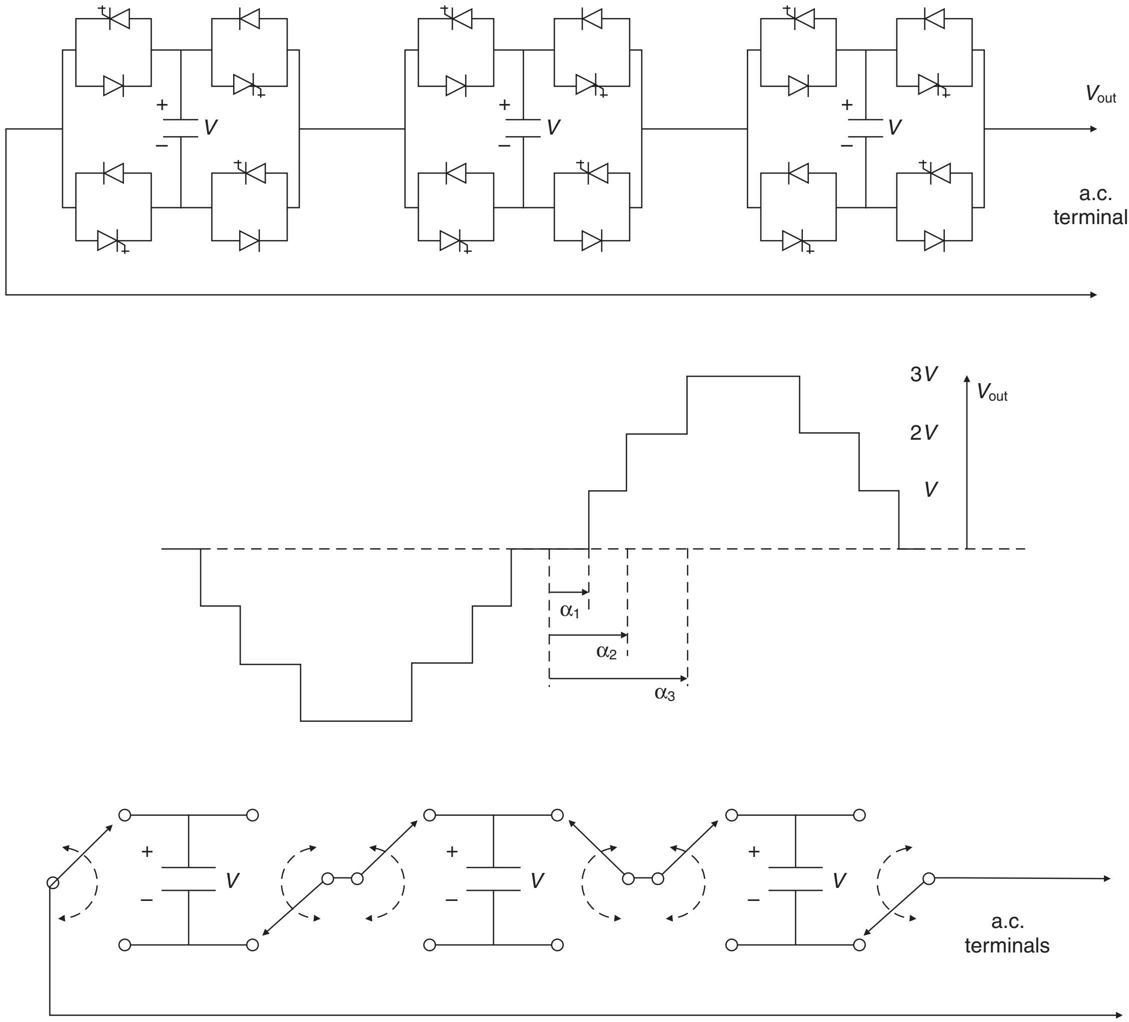


Figure 41.22 Single-phase chain circuit type of multi-level converter

(5) giving continuous fast control where transient stability is important, particularly on long-distance transmission; and

(6) on a tie line between two systems, damping out load swings which otherwise could lead to dynamic instability.

These different applications call for different characteristics of the variable var compensator. The types of compensator are reviewed in this section, with particular reference to their operating characteristics and applications, a comparative summary of which is given in *Table 41.1*.

41.6.2 Mechanically switched reactors and capacitors

Mechanically switched reactors and capacitors are used at all voltage levels, from the highest voltage transmission systems down to the correction of power factor on a works distribution system. Both technical and economic factors influence the decision to install switched reactive compensation. In many cases, operational considerations would permit the units to be permanently connected, but at the expense of continuous losses. The losses can be reduced by the installation of a switch suitable for fairly frequent operation, but this entails additional capital and maintenance costs.

If the compensation unit is permanently connected to other equipment (e.g. to a load or to a line), that equipment's protective circuit-breaker will also serve for clearing faults in the compensation plant. In assessing the economics of var balance it is desirable to know for what proportion of the time a unit could be switched out, but this data is often difficult to obtain from the existing loading information and even more difficult to predict for the future.

41.6.2.1 Shunt reactors

In transmission systems with very long lines, shunt reactors are often installed at the ends of each line section to control its voltage under energisation and light-load conditions. During times of heavy loading, when the line var gain due to the shunt capacitance is offset by its series var loss, it becomes desirable to switch these reactors out. This helps both to increase transmission capability and to minimise system losses.

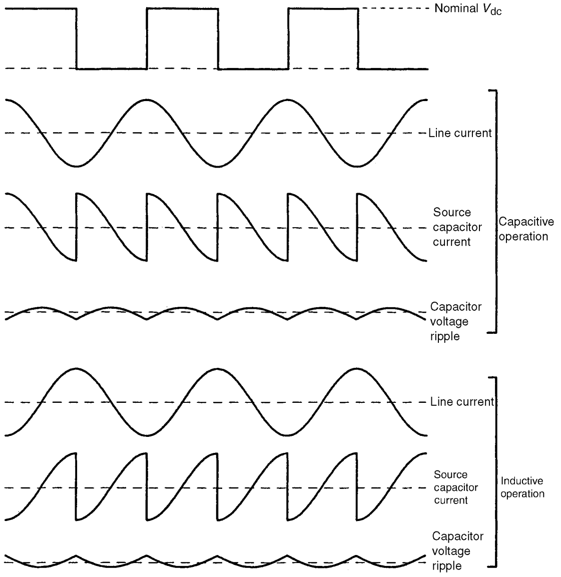


Figure 41.23 Converter output voltage and current waveforms and d.c. capacitor current and voltage during var generation and absorption

However, this requires transmission-voltage switches or circuit-breakers, which are expensive. Moreover, in the event of a large loss of load it would leave the system with a large excess of vars, leading possibly to excessively high voltage levels which could be unacceptable even for the short time necessary to reconnect the shunt reactors. As described later, a combination of fixed and continuously variable reactors may be a beneficial solution in these circumstances.

Where transmission or distribution systems include substantial lengths of cables, particularly at higher voltages, shunt reactors are sometimes installed at bulk supply points to compensate for the excess of vars under light-load conditions. To reduce the cost of losses inherent in all compensation devices, these are normally switched off as the load builds up.

41.6.2.2 Shunt capacitors

A very common and important application for shunt capacitor banks is in power factor correction (pfc) for industrial and commercial consumers. Here a straightforward economic evaluation is used to decide on the optimum reduction of var demand. The pfc capacitor banks usually consist of several separate units, or groups of units, each with its own contactor or switch. The load demand and power factor are, effectively, continuously measured and capacitor units are switched in and out as necessary to achieve the desired overall power factor for each successive tariff period, which may be only 15 or 30 min, for example.

Capacitor units have a very small inherent inductance and when units are switched back-to-back, without regard for the stored charges that each possesses at the instant of switching, there is a very large transient current flow between the units as the voltages are equalised. It is usual to limit the amplitude of the transient current by adding very small inductors into the circuit of each capacitor unit. These inductors may reduce the natural frequency of the capacitor circuits to perhaps 1 or 2 kHz. However, the installation of a pfc bank may result in a resonance between its capacitance and the system reactance at a prevalent low order harmonic, e.g. fifth or seventh; the switching inductors will not have a significant effect at such frequencies. Resonances of this kind can cause possible overvoltages or excessive currents. The capacitor bank may (because of its very low impedance at harmonic frequencies) act as a sink for harmonic currents produced by pre-existing voltage distortion in the network, and so become overloaded by them unless suitable series reactors are included. Before installing power factor correction banks and their associated series

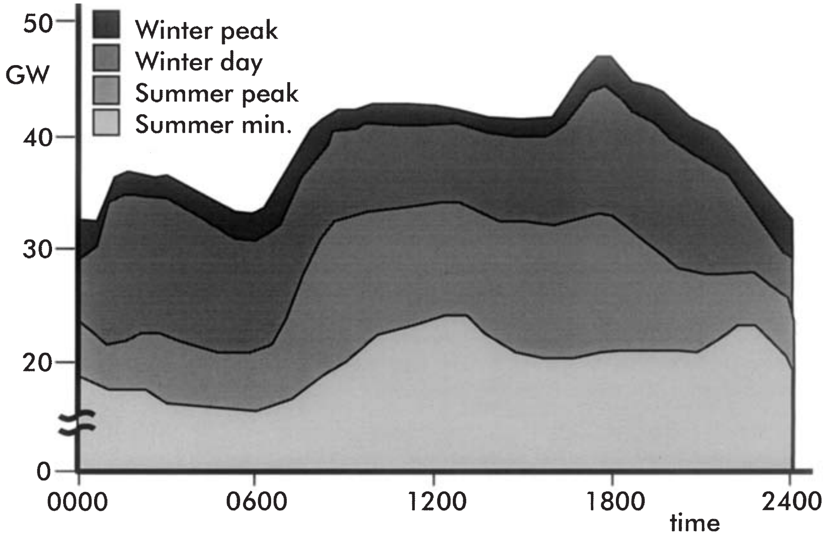


Figure 41.24

Table 41.1 Comparison of the characteristics of different var compensators

Item	*A*	*B*	*C*	*D*	*E*
Type of compensator	*Synchronous*	*Harmonic-compensated self-saturated reactor/ fixed capacitor*	*Thyristor-controlled reactor/fixed capacitor*	*Thyristor-switched capacitor/ thyristor-controlled reactor*	*STATCOM*
Features					
Steady-state characteristic	Automatic voltage regulator (AVR) easily adjustable	Restricted adjustment possible on site	Controller easily adjustable	Controller easily adjustable	Controller easily adjustable
Harmonic content of current	Negligible	Internally compensated in balanced three-phase system	Six-pulse design usually requires filters, particularly if reactance is <100% on rating	TSC, negligible TCR, smaller than in C	Low order harmonics usually small; high order harmonics can be large
Fault infeed	Approximately 4–6 times the rating	Nil	Nil	Nil	Nil
Overload capability	Gives both generation and absorption up to approximately two times normal during swings	Generation overload limited to capacitor bank rating unless special star–delta switching used			Generation and absorption overload only available if deliberately included
		Inherent very high absorption (e.g. three times normal)	Absorption limited by full conduction of thyristors		
Losses	High at full output; low at zero var output (float condition)	Low at full generation; high at zero var output (float condition) and at full absorption		High at full output; low at zero var output (float condition)	
Maintenance requirements	As for any large rotating plant	Small; as for conventional outdoor equipment	Moderate for special electronic (indoor) equipment when properly designed small for conventional outdoor equipment		
Approximate response time in typical complete system	0.2 s	0–0.05 s	0.02–0.06 s	0.02–0.06 s	0.01–0.04 s
Response to rapidly fluctuating loads	Relatively slow	Inherently fast	Normally less rapid than B, but some control improvement possible	Slower than C	Faster than C
Voltage control under load rejection (i.e. potential over-voltage condition)	Poor to fair, liable to self-excitation and loss of synchronism below minimum excitation limit	Very good with inherent fast response	Fair, controlled absorption limited by full conduction of thyristors	Poor, limited by the low rating of the reactor of the TCR	Limited by the installed rating of voltage-sourced converter (VSC)

cont'd

Table 41.1 (*continued*)

Item	*A*	*B*	*C*	*D*	*E*
Type of compensator	*Synchronous*	*Harmonic-compensated self-saturated reactor/fixed capacitor*	*Thyristor-controlled reactor/fixed capacitor*	*Thyristor-switched capacitor/thyristor-controlled reactor*	*STATCOM*
Voltage control under line outage (i.e. potential under-voltage condition)	Good, but relatively less rapid	May require switched capacitors to support voltage	May require switched capacitors to support voltage	May require switched capacitors to support voltage	Limited by the installed rating of voltage-sourced converter (VSC)
Behaviour following system fault	Machine inertia could cause swinging and loss of synchronism	Inherent response as a constant voltage device	Auxiliary controls could be used to damp load swings	Auxiliary controls could be used to damp load swings	Auxiliary controls could be used to damp load swings

reactors, measurements or estimates of pre-existing harmonic voltages in the network should be made and the effects of the capacitors on the remainder of the system should be studied.

In some networks it is the practice to install shunt capacitor banks at critical points of distribution systems with long overhead feeders, such as those supplying rural areas; the capacitors at the substations are usually switchable, whereas those part-way down the feeders are often pole-mounted and permanently connected. In the UK, owing partly to tariff incentives, load power factors tend to be high and there is no general application of this type of shunt capacitor compensation.

In the UK, power is delivered to distribution companies at bulk supply points, usually at 132 kV. Switched shunt capacitors are connected at some of these bulk supply points and are often used to compensate for var imbalance on the transmission network in addition to the var demand of the l.v. connected loads. Bank ratings are generally 60 Mvar and several are installed at some substations. In the last few years there has been an interesting development, resulting in the connection of switched capacitor banks of 150 Mvar and 225 Mvar directly on the 275 kV and 400 kV systems. In order to eliminate any possibility of adverse consequences due to amplification of pre-existing harmonic distortion, these banks are arranged as selectively damped harmonic filters, with very low losses at the rated power frequency but strong damping of any third harmonic currents.

41.6.2.3 Technical aspects of switching reactors and capacitors

Any switchgear installed to control a shunt capacitor bank or a reactor, be it circuit-breaker or load switch, must be carefully chosen, as the switching duties, even though the load currents are normally small, are not easy. The number of operations is often high.

With reactor switching, de-energisation may present difficulties if the current does not decrease smoothly to zero as the switch contacts separate and the arc is quenched. If there is current chopping and the current collapses to zero almost instantaneously, even from a very small current, the high rate of change of current can generate extremely high voltages across the reactor. In order to avoid damage to the reactor, the circuit-breaker and other equipment, it is necessary to ensure either that overvoltages cannot be generated, or that they are limited to levels which do not surpass the insulation withstand levels of the system, for example, by means of surge arresters.

When capacitors are energised, high transient inrush currents are produced which may severely stress both the switch and the capacitor. Circuit-breakers, or switches, that are designed to provide point-on-wave switching for each phase are able to reduce these inrush transients to very much lower magnitudes.

De-energisation is also a severe duty for any switch because the interruption process leaves the capacitor charged to the peak value of the system voltage. Within a half-cycle of the instant of interruption, the polarity of the system voltage will have reversed and the voltage across the opening contacts of the switch will be twice the normal crest voltage of the system. If the insulating medium does not recover sufficient dielectric strength, the arc re-strikes and there is a short pulse of current, which leaves an even higher level of voltage trapped on the capacitor (with reversed polarity). In the worst case, this process continues until there is a failure of major insulation on the capacitor or other equipment. Special tests are conducted on switchgear intended for capacitor switching to validate its suitability for the purpose. In some cases special discharge circuits are connected into the capacitor circuit to ensure that the capacitor voltage decays very rapidly after the current is interrupted and that, even if arc re-ignition occurs, there is no magnification of voltage.

41.6.3 Synchronous compensators

The synchronous compensator has a more positive action than most types of compensator in supporting a weak system, as it has both a generated voltage and inertia that enable it to act transiently as if it were coupled to a prime mover. The synchronous compensator can maintain full rated current independent of its terminal voltage and may be arranged to increase its var output temporarily to provide transient assistance to the system. This has been of value in certain applications; for example, where power is supplied by h.v.d.c. to a system having minimal generation, the receiving system has usually been strengthened by a synchronous compensator.

When synchronous compensators have been used for dynamic compensation of high voltage systems it has been necessary either to provide a dedicated transformer or, in some cases, to utilise a tertiary winding on a transformer which links two voltage levels within a system. With tertiary connection, the inter-winding reactance distribution of the transformer determines the high voltage bus-bar to which the compensator is most closely coupled.

The main limitations of synchronous compensators compared with static alternatives are a notably slower speed of response, typically 0.2 s, higher losses, inertia that can cause underdamped oscillations or loss of stability in some transient conditions, higher maintenance requirements (and lower availability) and higher capital costs. Because of their inherent limitations, synchronous compensators have now been almost completely superseded by static controllers.

41.6.4 Static var compensators

41.6.4.1 Saturated reactors

Saturated reactors are normally used in conjunction with shunt capacitor banks to provide var generation as well as absorption. Some of the shunt capacitors are configured as harmonic filters that are permanently connected in parallel with the SR; other shunt capacitor banks may be permanently connected or switched, as appropriate to the duty of the SVC.

A significant feature of the SR type of SVC is that, because of its transformer-like construction, the SR has very low maintenance/operational requirements and also a high short-time-rated overload capability of up to 4–5 times rated load for 1 second, or longer if necessary. This can be of use in networks with long transmission lines, where the Ferranti effect following a loss of load can produce excessively high temporary overvoltages. The SR, by swinging rapidly into a high var absorption mode, can balance the var production of the line and so reduce the inherent overvoltage.

The SR is normally operated as a fixed-voltage device, though adjustment of its saturation voltage over a range of a few per cent is possible in principle. If it is to be connected to a system whose voltage level must be adjustable, then a regulating transformer is normally connected between the SR and the system bus-bars, *Figure 41.25*. Where a step-up transformer is necessary to match the SR voltage (normally 6–69 kV) to the system voltage, this can include an on-load tap-changer to provide any regulation that is needed.

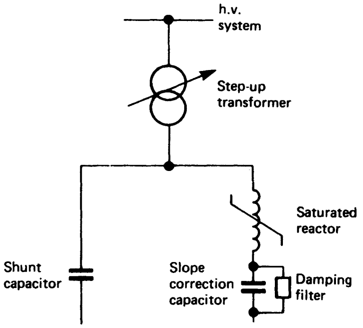


Figure 41.25 Saturated reactor compensator

The *V–I* slope typically required at the terminals of a SVC for transmission applications is around 3–5%, whereas the total slope obtained from a SR plus step-up transformer may be 15–25%. To compensate for this difference, a slope-correcting capacitor is connected in series with the SR as shown in *Figure 41.25*. To avoid the possibility of sub-harmonic oscillations after energisation, a damping filter is fitted across this capacitor.

The current drawn by a SR depends directly on the time-integral of the voltage applied to its terminals and its speed of response is therefore extremely fast, responding to voltage changes within the same half-cycle. Owing to the polyphase construction and the interconnection of its windings, a SR operates in a manner different from that in a linear shunt reactor. The latter absorbs the capacitive energy of the system as electromagnetic energy and returns it all every half-cycle, phase by phase. In contrast, the treble-tripler reactor, by virtue of its nine-active cores and interconnected windings, transfers about 90% of its energy from phase to phase as the magnetic flux commutates from one core to the next every 1.1 ms at 50 Hz (2 × ~~9 = 18~~ transitions per period). The inclusion of the slope-correcting capacitor slightly slows the effective overall response and a slope-corrected treble-tripler reactor SVC responds effectively to a step change within one to two cycles, dependent on the system impedance and on the rating of its sub-harmonic damping circuit. An alternative method of achieving the slope-correction effect without time delays has been used for lamp flicker compensation in *Figure 41.46*(*b*).

41.6.4.2 Thyristor-controlled reactors

The simplest design of TCR, *Figure 41.26*(*a*), uses three single-phase valves, usually connected in delta, giving a six-pulse unit: hence, it produces appreciable levels of fifth and seventh harmonic currents and in most cases will require the use of harmonic filters. Shunt capacitors are needed to provide var generation and, for large SVCs, it is usual for some of these banks to be thyristor switched. Although for some ratings it may be possible to connect the TCR directly to the power system bus-bar, for most applications an interconnecting transformer is required to match the optimised valve voltage to the system voltage; on-load tap-changers are very rarely included.

A few installations have used a step-up transformer with two phase-displaced secondary windings connected star and delta, each with a delta-connected set of thyristor valves and reactors, *Figure 41.26*(*b*). This 12-pulse design gives better harmonic compensation, the principal harmonics being now the 11th and 13th, and simpler filtering can be used.

The control system of a TCR can produce any desired voltage–current characteristic derived from various input signals. A basic characteristic for a transmission compensator having a closed-loop voltage control, with current compounding, is shown in *Figure 41.27*. When the valve reaches maximum conduction, the characteristic follows the slope set by the impedance value of the linear reactors plus coupling transformer (typically 0.7–1.0 p.u. on the basis of full load rating) up to the point of the thyristor current thermal limit, when by phasing back the firing angle the current is held constant up to the maximum voltage limit of the thyristors. The actual relative positions of these constant-current and continuous-conduction parts of the characteristic may be changed to suit the design requirements.

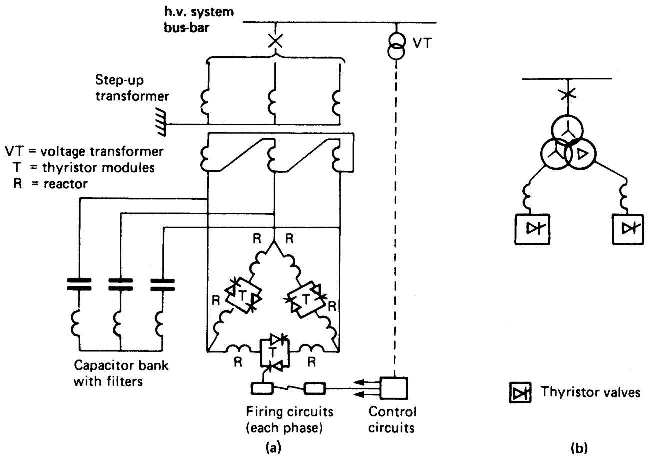


Figure 41.26 (a) Six-pulse TCR; (b) twelve-pulse TCR

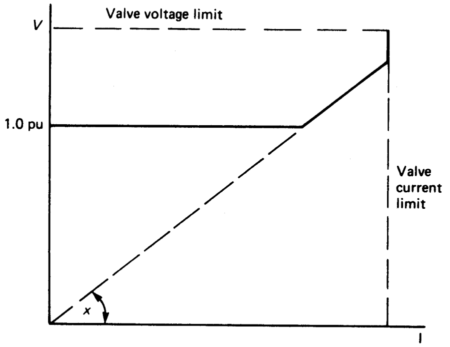


Figure 41.27 Voltage–current characteristics for a thyristor-controlled reactor

A speed of response to small disturbances of 1 to 3 periods is achievable with this type of control system. A slightly faster response may be obtained by using an open-loop current control; this can be useful for certain industrial applications where the highest speed is required and some sacrifice of control accuracy can be tolerated.

41.6.4.3 Thyristor-switched capacitors

Thyristors can be used to switch on a large bank of capacitors faster than is possible with a mechanical switch; therefore a potential application of a TSC is for fast boosting of a transmission-line voltage to maintain system stability. A TSC, perhaps having several steps, as in *Figure 41.15*, could also be used on its own to provide swing damping in a transmission system.

TSCs are commonly used in conjunction with TCRs to give much reduced losses at zero var output (float condition) compared with the losses of a scheme using a fixed capacitor and a larger TCR; they are also used to provide an increased operating range in the var generation region where speed of capacitor energisation is of importance and a high frequency of operation is required. In these applications the TSC provides the coarse steps, seldom exceeding four, whereas the TCR gives a fine continuous control in between.

In order to obtain a fine control of output using only TSCs, there must be a series of small steps, each with its own capacitor bank and thyristor switch. The number of steps and hence the fineness of control are dictated by economics in the cost of separate valves for each step and the complexity of controls. TSCs have occasionally been used for power-factor control of loads with frequently varying var demand; here the capacitor banks and thyristor valves were at a low voltage and a large number were used in parallel to approximate to a stepless control.

In the last decade, it has been found necessary to provide increased capacitive support for the National Grid in the UK. Because network requirements can change substantially over periods of a very few years, it is also desirable to be able to relocate equipment quickly to an alternative location. It is rarely possible to arrange for a high voltage connection to be quickly available, but transformers with 13 kV, 60 MVA tertiary windings are commonly available in Grid substations. The h.v. system may be either 400 kV or 275 kV and the l.v. may be 275 kV, 132 kV or 66 kV. Relocatable SVCs (RSVC) were specified, for connection to tertiary windings, and these only needed to provide capacitive Mvar; stepwise control of the output was permissible.

The specification for the RSVCs lent itself to an arrangement consisting of three TSCs, with binary ratings of 8.6, 17.1 and 34.3 Mvar, which are controlled in a co-ordinated way to provide the 60 Mvar output in seven steps of 8.6 Mvar, *Figure 41.28*. Twelve such RSVCs have been installed, in seven substations.

The ability of TSCs to reduce voltage fluctuations is very limited compared with that of other types of SVC and they cannot be controlled fast enough to be effective in reducing lamp flicker caused by arc furnaces.

41.6.5 STATCOM

41.6.5.1 Some practical implications

It would clearly be inappropriate to connect a voltage-sourced converter directly to the supply system, which will

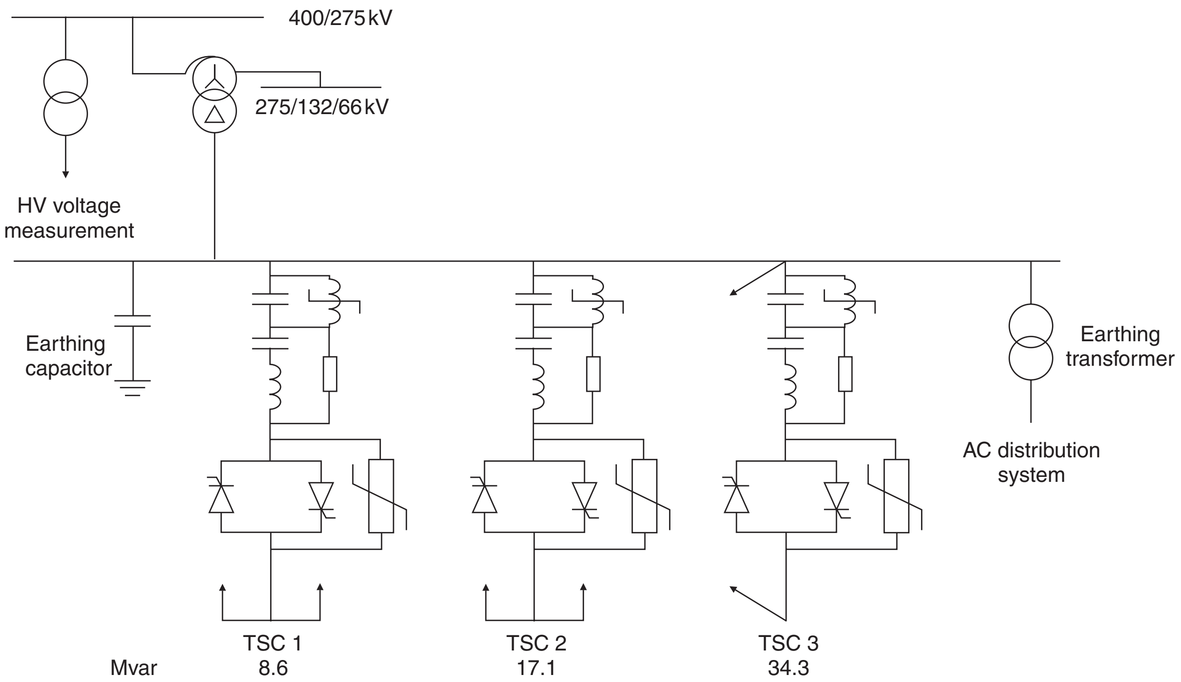


Figure 41.28 Relocatable SVC (RSVC) arrangement used for the National Grid in the UK

generally have a very much higher short-circuit power than the rating of the converter. As shown in *Figure 41.16*, there is normally coupling reactance between the system bus-bar and the converter terminals. When the converter is connected to the system via a transformer, this may provide sufficient inherent coupling reactance for satisfactory operation of the STATCOM.

In some applications where there is a stringent harmonic performance requirement, it may be necessary to include shunt capacitors or harmonic filters on the converter busbar, in which case buffer reactors (separate from the step-down transformer) may be needed to limit the flow of harmonic currents from the converter into the capacitors.

As well as limiting harmonic currents, the coupling reactance limits fault currents into the STATCOM; it also 'softens' the transient current response from the STATCOM when there is a fault on the supply system which collapses the voltage, or when there is a large magnitude dynamic overvoltage. Inevitably, the presence of the coupling reactance requires the voltage generated by the converter to be higher than that of the system when the STATCOM is acting as a capacitor and to be lower than the system voltage when the STATCOM is acting as an inductor. The vars absorbed by the coupling reactance (and the supply transformers) must be taken into account in the design and rating of the converters.

41.6.5.2 STATCOM operating characteristics

The 'natural' voltage–current characteristic at the terminals of a STATCOM, *Figure 41.29*(*a*), is entirely dependent on the converter source voltage V_e and on the coupling reactance, X_e, see *Figure 41.16*. In general the coupling reactance has a typical value between 10% and 20%, i.e. the voltage drop (or rise) across it is about 10 to 20% of the nominal system voltage at the rated current of the STATCOM. The continuous current rating of a GTO, IGBT or IGCT is almost independent of whether the current lags or leads the voltage. These devices usually also possess a short-time, or transient, overcurrent rating, which may exceed the safe turn-off (or turn-on) current for the STATCOM valve components. If, by accident or design, the safe turn-off current is exceeded, then turn-off signals must be prevented from being issued until the current returns to below the safe switching level.

The nominal maximum steady state voltage for a supply system is 1.1 p.u. of its nominal value and this presents no difficulties for the design and rating of a STATCOM. However, a STATCOM must also withstand dynamic overvoltages and transient overvoltages up to the protective level provided by the STATCOM's surge arresters. During transient conditions in which the instantaneous applied voltage exceeds the d.c. source voltage, the diodes of the STATCOM will allow current to flow, and this current will charge the d.c. source capacitors to a higher voltage.

For most practical applications, a SVC needs to operate with a slope reactance typically of between 2% and 5%, which is much lower than the value of the coupling reactance. Because the coupling reactance is fixed, the converter source voltage must therefore be changed as shown in *Figure 41.29*(*b*), raising it to obtain the desired loading (capacitive) current conditions or reducing if for lagging (inductive) conditions. This can be done very quickly, initially by changing the switching pattern and then followed, if appropriate, by changing the magnitude of the d.c. source voltage. The desired control characteristic is thus achieved by changing the source voltage in precisely the same way as for a synchronous compensator, but very much faster.

Both target voltage and slope reactance can be set in the STATCOM control. A set of voltage–current characteristics, for a range of target voltage settings with constant slope, is shown in *Figure 41.29*(*c*). For comparison, the *V–I*

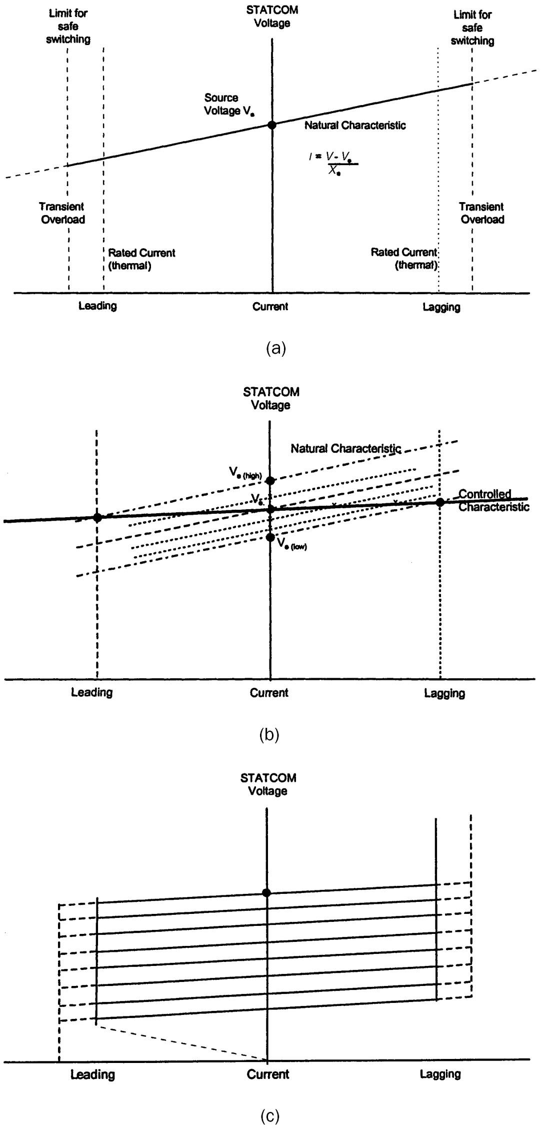

Limit for safe switching
STATCOM Voltage
Limit for safe switching
Source Voltage V_e
Natural Characteristic
$I = \frac{V - V_e}{X_e}$
Transient Overload
Transient Overload
Rated Current (thermal)
Rated Current (thermal)
Leading
Current
Lagging
(a)
STATCOM Voltage
Natural Characteristic
$V_{e\,(high)}$
V_e
Controlled Characteristic
$V_{e\,(low)}$
Leading
Current
Lagging
(b)
STATCOM Voltage
Leading
Current
Lagging
(c)

Figure 41.29 (a) Natural *V–I* characteristic of a STATCOM, (b) controlled *V–I* characteristic of a STATCOM; (c) family of *V–I* characteristics for a STATCOM

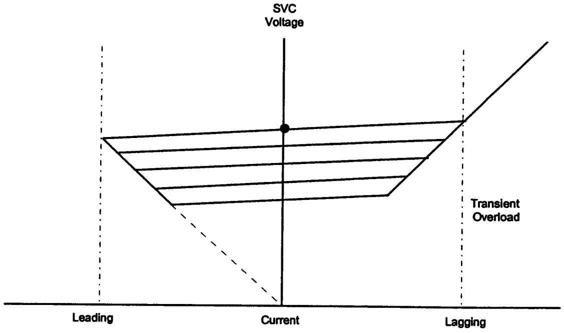

SVC Voltage
Transient Overload
Leading
Current
Lagging

Figure 41.30 Family of *V–I* characteristics for conventional SVC

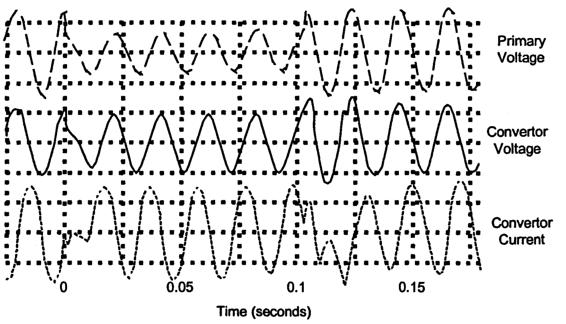


Figure 41.31 Response of a STATCOM to a depression in system voltage

characteristics of a conventional SVC are shown in *Figure 41.30*. At reduced voltage, the STATCOM can continue to be operated at rated leading (or lagging) current, with a constant transient overload current margin. These capabilities are available down to very low voltages. In contrast, the current limits for conventional SVCs are proportional to voltage. A STATCOM is better able to provide reactive/current support for a supply system whose voltage is severely depressed, whereas a conventional SVC can generally do more than a STATCOM to limit dynamic overvoltage.

41.6.5.3 Transient response

Because the operation of a STATCOM is based on the generation of a sinusoidal voltage, its response to transient disturbances is inherently good and extremely rapid. *Figure 41.31* illustrates how a STATCOM responds to a voltage disturbance. Prior to the voltage dip, the STATCOM is operating at about its rated lagging current. A dip of system voltage suddenly occurs, to about 50% of its steady state value. The STATCOM inherently responds to this disturbance by generating a capacitive current to support the system voltage but, even on the natural characteristic, there would be a capacitive overload current. To prevent this, the STATCOM control system detects the sudden change and reduces the target voltage to limit the STATCOM current to its rated capacitive value.

When the fault is cleared and the system voltage recovers to its pre-fault value, this will tend to cause an inductive overload current in the STATCOM. Again, the STATCOM control system is able to detect the change and adjust the target voltage appropriately. Although there is unavoidable transient distortion of the STATCOM current at each step change it can be seen that the changes from inductive to capacitive and capacitive to inductive current take place within a half cycle.

41.6.5.4 STATCOM losses

The forward voltage drop of GTO thyristors is greater than that of conventional thyristors because of the more complex system of semi-conducting junctions and the energy required for the turn-off duty. *Figure 41.32* shows the approximate variation of STATCOM losses (% of rated current) through the operating range from rated leading to rated lagging current.

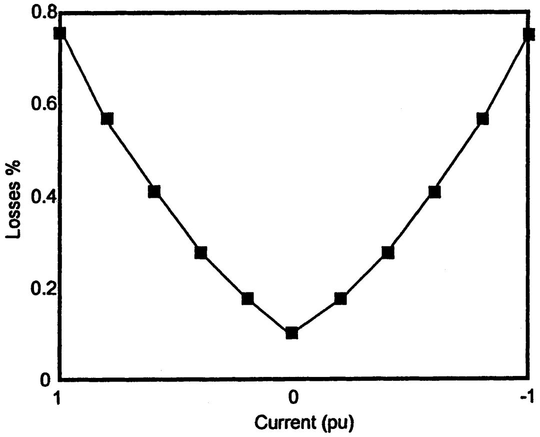


Figure 41.32 Typical loss curve for a STATCOM

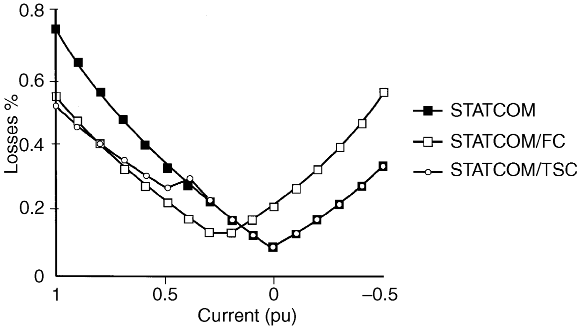


Figure 41.33 Typical loss curves for STATCOM applications

In many cases, the STATCOM output will need to be biased—generally towards the capacitive side for SVC applications. *Figure 41.33* shows the loss patterns for an output range +1.0 to −0.5 p.u. current.

If the STATCOM is rated for an output of ±1.0 p.u. current, but the upper half of the inductive range is not used, then the losses in the float condition (0 Mvar) and within the lagging range are quite low, but become quite high in the upper part of the capacitive range. These capacitive losses can be reduced by halving the rating of the STATCOM and combining it with a TSC of about 0.6 p.u. to reduce the losses (and generally also the overall costs).

An intermediate option is also illustrated with a STATCOM of ±0.75 rating, to cover the total dynamic range, biased by a fixed capacitor (or filter bank) of 0.25 p.u. output. This may give a better overall optimisation of cost and losses especially if the predominant range of operation of the compensator is in the typical range from about 0.1–0.6 p.u. capacitive current.

41.6.5.5 Other types of STATCOM source

In addition to voltage sources using batteries and capacitors, STATCOMs can be operated with an inductor, which provides a source of direct current rather than voltage. A three-phase, current-sourced converter then generates a set of three-phase output currents which, by appropriate switching action, lag or lead the system voltages. The basic output current is a square or block wave and harmonic reduction requires PWM, or multi-level or multi-pulse techniques, and/or harmonic filters. The output current magnitude may be controlled by diverting part of the source current through a bypass circuit. The energy in the current source can be sustained by drawing energy from the supply system or by using an external energy source. However the losses of a current-sourced converter tend to be higher than those of a voltage-sourced converter.

41.6.6 STATCOM applications

41.6.6.1 Transmission applications

In order to demonstrate the features of STATCOMs and to gain practical operating experience, several prototype STATCOM installations were put into service in transmission networks. In addition, several smaller STATCOM installations have been used (particularly in Japan) to reduce flicker caused by electric arc melting furnaces. However, many of the conventional applications of SVCs do not need to take advantage of many of the particular features of STATCOMs, such as equality of lagging and leading outputs, faster response time, possible active harmonic filtering capability and smaller site area.

Nevertheless, confidence in the STATCOM principle has grown sufficiently for several utilities to adopt them for normal commercial service. As an example, in 1996, the National Grid Company plc of England invited tenders for relocatable dynamic reactive compensation equipment for its 400 kV transmission network, capable of generating 0–225 Mvar at 0.95 p.u. system voltage. A particular requirement was the inclusion of a STATCOM of 150 Mvar range. The design adopted includes a ±75 Mvar STATCOM in conjunction with a 127 Mvar TSC and 23 Mvar fixed capacitor to provide a full controlled range of output +225 to −52 Mvar, *Figure 41.34*.

This STATCOM design is required to meet stringent harmonic emission levels and immunity to existing and future prospective harmonic levels. It uses multi-level converters in a chain circuit configuration. Typical voltage waveshapes are shown in *Figure 41.35*. The control system incorporates voltage control, reactive setpoint regulation, and a co-ordinating control for the STATCOM and the associated TSC. Provision is also made to include power oscillation damping control in the future, if it should be necessary.

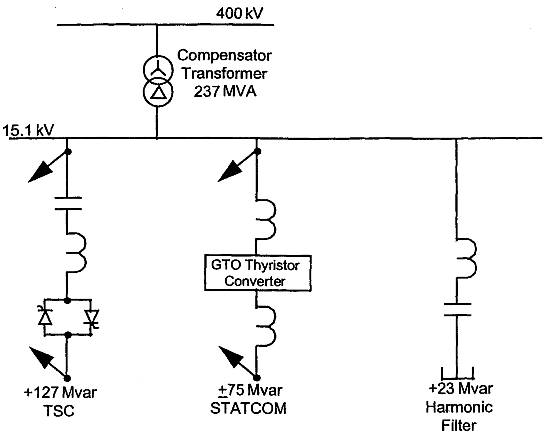


Figure 41.34 STATCOM arrangement for +225/−52 Mvar at 400 kV

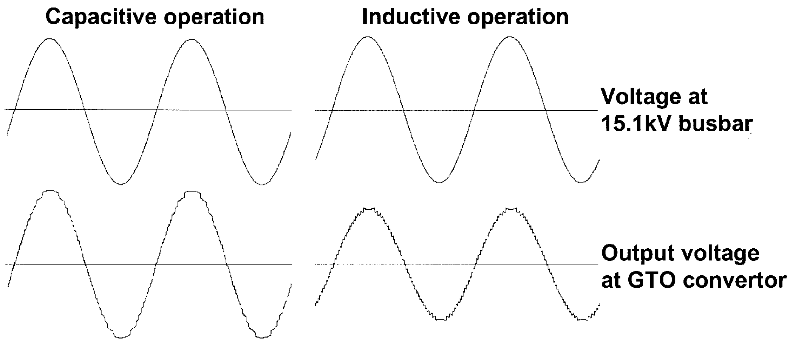


Figure 41.35 Voltage waveshapes of STATCOM at 15.1 kV bus-bar and at GTO converter

All the controls and power electronic equipment are housed in weatherproof, transportable GRP (glass reinforced plastic) cabins and the outdoor components are grouped together on frameworks to satisfy the requirement for easy relocation to another substation when this is required.

41.6.6.2 Specialised applications

Other applications of smaller STATCOMs, in service or under consideration, are for the reduction of lamp flicker due to arc furnaces, for voltage and var control for windfarms and for balancing of single-phase traction loads. These smaller units generally use PWM to obtain a satisfactory harmonic performance.

41.6.6.3 Energy storage applications

Some manufacturing processes require absolute continuity of supply to maintain product quality and/or safety, for example, float glass, paper, semi-conductor devices, and some chemical and nuclear processes. The cost of disruption may be so great that auxiliary or emergency power sources are economically justified. Uninterruptible Power Supplies (UPS) of moderate ratings are now very widely used to enable some processes to ride through brief voltage dips and interruptions. Where larger power ratings are required, STATCOMs, with enhanced energy storage or auxiliary power sources, offer a potential solution.

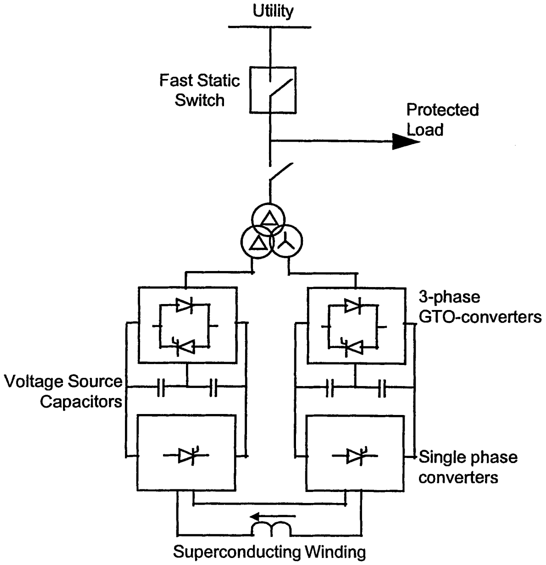


Figure 41.36 SMES circuit arrangement

Batteries have already been mentioned as voltage sources, but for stored energies of many megawatt seconds, present designs of battery are bulky and expensive. Very large capacitor banks might be used but are not efficient for bulk energy storage. Super-conducting magnetic energy storage sources (SMES), together with current-sourced converters have also been used in prototype format (*Figure 41.36*) and, in the future, might be very suitable for energy storage. A very attractive energy source of the future is the fuel cell.

41.7 Series compensation

Series compensation may consist of capacitor banks with a single fixed value, or adjustable in steps, capacitors with one or more stages that are continuously variable or it may be provided by a synchronous voltage source. All equipment that is intended for series connection must have suitable insulation to ground and be designed and rated for or protected against faults that can cause severe overcurrents through it.

41.7.1 Series capacitor compensation

Series capacitors have, for many years, provided the only practicable means of compensating the series inductance of transmission lines. Nevertheless, the application of series capacitors in large power systems requires caution at the system design stage, because series capacitors always introduce natural frequencies below the power frequency. Series capacitors resonate with the generator and line inductances at subharmonic frequencies and the subharmonic oscillations which follow any transient disturbances can lead to self-excitation of generators, to rotor hunting and to shaft oscillations. The possible danger of this phenomenon was illustrated in the early 1970s when the shaft of a large turbo-generator was twice damaged in service, owing to subsynchronous resonance (SSR) excited by series capacitors in the transmission system.

The SSR phenomenon is now well understood as being due to near coincidence of a shaft torsional resonance frequency and the complement of an electrical subharmonic frequency. It can be fully analysed using non-linear differential equations of the complete electrical–mechanical system. For studying such phenomena more economically, however, digital computer programs have been developed, based on either eigenvalue analysis or frequency-response analysis of the complete system of linearised machine and network equations, including any shaft torsional equations of the turbine prime movers.

41.7.1.1 Subsynchronous/subharmonic resonance damping

Methods of damping subsynchronous oscillations include the use of SVCs, damping filters, dynamic control of the series capacitor impedance and generator excitation control signals. The awareness of the possible danger of SSR, the improved methods of study and the available methods of damping such oscillations have enabled the application of series capacitors to continue successfully, albeit with some constraints.

Subharmonic oscillation phenomena may also occur when capacitors are in series with devices that have a non-linear reactive characteristic because they use saturable iron in the magnetic path, for example transformers, gapped-core 'linear' reactors and self-saturated reactors. These phenomena are amenable to non-linear or linearised analysis, but generally with only limited accuracy. Model studies using transient network analysers or simulators are employed in assessing individual applications involving transformers or linear reactors. Theoretical analysis of ferroresonance phenomena has not resulted in generally applicable conclusions. The performance of harmonic-compensated saturated reactors with slope-correcting series capacitors can be analysed satisfactorily and appropriate subharmonic damping filters can be designed accurately.

41.7.2 Controllable series compensation

41.7.2.1 Thyristor switched series capacitor (TSSC)

In a practical installation of series capacitors it is common for the overall bank to consist of several sections, each with its own protection, isolation and bypass switchgear. This subdivision provides an improvement in availability and more flexible adaptation to changing circumstances, including the possibility to avoid using capacitance values that would give undesirable resonance frequencies under particular system conditions.

It is impossible to obtain rapid or frequent bypassing and re-insertion of capacitor sections using conventional mechanically operated switchgear, but faster switching can be achieved using thyristor switches for one or more sections, *Figure 41.37*(*a*). There is a large discharge current when a capacitor is bypassed and a current limiting reactor (not shown) is needed to reduce the current to within the transient rating of the thyristor switch. Changing the capacitance value in discrete steps is suitable only for de-tuning SSR conditions but is not a satisfactory method of applying positive damping for SSR oscillations.

41.7.2.2 Thyristor controlled series capacitor (TCSC)

The TCSC is similar to the TSSC but the thyristor switch is converted into a TCR, *Figure 41.37*(*b*). By using point-on-wave switching of the reactor, two ranges of continuously variable impedance can be obtained, *Figure 41.37*(*c*). The controlled reactor has a significantly lower impedance than the capacitor so that when the thyristor valve is fully conducting, the overall impedance of the capacitor section becomes inductive; the current through the reactor is greater than the line current and the capacitor current is smaller. At the other extreme, when the thyristor valve is blocked, the line current flows only through the capacitor.

At some intermediate level of valve conduction, the effective inductive impedance of the reactor would be equal to the capacitor impedance and the combination would form a blocking circuit, with a very large voltage across it. Obviously, this condition must be avoided. Nevertheless, a range of satisfactory operating conditions can be obtained near to minimum conduction, giving the continuously variable capacitive effect that is desired from this circuit. There is another operating range near to continuous conduction, with a continuously variable inductance effect, but this is generally unimportant. *Figure 41.37*(*d*) illustrates the continuous and short-term capabilities of a single section of a TCSC.

Several TCSCs have now been installed in transmission systems. They can be controlled to apply positive damping to SSR conditions, if these arise.

41.7.2.3 Static synchronous series compensator (SSSC)

The advantages of series capacitors could be obtained and the major disadvantages could be avoided by using a FACTS controller to behave as a true negative inductance (i.e. with a reactance equal to $-j\omega L$ instead of $j\omega L$). This can be done by means of a voltage-sourced converter, as described in Section 41.5.6.2, and controlling it so that the voltage across it lags the current through it by 90°, *Figure 41.38*. This gives negative inductance, not only at the system frequency, but also at lower frequencies and for the more important of the low order harmonic frequencies.

If the d.c. source is only a voltage but not an energy source, the voltage across the controller will only deviate from 90° by an amount sufficient to supply the losses of the controller itself. The controller will cancel out part of the inductance of the line, reducing the voltage across the line and reducing the apparent surge impedance. The stability limit for the line will be increased and a greater power may be transmitted—or a longer line may be used at the same power level.

Other control functions could be used in response to a variety of system and network conditions and contingencies. The control could be used, for example, to prevent the line being overloaded by acting as a variable positive inductance. It could be used as a fixed voltage source to drive a circulating current around a particular part of the network in order to avoid overload in another circuit. With an energy source, other methods become possible for controlling power flow in the line. If the energy source is a shunt connected FACTS controller, the combination is called a universal power flow controller (UPFC) and is described in Section 41.8.2. The interline power flow controller (IPFC) uses another series connected FACTS controller, in a parallel line, as the energy source, as described below, *Figure 41.39*.

41.7.2.4 Interline power flow controller (IPFC)

It is not uncommon for two lines, that are of similar construction and rating, to follow different routes between major substations and so to have different lengths and different impedances. The sharing of current will be unequal. Furthermore, one or both circuits may link into additional load centres en route, with the result that one circuit becomes overloaded while the other has plenty of spare capacity. In such a situation an IPFC could be a far simpler and less disruptive solution than upgrading the overloaded circuit. It is very important in this application that the phase angle of the injected voltage should force sharing of the in-phase as well as the quadrature components of current in the circuits which are being equalised.

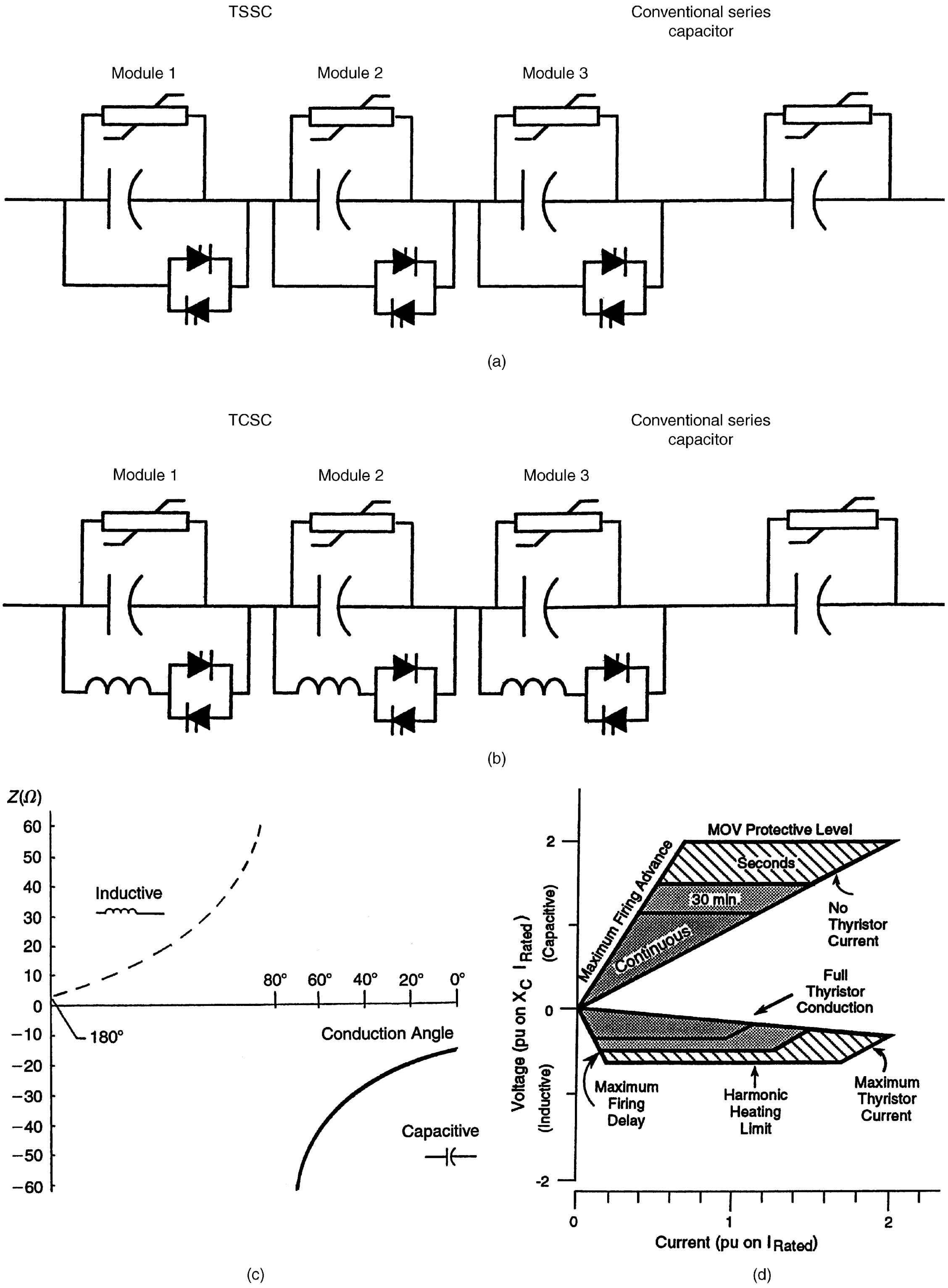

TSSC
Conventional series capacitor
Module 1
Module 2
Module 3
(a)
TCSC
Conventional series capacitor
Module 1
Module 2
Module 3
(b)
Z(Ω)
60
50
40
30
20
10
0
−10
−20
−30
−40
−50
−60
Inductive
180°
80°
60°
40°
20°
0°
Conduction Angle
Capacitive
(c)
MOV Protective Level
Maximum Firing Advance
Seconds
30 min.
Continuous
No Thyristor Current
Full Thyristor Conduction
Maximum Firing Delay
Harmonic Heating Limit
Maximum Thyristor Current
2
0
-2
Voltage (pu on X_C I_Rated)
(Capacitive)
(Inductive)
0
1
2
Current (pu on I_Rated)
(d)

Figure 41.37 (a) Thyristor switched series capacitor (TSSC); (b) Thyristor controlled series capacitor (TCSC), (c) Usable ranges of TCSC conduction; (d) Typical TCSC *V*–*I* capability characteristic

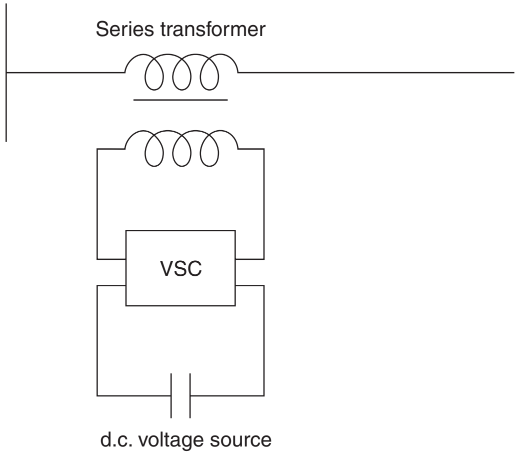


Figure 41.38 SSSC

41.7.3 Buffering of loads from system disturbances

There are many industries and industrial processes that are intolerant to voltage dips and even brief interruptions. If critical drive motors are shut down because the voltage falls below about 85% of rated voltage, or inverters suffer commutation failures because of voltage dips or distortion in one phase, the disruption can result in a serious loss of production. There may be an extended shut down of the plant while the rescue and re-start processes are undertaken. A device that sustains the voltage on the load bus-bar in the face of disturbances on the system has great potential value for protecting the sensitive loads.

41.7.3.1 Short-circuit limiting coupling (SLC)

One system employed in the 70s was the short-circuit limiting coupling (SLC, sometimes also called the 'resonant link'). The main elements were a series capacitor and reactor, of approximately equal ohmic values, connected between the consumer's bus-bar and the utility bus-bar. If a fault, or a major disturbance, occurred in the utility system, auxiliary components automatically changed the capacitor impedance and converted the link to a large magnitude inductive impedance. This buffered the customer's bus-bar from the effects of system faults and the critical plant continued to be supplied, without experiencing a disturbance, from customer's in-house generating plant. The design and implementation of the SLC included measures to guard against SSR effects between the series capacitor and the local generating plant.

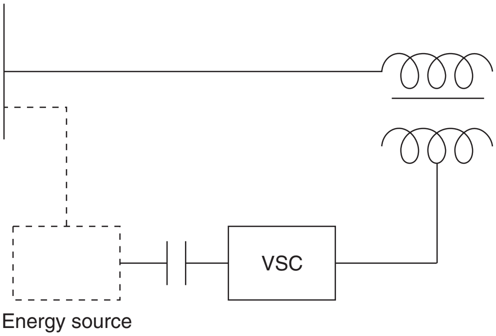


Figure 41.40 DVR

41.7.3.2 Dynamic voltage restorer (DVR)

The advent of voltage-sourced converter technology has enabled another buffering technique to be applied, which avoids the disadvantages of series capacitors and does not rely on local generation being present, the Dynamic Voltage Restorer (DVR), *Figure 41.40*. This can be designed to be effective for voltage dips down to about 50% of rated voltage, which is sufficient for most dips caused by lightning. Energy to maintain the level of the source voltage is usually derived from a simple rectifier on the input side. In order to deal with larger, prolonged voltage dips, it would be necessary to use an energy source of sufficient capacity to supply the load demand for the duration of the voltage dip.

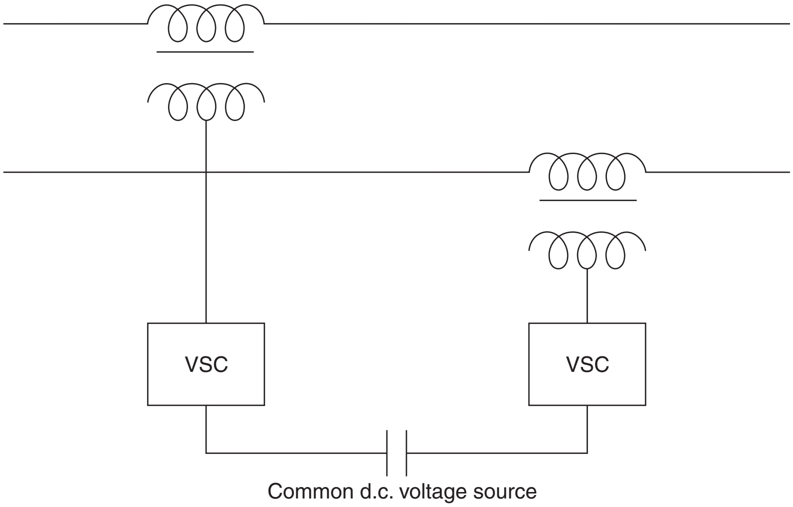


Figure 41.39 IPFC

41.8 Controllers with shunt and series components

It was shown in Section 41.3.5 that the power flow through a simple transmission line is given by

$$P = (V_S\,V_R/X)\cdot\sin\delta\varsigma$$

Although this power flow is dependent on the voltages at the ends of the line, these voltages can be varied only within a small range and therefore cannot be used to exert a strong control over power flow. Section 41.7 showed that the effective value of line reactance, X, could be changed by series compensation and that this provided a strong control over power flow for a given value of phase angle δ.

An alternative form of series connected controller for controlling the power flow is the phase angle regulator (PAR) *Figure 41.41*(*a*). The input and output voltages of the ideal PAR are equal in magnitude, *Figure 41.41*(*b*), but the phase angle, σ, between them can be controlled either to increase or decrease the effective phase angle across the line and hence the power flowing through it, *Figure 41.41*(*c*). Thus, with $V_S = V_P = V_R = V$

$$P' = (V^2/X)\cdot\sin(\delta\varsigma\pm\sigma)\Leftarrow$$

Figure 41.41(*d*) shows that, although the power flow for a given $\delta\varsigma$is changed by the addition or subtraction of σ, the maximum power transfer of the line is not changed. With a very fast, power electronic, tap-changer a rapid control of $\delta\varsigma$ could improve the transient stability margin by providing a valuable prolongation of maximum synchronising power through the line during the first post-fault swing. Fast control could also be used to make an effective contribution to the damping of system power oscillations.

In the practical implementation of a PAR, shunt and series connected transformers are used to inject a voltage in series with the line to give the desired change of phase angle (which, for small angles, is proportional to the injected voltage). The magnitude of the injected voltage is usually made variable by means of a tap-changer. The PAR is a special case of a more general range of phase shifting transformers for which the input and output voltage magnitudes are not always constrained to be equal; the phase angle of the injected voltage depends on the chosen winding configurations of the two transformers. By incorporating power electronics, the speed of control and the flexibility of the series/shunt arrangement can be greatly increased, including unrestricted control of the phase angle in addition to the voltage magnitude.

41.8.1 Quadrature booster transformer (QBT)

One of the simplest types of phase shifting transformer is the quadrature booster transformer; *Figure 41.42*(*a*) shows

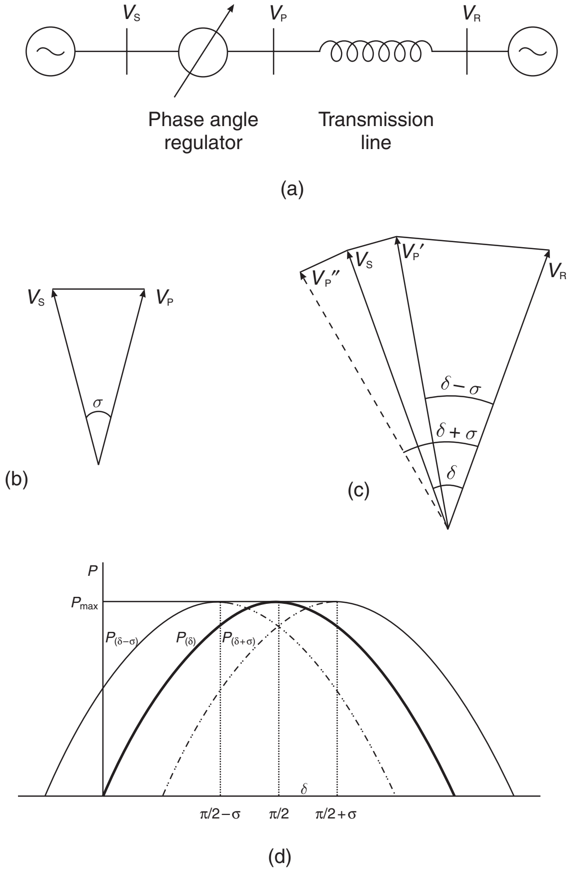


Figure 41.41 Phase angle regulator (PAR) circuit and operation

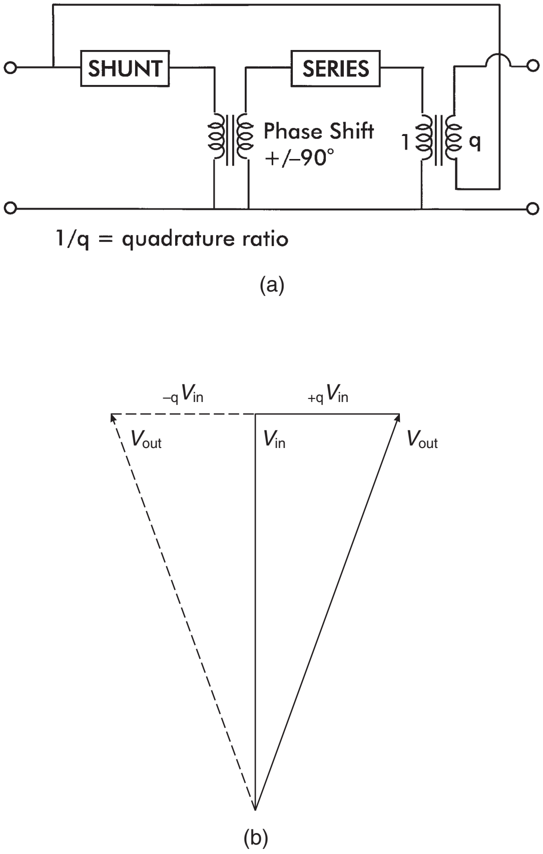


Figure 41.42 Quadrature booster transformer (QBT) circuit and operation

the basic arrangement. One transformer is connected with its primary winding in series with the transmission line, its secondary being fed from the secondary winding of a transformer connected in shunt to the system. The magnitude and polarity of the injected voltage is usually controlled by means of an on-load tap-changer.

The phase angle of the injected voltage is in quadrature with the system voltage, *Figure 41.42*(*b*), and therefore the magnitude of the voltage at the output terminals of the QBT is slightly different from that at the input terminals. For a typical maximum injected voltage equal to 20% of the shunt voltage, the difference will be about 2%; the phase shift is about 11°. The NGC has installed about a dozen QBTs, with ratings up to 2750 MVA, at critical points within the transmission network on England and Wales. Phase shifting transformers are commonly used by utilities in the US to control power flow in parallel, long distance, circuits.

QBTs are usually connected in substations where they can control the power flow in one or more circuits to prevent thermal overloading. As thermal time constants of lines are measured in minutes, a conventional mechanical tap-changer can be operated quickly enough to prevent such overloads.

If there were a requirement to improve transient stability with the aid of a phase shifting transformer, the injected voltage would have to be changed much more rapidly, for example, using a thyristor-based tap-changer instead of a mechanical tap-changer.

41.8.2 Unified power flow controller (UPFC)

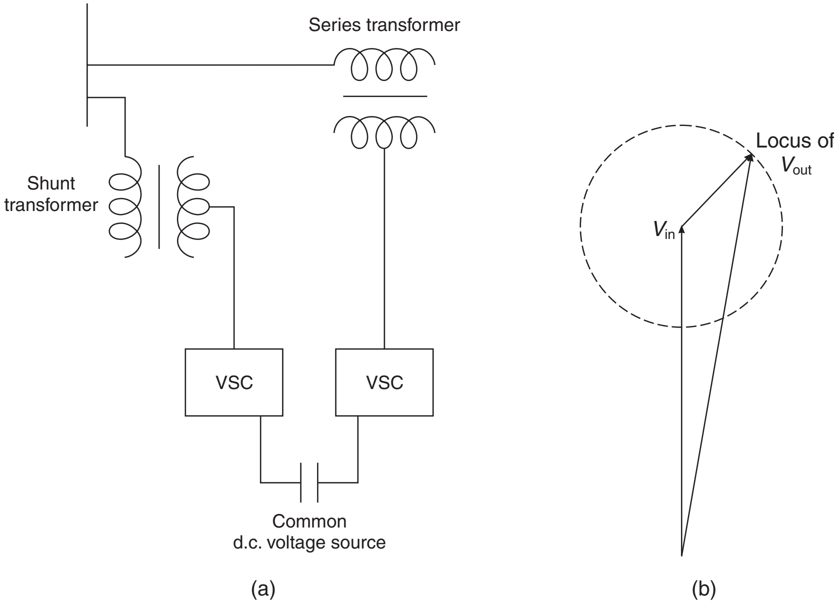
The UPFC, *Figure 41.43*(*a*), is a combination of a STATCOM and an SSSC and offers great flexibility of operation. In this case the secondary winding of each transformer is connected to a three phase voltage-sourced converter. The two converters are connected back-to-back as a d.c. link with their d.c. capacitors acting as a common d.c. voltage source so that, with an appropriate co-ordination between the control systems for the two converters, power can be exchanged in either direction between the shunt and series systems. In addition, each converter can be controlled to supply or absorb Mvar independently to the series and shunt systems.

Thus, whereas a series impedance device, such as SSSC can only develop a voltage in quadrature with the current, the UPFC can generate a voltage in the series circuit with a phase angle that is controllable throughout the full range of 360°, as shown in *Figure 41.43*(*b*). When the voltage is not in quadrature with the current, power will flow out of or into the series connected converter, via the d.c. link, into or out of the power system through the STATCOM. The series converter can therefore provide impedance compensation and/or phase shifting. The STATCOM can provide voltage regulation at its point of connection independently of the power being transferred, subject only to its limits of rating.

A UPFC was commissioned in 1998 at the Inez substation of American Electric Power (AEP) in Kentucky, USA. The series and shunt units can be operated together as a UPFC or may be operated independently as STATCOM and SSSC respectively. The converters are rated at ±160 MVA and use GTOs; they are identical. A spare shunt transformer has been provided so that the converter provided for the series application is also capable of being operated as a STATCOM.

41.9 Special aspects of var compensation

The power demand of industrial loads, such as arc melting furnaces or thyristor drives for rolling mills and mine winders, varies rapidly in both magnitude and power factor. In an arc furnace the impedance of the arc can change completely from one half-cycle to the next. The changes may be particularly large during the early part of a melt; the arc can become short-circuited by scrap metal, when the furnace current is limited only by the circuit reactance, or may be interrupted completely for brief periods. This gives



(a) (b)

Figure 41.43 Unified power flow controller (UPFC) circuit and operation

a current demand which fluctuates very rapidly, in an unpredictable random manner, and may take any value between zero and short-circuit current.

In thyristor drives, very large changes in energy are required during acceleration and deceleration. For instance, in a rolling mill, the drive motor is initially running at low speed and therefore at a low voltage. When an ingot enters the rolls, the drive motors are accelerated to full speed by phasing back the thyristors; this increases the current very rapidly to its rated maximum value, typically 2 to 2.5 times rated current at about 0.2–0.3 power factor lagging. The current demand falls to its full load value when the motor has accelerated to full speed. Unlike the situation in the arc furnace, the changes in power and var requirements are much more predictable, as they occur in a regular cycle and most are relatively slow.

Both kinds of load can cause voltage fluctuations due to the fast changes of var flow through the system reactance. These voltage fluctuations may be unacceptable unless the network is strengthened or reactive compensation applied. The severest limitations usually arise due to lamp flicker and sometimes due to interference with electronic control circuits of other equipment. Different types of lamp have different sensitivities to voltage change. For lamps used in a domestic environment, filament lamps are the most sensitive; their percentage brightness change is about 4 times the percentage voltage change. In comparison, for fluorescent lamps, the ratio of brightness change to voltage change is about 1. Sodium lamps are even less sensitive with a ratio of about 0.5; although high-pressure mercury vapour lamps have a ratio of about 3 they are normally found only in an industrial and not a domestic environment. Because lamp flicker caused by arc furnaces is very difficult to deal with, it will be emphasised in this section, but the general principles are applicable to the compensation of other fluctuating loads.

41.9.1 Lamp flicker compensation

The spectral density of voltage fluctuations produced by an arc furnace is approximately in inverse proportion to the square root of the frequency. People experience a subjective response to lamp flicker; generally, human sensitivity peaks just below 10 Hz for 230 V filament lamps. As can be seen from *Figure 41.44*, a weighted combination of these characteristics shows that the frequencies most liable to cause visual annoyance lie in a band from about 2 to 25 Hz. If the voltage fluctuations at 10 Hz are greater than about 0.2%, then they are likely to cause a noticeable flicker on the luminous output of a 230 V filament lamp. A 110 V lamp of the same wattage has a heavier filament with a greater thermal capacity, which results in a smaller response to voltage fluctuations and the most disturbing frequency is reduced to about 5 to 6 Hz.

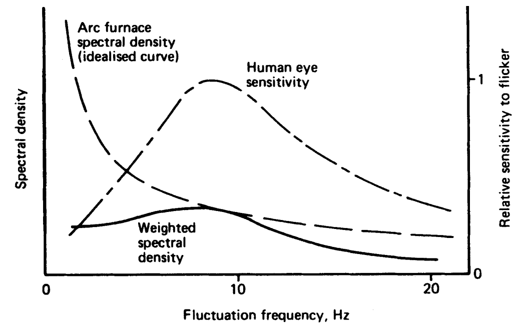


Figure 41.44 Eye sensitivity to arc-furnace-induced flicker

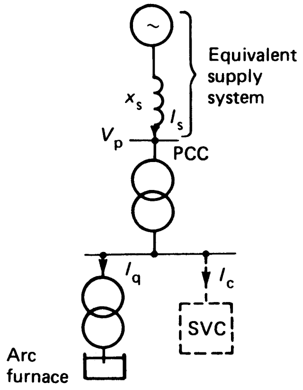


Figure 41.45 Simplified arc furnace supply circuit. PCC is point of common coupling (with other consumers)

A circuit supplying an arc furnace can be simplified to that shown in *Figure 41.45*, where the point of common coupling (pcc) is the point in the network at which other consumers are connected. The resistance of the supply is usually small compared to the reactance, X_s, and the voltage dip at this point, V_p, is predominantly due to the variation of arc furnace var demand. If there is no SVC installed, the reactive current, I_s, in the supply is the same as the furnace reactive current, I_q, and we get:

$$V_p = I_s \cdot X_s = I_q X_s$$

It is thus relatively easy to estimate the magnitudes of the voltage dips caused by var fluctuations, but the annoyance caused by a sequence of rapid voltage dips is difficult to assess. In order to assess and quantify the effects of fluctuating voltage dips on the human eye and brain, a flicker meter has been developed by the International Union for Electroheat (UIE) and has been accepted by the IEC. The flickermeter measures the successive voltage fluctuations and, by means of algorithms worked out from first principles, converts them into numerical values that are compared with what 50% of the population would regard as the threshold of perceptibility for lamp flicker. For this threshold level of lamp flicker, the UIE flickermeter would give a numerical output of 1.0 for 'short term flicker severity' (Pst).

A flickermeter can only be applied when a furnace comes into service and cannot be used directly to predict flicker levels. However, a simple estimating procedure for planning purposes, has been derived empirically from records of complaints of flicker on many installations. This procedure evaluates the 'short-circuit voltage depression' (SCVD) for a proposed arc furnace; this is the change of voltage at the pcc that would be caused by a change in furnace var demand from no-load to a steady three-phase short-circuit on the electrodes. If the SCVD is greater than about 2%, consumers are likely to suffer sufficient annoyance to complain about lamp flicker. For an arc furnace installation with an SCVD of about 1.3%, the UIE flickermeter would typically indicate a maximum Pst value of about 1.

The SCVD criterion can be used to assess the maximum furnace rating that should be connected to a given system, but it can only be used to determine the rating of a compensator for flicker reduction provided that the compensator is capable of reducing all flicker frequencies in the visual annoyance range reasonably equally. Where a compensator has an acceptably linear fluctuation frequency versus speed of response characteristic up to about 25 Hz, then, if connected as shown in *Figure 41.45*, a steady state calculation of SCVD can be used to estimate its rating, i.e. the compensator current jI_c makes up the difference between the allowable $-jI_s$ and the value of $-jI_q$. A high speed of response is essential for flicker reduction. It has been shown that if the compensator has a control-time delay of 10 ms, no matter what its rating, it can give very little reduction of flicker; with a time delay of 20 ms or greater, a range of frequencies within the visual annoyance range will be strongly accentuated. A thyristor switched capacitor compensator, for example, cannot achieve the necessary speed of response to reduce the arc-furnace flicker in the frequency range above 5 Hz, where the human eye is most sensitive.

The harmonic-compensated saturated reactor without a slope-correction circuit has been used to give flicker reduction of up to 3:1. It has been employed successfully in many installations as a bus-bar compensator (*Figure 41.46*(*a*)), having been designed on the basis of the SCVD criterion. The tapped-reactor/saturated-reactor scheme (*Figure 41.46*(*b*)) has been used to obtain a flicker reduction of up to 7:1 for a single arc-furnace. In this scheme the saturated reactors are single-phase devices and the slope correction is achieved by the tapped reactor winding ratios; this compensator inherently compensates for the unbalance loadings of the arc furnace and gives an instantaneous response. It produces considerable harmonic distortion and requires substantial filtering.

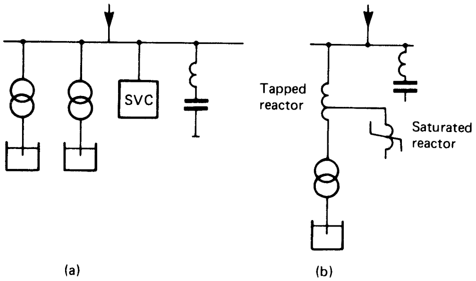


Figure 41.46 Arc-furnace compensation: (a) bus-bar compensation; (b) tapped reactor/saturated reactor compensator

The TCR used as a bus-bar compensator can be made suitable for arc-furnace compensation with a flicker reduction of about 2:1. Voltage-sourced converters, because of their lower reactances and capability for much faster response, can outperform conventional TCRs; present indications point to a flicker reduction of about 4:1 being obtainable.

41.9.2 Phase balancing

Three-phase a.c. systems are intended to operate in a balanced mode. Unbalance can have detrimental effects on equipment connected to the system; rotating plant is particularly sensitive due to additional losses and consequent excessive heating. Most rotating plant is able to accept only 1% of negative phase sequence voltage continuously and 2% for short periods.

Most loads of significant size are 3-phase and operate in a balanced manner. The main exceptions are arc furnaces, mentioned above, and substations supplying a.c. traction loads. Unbalance is also caused by the physical asymmetry of transmission lines; on long lines it is usual to minimise this unbalance by regular transposition of the conductors. In rural distribution systems, unbalance is sometimes caused by unequal distribution of the many single-phase loads between the three phases.

Arc furnaces are unlikely to cause adverse unbalance effects without first causing complaints of lamp flicker. The corrective action of any flicker suppression equipment will normally also limit unbalance contributions to acceptable levels.

The unbalance caused by traction systems has sometimes been corrected by means of conventional thyristor controlled equipment with one phase providing conventional power factor correction and the other two acting as a controllable Steinmetz circuit, *Figure 41.47*, for balancing the power component of traction load. The asymmetrical

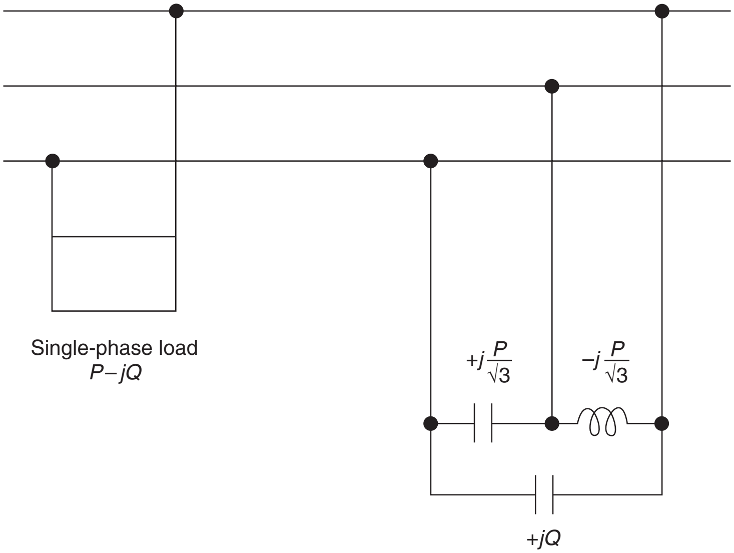


Figure 41.47 Steinmetz circuit for balancing single-phase loads

operation of the three phases of the SVC requires the shunt capacitors to be arranged to filter low order harmonics.

Converter technology now provides an alternative method of phase balancing with a nominally lower rating of equipment. A voltage (or current) sourced converter is used to generate an nps current which is equal in magnitude but in phase opposition to the nps current of the traction load. It thus acts as an active filter for unbalance.

41.9.3 Switched resistors

Switched resistors do not come under the strict heading of reactive power plant but have been used in certain situations to complement such plant. In situations where system transient stability is threatened by loss of synchronism due to the acceleration of a suddenly unloaded generator, switched braking resistors can provide a temporary substitute load for the generators and prevent excessive acceleration. By using thyristor switches instead of mechanical switches, fast repetitive switching can be used to aid in damping system swings.

Thyristor switched resistors have also been used with an arc furnace which comprised a significant proportion of the load on a set of gas turbine generators. In order to reduce maintenance requirements, it was desirable to operate the generators at a fairly steady load and to smooth out the rate of change of load when the furnace was suddenly switched off. This was done by thyristor switching of shunt resistors.

41.9.4 Harmonic currents and harmonic filter design

Many types of plant draw a distorted, non-sinusoidal current from the supply system, as, for instance, large rectifiers (aluminium smelters), controlled converters (traction drives, mine winders, rolling mills) and discharge lamps. The millions of low power, harmonic-producing devices used in the home and workplace, such as light dimmers, television receivers and personal computers also make a surprisingly large contribution to the total voltage distortion on many systems. Power transformers (due to magnetising currents), h.v.d.c. converter stations, SVCs and many other FACTS controllers are sources of waveform distortion in the network itself. These currents and the consequent voltage waveform distortions can be described in terms of fundamental and harmonic frequency components using Fourier analysis.

The harmonic content of voltages and currents can have undesirable disturbing effects on the system and its loads and on adjoining telecommunication circuits. Inductive or capacitive coupling between telephone and power lines induces harmonic interference over the audio-frequency range, and appropriate weighting factors are applied to the individual frequencies when assessing total distortion effect using measures such as telephone interference factor (TIF). The increased use of fibre-optic communication circuits has reduced the extent of this telephone interference.

Rotating generators and induction motors experience increased power loss due to the flow of harmonic currents. Waveform distortion can sometimes cause maloperation of electromagnetic as well as static protective relaying equipment. Shunt capacitor banks primarily for power factor correction are particularly affected by harmonics, as their impedance falls inversely with frequency so that they tend to act as 'sinks' for high frequency harmonic currents and may overload due to increased losses. Shunt capacitors form a resonant circuit with respect to the rest of the system at some frequency above the fundamental (i.e. at a super-synchronous frequency); if the resonant frequency is close to that of a local source (or sometimes even a remote source) of harmonic current, large and potentially damaging harmonic overvoltages and current magnifications can occur. For these reasons, harmonic distortion must be limited to acceptable levels by careful system design and, where necessary, by installation of shunt harmonic filter equipment at appropriate system bus-bars.

In order to minimise adverse harmonic effects, supply authorities apply harmonic current and voltage distortion limits, based on observed disturbances and on a desire to protect other consumers and telecommunication services. Compliance with the limitations at the point of common coupling with other loads may require reduction of system impedance by addition of lines or transformers, or the connection of the distorting load to a different point in the network. In case of converter loads the pulse number may have to be increased from 6 to 12, or even 24.

Connection of a properly designed harmonic filter consisting of a combination of inductive, capacitive and resistive elements at a suitable bus-bar will provide a low-impedance shunt path to the flow of harmonic currents generated by the distorting loads. The filter and the system will share the harmonic current injected by the load in the inverse ratio of their respective impedances. The desired impedance pattern of the filter circuit can usually be achieved by an appropriate combination of single-frequency tuned arms and broadband damped arms, *Figure 41.48*.

The system impedance varies with harmonic frequency in a complex manner, as most systems are not purely inductive. Systems have intrinsic resonant frequencies and, at these frequencies, there are dramatic changes in impedance from inductive to capacitive for only a few per cent variations in frequency. The supply system will, in general, have many different operating configurations and, to take account of the worst conditions, the system impedance vector is sometimes assumed to lie within a circular segment limited by maximum impedance angles. In simpler industrial circuits, the supply impedance is usually dominated by transformer impedance; this transformer impedance is assumed to increase in proportion to frequency to give the system harmonic impedance. This simplifying assumption can also be made (in the absence of other data) if the system up to the point of connection of the filters is not primarily composed of long lines or cables.

The system may contain several harmonic sources other than the particular known loads, or SVCs, or h.v.d.c. converters; in the assessment of the voltage distortion levels and filter component current ratings the effects of such 'pre-existing' harmonics must be included.

The high-pass damped filter shown in *Figure 41.48* is sometimes advantageous compared with sharply tuned arms if the filter is to be satisfactory over a wide range of

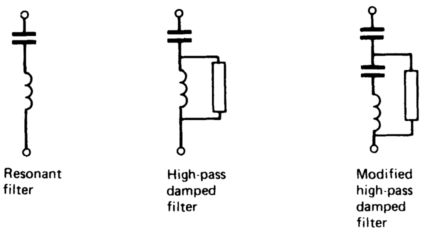


Figure 41.48 Harmonic filters

system operating frequency (e.g. from 49 to 51 Hz); but such damped arms can have greater losses. The modified damped filter is of particular use if the frequencies of the harmonics to be passed are low, as the loss in the resistor can then be excessive at the fundamental frequency unless it is shunted by the series L–C circuit tuned to the fundamental frequency. Such a filter is specifically useful for suppressing unintended resonances at 'non-characteristic' frequencies (e.g. second harmonic) which are inherently low in current magnitude and thus would not give large loss in the resistor at these harmonic frequencies.

The optimum design of a combination of different filter arms (tuned, damped, etc.) to achieve a particular harmonic impedance pattern requires an economic assessment of the effects of choosing different relative proportions (in terms of fundamental Mvar) of the filter arms. The choice is governed not only by the harmonic performance requirements, but also by the fundamental plus harmonic ratings of the filter components—in particular, the capacitor banks. The effects of variations in the filter impedance with system frequency, temperature and component tolerances should be taken into account. The choice of the filter combination should also ensure that the non-characteristic harmonics (e.g. second, third or fourth order), which may be relatively small, are not excessively magnified by unintended resonances between the filter itself and the system. The need to minimise fundamental plus power–frequency losses in the filters also plays an important role in the final choice of the design.

41.10 Future prospects

41.10.1 Recent progress

There has been considerable progress in the field of var control during the last few years. A number of 'future prospects' of only 10 years ago have now been developed into practical FACTS controllers. In the field of shunt var compensation, several different types of voltage-sourced converter are being used in the various STATCOMs that are now in service in transmission systems. TCSCs are being used to provide series compensation of long transmission lines. The concept of the SSSC combined with a STATCOM to form a UPFC has been realised in a unit rated at ±160 MVA and progress has been made towards the installation of an IPFC. In distribution systems there are a number of DVRs providing protection against voltage dips for sensitive loads. Several active filters have been put into service.

41.10.2 Further advances

41.10.2.1 Development of power electronic devices

Many of the developments of power electronic switching devices that have already taken place have been in response to the needs of the large market for industrial drives rather than the needs of the supply industry; this situation is likely to continue. Developments are generally directed towards the reduction of forward voltage drop and overall losses, improved ratings, increased switching speeds and smaller physical size.

41.10.2.2 Development of converter designs

It must be anticipated that converter designs for FACTS controllers will evolve to incorporate the improved devices as they become available. Switching and control using conventional thyristors is mature technology and is likely to continue to be applied in FACTS controllers for many year, with relatively minor improvements. In contrast, voltage sourced converters will continue to be developed and improved as their application becomes more widespread. A particular target for improvement is likely to involve the reduction of harmonic generation by the converters and/or active filtering of harmonic distortion. In order to accommodate some types of energy storage systems, there may be increased activity in developing current-sourced converters.

41.10.2.3 Development of energy sources and storage systems

With sufficient capacity, energy storage systems could contribute to spinning reserves, peak lopping requirements and system frequency control. Battery energy storage systems (BESS) lend themselves to providing energy input to the voltage sourced converters of STATCOMs and superconducting magnetic energy storage (SMEs) is under development but improvements in the economics of energy storage are still needed. Fuel cells offer an alternative type of energy source and regenerative fuel cell systems are currently being developed.

41.10.2.4 Thyristor tap-changers

The technique of using thyristors instead of mechanical switches is available for controlling transformer tap ratios and hence for fast control of voltage and vars. The magnitude as well as the phase angle of the voltage can be made rapidly controllable, but the considerably higher cost has so far not justified their use. New types of power electronic device might improve the viability of thyristor tap-changers.

41.10.3 Future applications of FACTS controllers

41.10.3.1 Long distance transmission

Series capacitor compensation is an economic way of increasing the power transfer capacity of a line, but some of the potential gain in additional capacity may be lost when linear shunt reactors are permanently connected. Subsynchronous resonance conditions must be evaluated at the design stage, but techniques are now available for damping out SSR.

Shunt compensation using SVCs provides good voltage control along the line and at its terminals and can also result in increased transmission capacity. Further benefits arise from the combination of shunt and series compensation, which is likely to be more widely used as transmission companies seek to maximise the utilisation of their assets.

41.10.3.2 HVDC applications

It is increasingly difficult for utilities to obtain permission for the construction of new circuits, or even the up-grading of existing a.c. transmission lines. Undergrounding is extremely expensive. In such situations, the only practical way to increase the power transfer along a particular corridor may be to convert circuits from a.c. to d.c. operation. The easy controllability of d.c. can also enhance the flexibility of the rest of the a.c. system and can contribute to the damping of a.c. system phase angle oscillations. H.v.d.c. can also be used to couple adjacent, non-synchronised systems to provide mutual benefits from shared reserve generating capacity.

These applications of d.c. technology should become increasingly common.

41.10.3.3 Power quality

Power quality is judged on a range of criteria including interruptions to the continuity of supply and disturbances to the voltage magnitude, waveshape and frequency. Customers are increasingly concerned about these aspects and demanding improvements (and reduced tariffs). FACTS controllers will increasingly be used to assist in improving power quality.

41.10.3.4 Energy management systems

Modern energy management systems (EMS) provide an operator with up-to-date information on the state of his system and enable him to obtain very rapid answers to his 'What if?' questions. These systems can provide valuable information for longer-term system planning purposes, especially in respect of the requirements for voltage control and var management. They can also be used to supply control data for the FACTS controllers already operating within the system so that system operation can be made more robust and reliable under both normal and emergency conditions.

Bibliography

CIGRE WORKING GROUP 31-01, *Modelling of static shunt var systems (SVS) for system analysis*, ELECTRA, 51 (March 1977)

CIGRE WORKING GROUP 38-01, TASK FORCE NO 2, Brochure 25, *Static Var Compensators*, CIGRE, Paris (1986)

CIGRE WORKING GROUP 38-01, TASK FORCE NO 3, Brochure 30, *Reactive Power Compensation Analyses and Planning Procedure*, CIGRE, Paris (1989)

CIGRE WORKING GROUP 14-19, Brochure 144, *Static Synchronous Compensator* (*STATCOM*), CIGRE, Paris (1999)

CIGRE TASK FORCE 14-27, Brochure 160, *Unified Power Flow Controller* (*UPFC*), CIGRE, Paris (1997)

CIGRE JOINT WORKING GROUPS 14/37/38/39-24, Brochure 183, *FACTS Technology for Open Access*, CIGRE, Paris (2001)

CIGRE, LONDON SYMPOSIUM, *Working Plant and Systems Harder*, CIGRE, Paris (June 1999)

CIGRE/IEEE FACTS WORKING GROUPS, *FACTS Overview*, IEEE 95 TP 108 (1995)

ELECTRICITY ASSOCIATION, *Planning Levels for Harmonic Voltage Distortion and the Connection of Non-linear Equipment to Transmission Systems and Distribution Networks in the United Kingdom* (*Engineering Recommendation* G5/4), London (2001)

ELECTRICITY COUNCIL, *Planning Limits for Voltage Fluctuations caused by Industrial*, *Commercial and Domestic Equipment in the United Kingdom* (*Engineering Recommendation* P28), London (1989)

HINGORANI, N. L. and GYUGYI, L., *Understanding FACTS*, IEEE Press, New York (2000)

IEC/TR2 60868(1986-09), *Flickermeter—Functional and design specifications*

IEC/TR2 60868-0(1991-05), *Flickermeter—Part 0: Evaluation of flicker severity*

IEC/TR2 61000-3-5(1994-12), *Electromagnetic compatibility* (*EMC*)—*Part 3: Limits—Section 5: Limitation of voltage fluctuations and flicker in low-voltage power supply systems for equipment with rated current greater than 16A*

IEC/TR3 61000-3-6(1996-10), *Electromagnetic compatibility* (*EMC*)—*Part 3: Limits—Section 6: Assessment of emission limits for distorting loads in MV and HV power systems—Basic EMC publication*

IEC/TR3 61000-3-7(1996-11), *Electromagnetic compatibility* (*EMC*)—*Part 3: Limits—Section 7: Assessment of emission limits for fluctuating loads in MV and HV power systems—Basic EMC publication*

IEEE COMMITTEE REPORT, *Proposed terms and definitions for subsynchronous oscillations*, IEEE Transactions on Power Apparatus and Systems, **PAS-99**(2), (March/April 1980)

IEEE COMMITTEE REPORT, *Var management*, *problem recognition and control*, IEEE Transactions on Power Apparatus and Systems, **PAS-103**(8), (August 1984)

IEEE COMMITTEE REPORT, *Static VAR Compensators Planning*, *Operating and Maintenance Experiences*, IEEE Power Engineering Society Winter Meeting, **90THO320-2PWR**, (February 1990)

IEEE Std 1031-2000, *Guide for the functional specification of transmission static var compensators*

IEEE Std PC57.135, *Guide for the application*, *specification and testing of phase shifting transformers*

KNIGHT, R. C. *et al.*, *Relocatable GTO-based static var compensator for NGC substations*, CIGRE Paper 14-06 (1998)

MATHUR, R. M. (Ed.), *Static Compensators for Reactive Power Control*, Canadian Electrical Association, Canada (1984)

MILLER, T. J. E. (Ed.), *Reactive Power Control in Electric Systems*, Wiley, New York (1982)

SONG, Y. H. and JOHNS, A. T. (Ed.), *Flexible a.c. transmission systems*, IEE Power and Energy Series 30, London (1999)

42

Electricity Economics and Trading

B J Cory DSc(Eng), FIEE, Fellow IEEE, CEng
Imperial College, London

Contents

42.1 Introduction 42/3

42.2 Summary of electricity pricing principles 42/3
  42.2.1 Economic efficiency 42/3
  42.2.2 Marginal pricing and risk 42/3
  42.2.3 Delivery charging 42/4

42.3 Electricity markets 42/4

42.4 Market models 42/5

42.5 Reactive market 42/5

42.1 Introduction

Like any commodity or good, electrical energy can be bought and sold in an established market place. Hence buying and selling infrastructures are required to enable quantities and prices to be publicised and contracts negotiated between various groups or individuals. The difference between electrical energy and other commodities is the delivery system—rather than delivering by some form of road, rail, air or sea transport, once generated, electricity can only be delivered by wires over a transmission and distribution (T&D) system (see Chapter 39 for details). For such a system to function, at all times the energy input must be equal to the energy taken out plus the system losses (10% or so). This is an added complication for which the consumer has to pay extra to a system operator as well as a charge to the T&D asset owners for use of the system (UoS). In many countries the T&D asset owner is also the system operator, but this need not be so (compare the transporter as a separate entity to the manufacturer, retailer or road owner). Consequently the economics of electricity supply has not only to include the cost of providing a primary fuel (e.g. for gas, coal or oil fired power stations) or even for the provision of renewable energy production (wind generators, solar cells, hydropower, etc.) it must also cover the cost of delivery over a reliable and secure network, right down to each consumer's supply terminals. There are many components to electricity supply chain economics which will be identified in this chapter, finishing with processes for trading in electrical energy markets.

It should be noted that in many developing countries the production and delivery of electricity is still centrally controlled by the government through a state run industry where planning, operation and prices are organised according to the economic and social welfare priorities of the legislature. Until the 1980s this form of electricity supply industry (ESI) control was common in Europe, S. America and Japan but since then many governments have realised that to obtain investment from the financial markets to develop and expand the ESI some 'unbundling' and 'privatisation' of the businesses bound up with electricity production and supply could be beneficial in reducing prices for the consumer through competition and kick-starting new kinds of economic activity. As a result, deregulation and restructuring of the ESI is taking off in many countries where separate utilities have been run as businesses with private investors for many years, but subject to government or state imposed regulation.

42.2 Summary of electricity pricing principles

42.2.1 Economic efficiency

It is obvious that the price of any commodity in a 'free' market will determine the purchase and use of that commodity. With electrical energy, since storage by the consumer is prohibitively expensive, the price per kWh is important but so is the time of day that it occurs. Over a 24 hour period, electricity is usually priced in one hour or half hour 'slots' for prescribed or transacted amounts delineated in kWh. If this amount is exceeded at the end of the slot, a penalty price could be charged for the excess. To calculate the resulting bill the actual energy taken over the time slot must be metered in quantity and time; this requires a modern electronic meter which can be interrogated for all the slots since the last bill was calculated. Such a meter, preferably with a 2-way communication link to the reading and billing agency, is 5 to 6 times the cost of a simple kWh meter as fitted in residential and small business premises. Consequently, time of day pricing is normally only available where the metering costs are justified such as for customers with annual energy bills of £1000 ($1500) or more. Any customer wishing to keep control of business running costs will no doubt be influenced by the time-of-day energy price and adjust energy use accordingly. As an alternative, the limitation of energy use may be delegated to the supplier for a negotiated reduction in tariff, e.g. off-peak night storage heating tariff. The electricity generator will be charging a price for energy to cover the cost of production plus overheads and profit but over a period of time the customer will adjust requirements to minimise total costs, so in the absence of other factors, a balance between generation and consumption will be achieved. This is called a *Pareto* balance named after the 19th century economist, Alfredo Pareto. It is also one definition of *economic efficiency* as sought after by economists with a free market.

Unfortunately, although the market may be efficient it does not follow that everyone using the market is getting good value for money. Some participants may be constrained in their ability to offer lower prices because of their geographical position in respect of the delivery of their goods or alternatively consumers may be paying more for their energy than if they installed and ran their own generator (e.g. a CHP scheme) provided they could find the capital to purchase it. *Social welfare* is an attempt to measure the benefits of having an infrastructure in place which maximises the sum of the monetary benefits accruing to every individual or group (producers and consumers) in the country or state, including any tangible but useful benefits to which a monetary value can be given. A good measure of useful benefits is the willingness to pay, thereby suggesting that the optimal alternatives or *choice* for both parties as brought about by *competition* is desirable. It is at this stage that the government, state or regulator may step in to limit choice so that some societal group (e.g. the poor) is not disadvantaged. This can be done by *subsidy* from government or by *price capping* by a regulator, just two among many possibilities.

42.2.2 Marginal pricing and risk

Any commodity sold in the market place will become dearer as its availability diminishes. Electrical energy is no exception because its price per kWh for a particular period will rise as generators contract to produce more and more energy, starting with the cheapest plant until the most expensive plant is contracted during some periods. If some production capability is not contracted over, say, a season, it can be mothballed or scrapped. Consequently, at any particular time slot not only can the price per kWh available be recorded but also its rate-of-change with total demand. This rate-of-change is called *short-run marginal cost* (SRMC) and indicates the cost of supplying the next kWh of energy. It can be applied to a single generator, a portfolio of generators or to a whole integrated system. In the latter case it is known as *system marginal price* (SMP) or *system lamda* recognising that it is derived from equalising the marginal costs of all part loaded generators by differential calculus. (Note: in a centrally controlled and integrated network such plant would be loaded by the system operator in ascending order of notified or bid prices—the so-called *economic dispatch*.)

In practice the SRMC, if charged, although more than the average cost of generation, will not bring in sufficient revenue to cover the repayments on the invested or asset cost to enable producers to survive and make a reasonable profit. When these necessary repayment costs are included in the price per kWh, they are called *long run marginal costs* (LRMC). If sufficient revenues are to be made to include LRMC, economic theory requires that over several years SRMC and LRMC should be equal for profit maximisation. However, the usual case is that SRMC is less than LRMC requiring that the producer or supplier must charge a 'supplement' to cover the capital repayment cost of installed plant and any new plant—this supplement is known as *revenue reconciliation*. In market trading it is expected that the seller will include a *mark-up* in the negotiated price to enable the business to continue. The mark-up is a matter of judgement depending upon competition in the market and is part of the *risk taking* strategy adopted by the business. There is a wealth of economic literature and many useful theories such as Ramsey pricing, inverse elasticity rule, benefit maximisation etc. which businesses indulging in markets should take on board. (see Hunt & Shuttleworth, 1996, Schweppe *et al*., 1989).

42.2.3 Delivery charging

With physical commodities such as grain, coal, packaged goods etc. a haulier or transporter is engaged at a contract price to carry out delivery to the customer. Who pays for the delivery is a matter for the contract to specify. With electrical energy there are several complications to take into account, the most important being the *instantaneous* nature of electricity and the need for the existence of a continuous wire path between generators and consumers. A further complication is that as the power system network (at least as far as transmission is involved) is a *meshed* system there is no defined path between a producer and a consumer. Consequently, any delivery charge (known as *use-of-system* charge) in all fairness must depend upon some measure of quantity, based on the provision of installed plant and circuits with adequate capacity and the 'wear & tear' on the system due to the flow of power through the wires. The only reason for system equipment, including overhead lines, cables and transformers, to deteriorate and require maintenance or replacement is that the power flow involves some heating due to losses, the raised temperature thereby causing faster deterioration of insulation than if no heating occurred. Over the power system as a whole, the owners and operators require to recover their costs in an equitable manner. This is done by a use-of-system (UoS) charge agreed with the regulator as competitive delivery systems are unlikely and delivery is a monopoly business.

It is usual for the power network to be split between various owners through which power flows by a path determined by basic electrical laws from the generator to the consumer. Each owner will expect to recover the cost of installing and maintaining equipment (amortised over say 20 years) and for controlling the flows, including voltage control at all parts of the owned network. This requires metering at input and output points such that the delivery charge can be apportioned dependent upon the chosen or agreed rate. Suffice to say, there are a number of established methods for charging based on capacity (kW) and metered energy (kWh), the main requirement being the recovery of sufficient monies to keep the utility owner in business or to satisfy any governmental or regulatory target set for rate-of-return on the service. Details of UoS charging methods can be found in Schweppe, *et al*. 1989 or Weedy & Cory, 1998.

42.3 Electricity markets

A commodity market is where buyers and sellers can negotiate a contract for a designated 'block' of electrical energy. It also enables everyone trading in the market to 'discover' the price that other traders are paying or are prepared to offer their blocks of energy so that an auction can be conducted under established rules. An electricity market has effectively only one commodity on offer but its price is very dependent on the period over which it has to be delivered. This is because the price of electricity is higher at peak times than at other periods of the day, week or season (see Chapter 39 on Planning). In the market through negotiation, contracts can be drawn up with the usual penalties for non-delivery etc. under market rules. The market must be efficiently organised and include mechanisms for collecting revenues due under the contracts and any penalties for non-compliance. This can be a complicated procedure, often requiring legislation to enforce the rules and to deal with disputes.

One of the most difficult features markets need to enforce is that of preventing particular traders setting prices because they have monopoly or near monopoly power due to their size or the lack of effective competition. Also they must police the traders to ensure that collusion on price is outlawed. If natural monopolies exist e.g. in transmission and distribution, then a regulator is appointed to control prices that can be charged. In many cases the same market may deal with a number of different businesses such as generators (producers), shippers (transporters) and suppliers (retailers) to end users. In electricity the markets are now used to dealing with four types of traders, namely generators (single or multiple utilities), transmitters (usually high voltage transmission companies), distributors (lower voltage) and suppliers (retailers to the individual consumer, factory, commercial building, etc.). Up to now, transmitters and distributors have been thought of as natural monopolies implying that other shippers would find it too expensive to offer an alternative means of delivery, but with the growth of small embedded generators who might find it worthwhile to construct their own local distribution network, this natural monopoly could be gradually eroded.

Nowadays, everyone expects markets in commodities or services to be run over the internet by established marketeers and the electricity forward market is no exception. Traders can either switch around to various internet sites to discover what prices and trades are available, or they can stay with one market maker if they have a good reason. The internet market is often global in extent although trades between producers and suppliers would need to ensure that delivery could be made and the delivery charge accounted for. Since some markets include other types of energy e.g. gas, oil, coal, bulk shipping across the world or through another country is well established. Delivering electrical energy via another utility's or country's power system (known as 'wheeling') is now becoming possible with the opening up of many grids for third party access under liberalisation rules or legislation.

42.4 Market models

As the ESI in most industrialised countries begins to open up to trading, the 'model' of generation, transmission, distribution and supply moves from a 'vertically' integrated system often with central government control to the separated 'unbundled' system with each part being run as an independent business (see Hunt & Shuttleworth, 1996). In this latter case, economists insist that in a truly competitive market with many buyers and sellers for all time periods, prices for energy and delivery will become based on marginal cost and the most economic means of providing all consumers with electricity will ensue. Unfortunately, as we have seen, transmission and distribution (T&D) are never likely to become fully competitive and, as monopolies, they will have their prices set by a regulator who is required by legislation to set fair prices with a fair return on assets for the owners. With T&D prices added, the market operates with two prices—a generation price and a delivery price for each geographical zone of the power system. The price differential between zones reflects the cost of delivery including payment for losses and service costs for arranging the delivery by the T&D operators. It is implicitly assumed that all trades can be delivered, but if the delivery system is congested (usually on the transmission network) then the system operator can order up (at a price) extra generation to clear or balance the market. The extra cost of this service must be added to the delivery charge in an agreed manner. In some systems it may be possible to buy priority transmission rights, thereby ensuring that at congested times a trade can be completed. In the extreme event that the negotiated energy quantities in a given time slot cannot be delivered, then some form of load shedding is necessary to save the system from collapse. This will certainly incur penalty payments from the supplier to the consumer for lost production or inconvenience, thereby acting as a ceiling on any traded prices. Although most ESIs are now in some transition stage between government controlled vertically integrated operation and a form of generation and supply competition, the final form of the ESI may well take on different features to those expounded here.

42.5 Reactive market

Besides ensuring that trades can be delivered, the system operator (SO) must keep the 'nodes' or 'bus-bars' in the system running within at least ±10% of their designated voltages. This is a tricky job when the power flows through the network can vary by up to 60% during the course of the day. The mechanism for voltage control is through the injection or extraction of volt-ampere reactive (Var) at the main network nodes. Briefly, Var is not real power but is a controlled oscillation of electrical energy between the capacitance and inductance of the network. Any imbalance of Var must be corrected on a local basis through Var producing or absorbing devices, because if Var has to be transmitted over any distance (e.g. 50 km or more) it can produce a considerable voltage change between adjacent nodes, thereby worsening the voltage control problem. Although no net energy is required to produce or absorb Var, it can lead to small extra losses due to the resistance of the network or inherent in the devices it flows through. It can be measured as kVarh by a suitable meter in a similar fashion to kWh and is conventionally considered positive if injected into the network or negative if extracted from the network. A surplus of Var occurs naturally in most networks (consisting of cables, overhead lines and transformers) under light load conditions leading to higher than normal voltages and a deficit of Var occurs under heavily loaded conditions, leading to lower than nominal voltages.

Fortunately, all synchronous generators are designed to run either underexcited or overexcited through control of the rotor currents from the excitation system associated with each generator. When they are overexcited they produce Var (or operate at a lagging power factor in order to satisfy an inductive load); underexcited they have a limited capacity to absorb Var (run at a leading power factor for a predominantly capacitive load). The penalty, as far as the generator plant owner is concerned, is that overexcitation being the more usual of the two running conditions, incurs extra heating of the generator rotor and a drain on the excitation supply, whilst in the long term it causes a more rapid deterioration of the rotor and stator conductors. Both of these considerations require that the plant owner should be compensated based on the net kVarh metered at the generator terminals. At the consumer's terminals, loads normally consist of circuits requiring magnetisation in some form or other e.g. a motor, power supply with input transformer, thereby demanding a lagging current (Var absorbing). The kVarh absorbed can be metered and charged for if there is an economic case to do so, particularly for large 3 ph. loads supplied at high voltage.

However, unlike real power, Var can be locally produced by inserting a capacitor in shunt with the system i.e. in parallel connection to the load. Hence it is considered that a reactive market (£/kVarh) could encourage consumers to reduce their reactive demand, thereby aiding voltage control of the system. In any case, the SO will need to recover any expenditure on kVarh and this would obviously need to come from consumers. Additionally, the transmission & distribution owner could also enter the reactive market by installing compensation devices (known traditionally as power factor correction) at strategic nodes in the system and covering their amortisation costs with a negotiated charge. In most cases, it is expected that the price per kVarh will be of the order of one tenth that of the average kWh energy price.

Bibliography

1 SCHWEPPE, F. C., CARAMANIS, M. C., TABORS, R. D. and BOHN, R. E., *Spot Pricing of Electricity*, Kluwer, (1989)

2 HUNT, S. and SHUTTLEWORTH, G., *Competition & Choice in Electricity*, John Wiley (1996)

3 WEEDY, B. and Cory, B. J., *Electric Power Systems*, 4th Edition, John Wiley (1998)

43

Power Quality

J Stones

Contents

43.1 Introduction 43/3

43.2 Definition of power quality terms 43/3
  43.2.1 Voltage dip 43/3
  43.2.2 Voltage swell 43/3
  43.2.3 Short interruptions 43/3
  43.2.4 Transients 43/3
  43.2.5 Harmonics 43/4
  43.2.6 Inter-harmonics 43/4
  43.2.7 Flicker 43/4
  43.2.8 Voltage imbalance 43/4
  43.2.9 Frequency deviation 43/4

43.3 Sources of problems 43/4
  43.3.1 Power electronic devices 43/4
  43.3.2 Arcing devices 43/4
  43.3.3 Load switching 43/5
  43.3.4 Large motor starting 43/5
  43.3.5 Embedded generation 43/5
  43.3.6 Sensitive equipment 43/5
  43.3.7 Storm and environment related damage 43/6
  43.3.8 Network equipment and design 43/6

43.4 Effects of power quality problems 43/7

43.5 Measuring power quality 43/7

43.6 Amelioration of power quality problems 43/7
  43.6.1 Earthing practices 43/7
  43.6.2 Standby UPS 43/7
  43.6.3 On-line UPS 43/8
  43.6.4 Hybrid UPS 43/8
  43.6.5 Local or embedded generation 43/8
  43.6.6 Transfer switches 43/8
  43.6.7 Static breakers 43/8
  43.6.8 Active filters and SVCs 43/8
  43.6.9 Passive filters 43/8
  43.6.10 Energy storage system 43/8
  43.6.11 Ferro-resonant transformers 43/8

43.7 Power quality codes and standards 43/8

43.1 Introduction

Power quality is an issue that is becoming increasingly important to electricity consumers at all levels of usage. Sensitive equipment and non-linear loads are commonplace in both the industrial and the domestic environment, because of this a heightened awareness of power quality is developing. Occurrences on the supply network that were once considered ‘normal’ by electricity companies and users are now considered a problem to the users of more sensitive equipment.

43.2 Definition of power quality terms

43.2.1 Voltage dip

By definition a voltage dip is a reduction in the rms voltage in the range of 0.1 to 0.9 p.u. (retained) for duration greater than half a mains cycle and less than 1 minute.

Often referred to as a ‘sag’. Caused by faults, increased load demand and transitional events such as large motor starting (*Figure 43.1*).

43.2.2 Voltage swell

By definition an increase in the rms voltage in the range of 1.1 to 1.8 p.u. for a duration greater than half a mains cycle and less than 1 minute. Caused by system faults, load switching and capacitor switching.

The magnitude of a voltage swell is dependent upon the fault location, relative to the point of measurement, the system impedance and system earthing. An earthed system with delta–star connected substation transformer will ensure little change in the un-faulted phases voltages due to a low impedance path from fault to transformer (*Figure 43.2*).

43.2.3 Short interruptions

An interruption is defined as a reduction in the supply voltage, or load current, to a level less than 0.1 p.u. for a time of not more than 1 minute. Interruptions can be caused by system faults, system equipment failures or control and protection malfunctions.

Because the magnitude of the supply voltage drops below 0.1 p.u. during an interruption, the magnitude of the interruption is considered to be zero, the interruption is defined only by duration. The length of short interruptions is generally governed by the type of protection equipment utilised on the system such as auto-reclosers (*Figure 43.3*).

43.2.4 Transients

A transient is an undesirable momentary deviation of the supply voltage or load current. Transients are generally classified into two categories, impulsive and oscillatory. An impulsive transient is an abrupt change in the steady state voltage and/or current, that occurs outside the power frequency range and is unidirectional in polarity.

Impulsive transients are characterised by the rate of change of voltage or current magnitude, i.e. the rise and fall times. Oscillatory transients are generally characterised by their frequency content and duration. They are often due to the network response to an impulsive transient (*Figure 43.4*).

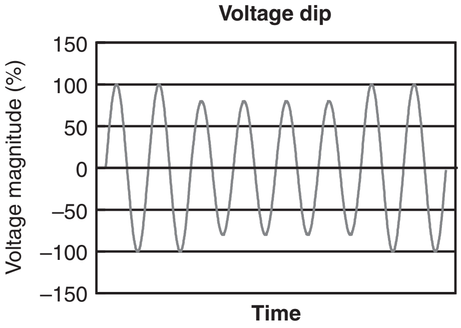


Figure 43.1 Waveform showing a voltage dip

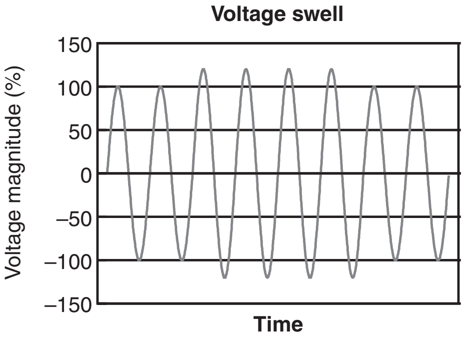


Figure 43.2 Waveform showing a voltage swell

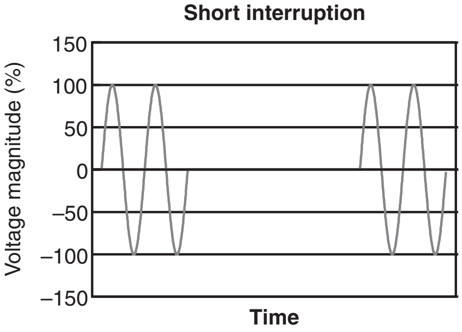


Figure 43.3 Waveform showing a short interruption

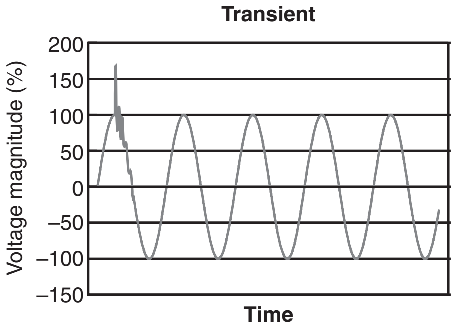


Figure 43.4 Waveform showing an impulsive transient event

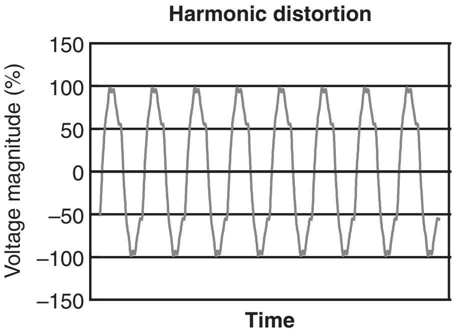


Figure 43.5 Waveform showing harmonic distortion

43.2.5 Harmonics

Harmonics are periodic sinusoidal distortions of the supply voltage or load current caused by non-linear loads. Harmonics are measured in integer multiples of the fundamental supply frequency. Using Fourier series analysis the individual frequency components of the distorted waveform can be described in terms of the harmonic order, magnitude and phase of each component. The quantity total harmonic distortion (THD) is commonly used, it represents a summation of all the harmonic components present in a waveform ($n \neq 1$) (*Figure 43.5*).

43.2.6 Inter-harmonics

Distorted voltage or current waveforms containing periodic distortions of a sinusoidal nature that are not integer multiples of the fundamental supply frequency are termed inter-harmonics.

43.2.7 Flicker

A term used to describe the visual effect of small voltage variations on electrical lighting equipment (particularly tungsten-filament lamps). The frequency range of disturbances affecting lighting appliances, which are detectable by the human eye, is in the range of 1 to 30 Hz.

Flicker severity is the intensity of flicker annoyance defined by the UIE-IEC flicker measuring method and evaluated by the following quantities:

- short term severity (P_{st}) measured over a period of ten minutes
- long term severity (P_{lt}) calculated from a sequence of 12 P_{st} values over a two hour interval, according to the following expression:

$$P_{\mathrm{lt}} = \sqrt[3]{\sum_{i=1}^{12} \frac{P_{\mathrm{st}i}^{3}}{12}}$$

43.2.8 Voltage imbalance

Voltage imbalance is defined as a deviation in the magnitude and/or phase of one or more of the phases, of a three-phase supply, with respect to the magnitude of the other phases and the normal phase angle (120°). Voltage imbalance is more commonly a problem in rural areas where one, or two, phase(s) of a three-phase supply is/are being loaded more than the other(s). Can be expressed as a percentage of the ratio of either zero or negative phase sequence components to the positive phase sequence.

43.2.9 Frequency deviation

A variation in frequency from the nominal supply frequency above/below a predetermined level normally ±0.1%.

43.3 Sources of problems

43.3.1 Power electronic devices

Power electronic devices both cause and are susceptible to power quality disturbances. The most common 'economically damaging' power quality problem encountered involves the use of variable speed drives (see below). All computers contain a power electronic switched mode power supply (SMPS, see below) which is a cheap and convenient method of converting mains supply into low voltage d.c. without expensive transformer windings. These supplies are the cause of a significant increase in the level of 3rd, 5th and 7th harmonic voltage distortion.

43.3.1.1 Variable speed drives

Variable speed motor drives or inverters are highly susceptible to voltage dip disturbances and cause particular problems in industrial processes where loss of mechanical synchronism is an issue. The ideal solution to problems of this nature would be for planning engineers to install equipment that has a 'reasonable level' of susceptibility to voltage dips from the outset, this does not happen for two reasons:

(i) Manufacturers of these drives often don't publish detailed information concerning their equipment's level of susceptibility to voltage dips.
(ii) Where optional equipment filters are offered by manufacturers at the installation stage, many customers opt not to fit them for economic reasons.

43.3.1.2 SMPS (including IT equipment)

A large range of equipment, mainly office and domestic, use switch mode power supplies to convert mains to the required d.c. level. Many of these converters draw a non-linear current from the supply which is high in third and fifth harmonic content.

Because the third harmonic is a 'triplen' harmonic it is of zero order phase sequence and therefore adds in the neutral of a balanced three-phase system. The increasing use of IT equipment has led to concern of the increased overloading of neutral conductors and also overheating of transformers.

Recent developments have seen the use of switch mode power supplies in fluorescent lighting applications, these lighting applications typically represent in the region of 50% of a modern building's load. Many commercial modern buildings have large neutral conductors to cope with the levels of third harmonic, which can theoretically reach three times the magnitude of the fundamental.

43.3.2 Arcing devices

Electric arc furnaces, arc welders and electric discharge lamps are all forms of electric arcing device. These devices are highly non-linear loads the current waveform of which is characterised by an increasing arc current limited only by the network impedance. Large arc furnace installations have typical current requirements of 10s of thousands of amps, welding sets draw current in the range of 100s of amps, individual electric discharge lamps draw only fractions of an amp but when it is considered that a large percentage of the domestic and commercial load requirement is contributed by lighting requirements this has a significant impact.

All arcing devices are sources of harmonic distortion, the arcing load can be represented as a relatively stable source of voltage harmonics. The effects of arc furnaces are difficult to mitigate, balancing the phases with other furnaces will not always be effective as arc furnaces are operated in various modes leading to phase imbalance. Arc welders commonly cause transients in the local network due to the intermittent switching, some electronic equipment should be protected from the effects of these impulsive spikes. Because of the requirement to limit the current within fluorescent lights a ballast is fitted which can add to the level of harmonic distortion of the supply, of particular concern is the level of the third (or higher order triplen) harmonic. Balancing the phases to have equal harmonic load is a good way to minimise the level of the triplen harmonics, but on a star–star connected transformer connection no cancellation will occur.

43.3.3 Load switching

The effect of heavy load switching upon the local network is a fairly common problem causing transients to propagate through to other ‘electrically close’ equipment. These transients can be of surprisingly large voltage magnitude, but have very little energy due to their short duration which is normally measured in terms of milliseconds. Electronic devices which may be sensitive to these voltage impulses can have their operation impaired.

The effect of load switching on the voltage is typically encountered in the form of transient activity (as seen in *Figure 43.4*). This type of transient might occur as the result of switching in a heavy single-phase load, the effect seen on the voltage measured nearby. Other equipment can be protected from these switching transients by electrically isolating them from the affecting equipment.

43.3.4 Large motor starting

Because of the dynamic nature of an induction machine it draws a current depending upon the mode of operation, during starting this current can be as high as six times the normal rated current. This increased loading on the local network has the effect of causing a voltage dip, the magnitude of which is dependent upon the system impedance. It can take several seconds for motors to reach their rated speed, for this reason measures are taken to reduce the level of current drawn during motor starting. These measures are dependent upon the type of motor and drive, most modern motors employ a sophisticated power electronic converter ‘drive’ which in most cases will control the motor’s starting current to a reasonable level. Some lower cost types of motors use series capacitors or resistors to reduce the starting current, these components are then switched out once the motor’s rated speed has been reached. Autotransformers are used to start some older motors, these have a variable secondary winding which allows the motor stator voltage to be controlled and hence the current drawn from the supply.

43.3.5 Embedded generation

Increasing levels of embedded generation predicted in the future are likely to have an effect on power quality. Although it can not be stated that this increased level of dispersed generation on public distribution networks will degrade or improve power quality (this is an issue of some contention) it can be said that there are both advantages and disadvantages to the more widespread application of embedded generation where power quality is concerned.

An increased amount of embedded generation at substation level and below will lead to increased fault levels in the feeders. This increased fault level is one of the major concerns when considering embedded generation issues leading to calls for increased protection.

The voltage of embedded generators, when located at some distance from a substation, must be controlled to ensure power flow from the high voltage (substation bus) to the lower voltage (embedded generator connection).

With wind turbines voltage fluctuation due to variations in wind speed leads to problems with voltage regulation and therefore potential power quality problems on a local level. When a wind turbine blade, rotating at relatively low speed, passes the supporting pylon, the blade is momentarily inefficient in converting the airflow into rotational velocity, this leads to a low frequency ‘pulsing’ causing a continuous flicker problem at the generator output. This effect can be quite extreme in areas electrically close to a windfarm connection.

43.3.6 Sensitive equipment

If it were not for sensitive equipment, power quality would not have become such an issue in recent years as it has. Equipment manufactures are designing and manufacturing ever more sophisticated equipment as time goes on much of which is increasingly susceptible to variations in power quality. There are many issues relating to the subject of equipment sensitivity, the main areas of concern are:

- Reduced equipment operating life.
- Instantaneous equipment malfunction.
- Equipment malfunction to data corruption.
- Reduced process quality.
- Process stoppage.
- Equipment damage.
- Economic damage to operators.
- Safety issues.

As mentioned previously variable speed drives are a major problem. These drives are power electronic converters and are prone to voltage dips of relatively small magnitude.

Any device which depends upon a volatile memory chip for information storage is potentially at risk from power quality events, IT equipment therefore requires protection. Computer central processing units are prone to damage from power quality and quality of supply disturbances, the repeated thermal shock of being energised and de-energised can lead to permanent damage and early failure. Many processes in industry depend upon automated microprocessor control systems. Solutions to the power quality problems encountered with this type of equipment often consists of protection for the control system alone, the actual process not being sensitive to the more common disturbances.

A large part of the problem with sensitive loads is down to inadequate legislation. The EMC limits EN 61000-3 series cover many electromagnetic compatibility requirements that manufacturers of equipment must meet.

The area of susceptibility to voltage dips is not covered in enough depth, manufacturers are being allowed to sell equipment that is too sensitive to voltage dips. There are reasons why these 'loop-holes' in the EMC directive exist, one is that the regulations are based on statistical occurrences, another is that the regulations have to account for existing equipment.

43.3.7 Storm and environment related damage

Lightning strikes are a cause of transient overvoltages often leading to faults. Lightning does not have to strike a conductor in order to inject transients on to the local network, 'impulses' can be induced if lightning strikes near a conductor. The local ground potential can be raised by a nearby strike leading to neutral current flowing to earth via a remote ground, this can have destructive effects on sensitive equipment. Lightning strikes which hit overhead lines often cause 'flashovers' to neighbouring conductors as the insulators break-down, the strike will therefore not only consist of a transient overvoltage but also fault clearing interruptions and dips.

High winds and storm conditions cause widespread disruption to the supply networks. Where disruptions are caused by faults that can be cleared in less than one minute, by the use of auto-reclosers for example, the effect on the network is seen as a power quality issue. Long interruptions, above one minute, are generally seen as reliability or quality of supply issues.

Snow and ice build-up have a severe effect on the reliability of overhead lines, this has obvious power quality/quality of supply consequences.

Sea mists in the vicinity of overhead lines can lead to flashover between conductors, insulators must be cleaned on a regular basis in these areas to avoid these problems. In hot and humid climates dust and heavy dew can cause similar flashover problems requiring non-intrusive insulator cleaning methods.

Damage due to wildlife and trees is common in rural areas, particularly in the spring. As with any faults these are potential causes of power quality problems.

43.3.8 Network equipment and design

Auto-reclosing circuit-breakers are commonly used on rural radial networks to allow faults time to clear and supply to continue without the requirement for manual intervention. The circuit-breakers have the purpose of increasing the security of supply, fuse saving and reducing outage time. As far as customers are concerned they can however cause power quality problems due to the way in which they operate. Fast tripping is the use of circuit-breakers or line reclosers to trip in a very short period of time under fault conditions, this is another method of saving fuses, but has an adverse effect on power quality. As most faults on rural networks are of a transient nature, when such a fault occurs the auto-recloser will trip, after a predetermined time delay the circuit-breaker will reclose. If the fault has cleared then the supply is fully restored and no more disruption will occur, if the fault was not cleared then the breaker will trip again (for up to 4 times). It is this repeated reclosing that is a power quality issue, customers will have their supplies reconnected for each of the instants when the auto-recloser re-connects the supply, this can lead to equipment damage in extreme circumstances. To reduce the level of customer annoyance caused by the repeated operation of auto-reclosers these devices are normally set with a built in 'dead-time' which allows the fault more time to clear prior to recloser operation. A 'lock-out time' can also be set which will stop the unit from reclosing if a certain number of operations have taken place within a pre-set time window.

Capacitor switching is a major cause of transients on the supply network. Capacitors are used to provide reactive power compensation (vars) hence reduce system losses. Some capacitors are permanently connected to the network, others are switched to suit the load conditions. When capacitors are switched in to a supply, voltage transients occur due to the interaction of the network inductive elements and the additional capacitance. Where power factor correction capacitors are installed at a customer's site and utility capacitor switching is taking place the effect of the switching transient can become magnified and oscillatory in nature. This 'voltage magnification' is a function of the impedance of both the capacitance's, the network, end user circuit inductance and the capacitor switching 'frequency'. If the resonance frequency of the end-user circuit and the network resonance in response to the capacitor switching are equivalent, then the maximum voltage magnification will occur. This can amount to double the nominal supply voltage in extreme cases.

Tap-changing transformers are used for the purpose of voltage regulation. They can have either mechanical or electronic devices for changing the transformer tapping to vary the secondary voltage level. Electronic tap changers can respond to load changes in a very short time frame, mechanical devices are more suited to more predictable slower changing loads.

Current limiting fuses are used where the fault current is high and have the effect of improving the overall power quality by isolating such a fault in a very short time frame. Typically rated in thousands of Amps the fuses will isolate a faulted connection in less than half a mains cycle.

Transformer energisation causes large oscillatory inrush currents that have an adverse effect on power quality each time a transformer is switched in to the network. The energisation of a transformer can cause dynamic overvoltages for up to one second after it is switched in, the inrush current is highly distorted. Particular problems are encountered when transformers are connected to power factor correction capacitors, the problem can be eliminated by switching the devices in separately. Ferro-resonance can occur in distribution networks, at frequencies below 300 Hz, as the result of transformer inrush current harmonic components and series connected capacitors resonating with the transformer magnetising inductance. To combat this effect capacitors are de-tuned from known resonance frequencies.

Network sectionalising on a radial distribution network consists of the use of line reclosers in strategic positions which will allow increased probability of continued supply to more critical loads under fault conditions.

Surge arresters are used in areas where lightning strikes are a frequent occurrence. They have non-linear characteristics which allow current surges, induced by lightning to be 'bled-off' to earth. Surge, or lightning, arresters are typically fitted on every few poles on LV systems in lightning prone areas.

Line shielding is another method of reducing the effect of lightning on overhead lines in storm-prone areas. More commonly used on transmission lines, the ground conductor is supported above the phase lines thus reducing the possibility of faults caused by lightning strikes. Line shielding adds substantial cost to a distribution line because the poles must be higher and the benefits of having neutral conductors beneath the phase lines are removed.

43.4 Effects of power quality problems

Power quality can have a large detrimental effect on industrial processes and the commercial sector. Industrial processes differ in their requirements, from a power quality perspective, each having particular 'weaknesses' in terms of power quality attributes. The important power quality considerations to be accounted for to the industrial end-user centre around costs associated with machine down-time, clean-up costs, product quality and equipment failure. Solutions to power quality problems must be implemented by industrial end-users which reflect the cost versus benefit case for implementation.

Domestic customers tend not to be so adversely affected by power quality problems in that equipment in the home tends to depend less upon a high degree of power quality in its normal operation. However, trends in home ownership of IT equipment and sophisticated communications equipment for home entertainment purposes will mean a shift in power quality requirements for worst affected customers in the near future.

Table 43.1 shows power quality quantities and their most common effects on end-users.

43.5 Measuring power quality

The first consideration after having identified a PQ problem is the economic case for solving the problem, it is therefore important to try and quantify the cost of the problem to the end user. It would be useful if the customer could provide information of:

- The nature of the power quality problem
- The estimated cost of the disturbance
- The frequency of occurrence
- The times at which the problem occurred
- Information about the equipment affected
- Information about equipment nearby
- Previous monitoring carried out at the site

Table 43.1 Power quality, quantities and effects

Quantity	*Effect*
Voltage dips	Machine/process downtime, clean up costs, product quality and repair costs all contribute to make these type of problems costly to the end-user.
Transients	Component failure, hardware reboot required, software problems, product quality.
Harmonics	Transformer and neutral conductor heating leading to reduced lifespan. Audio Hum, Video 'flutter', software glitches, power supply failure.
Flicker	Visual irritation.

In order to quantify the less obvious or hidden costs associated with a PQ problem it is necessary to implement a scheme of monitoring at the site of the problem.

A thorough site survey should be carried out prior to adopting a scheme of monitoring, this should consist of:

- Review of site circuit diagrams.
- Obtain detailed information of loads affected.
- Obtain detailed information of suspect loads.
- Equipment failure/malfunction logs.
- Identify the 'significant' problem areas.
- Discussion with relevant personnel involved.
- Preliminary monitoring where necessary.

After all the above information has been collected and considered, if no obvious cause is apparent, then a scheme of monitoring should be adopted to suit the power quality problem. A detailed scheme of monitoring will require some important decisions to be made from the outset, if the outcome of the monitoring is to be of maximum benefit:

- The type of monitoring equipment to be used.
- Where to connect the monitoring equipment.
- Over what a period to monitor.
- What quantities to measure.
- Analysis of disturbance data.

Measurable quantities in power quality include; supply voltage variations, short/long interruptions, voltage dips and swells, harmonics, inter-harmonics, flicker, voltage imbalance, frequency deviation.

43.6 Amelioration of power quality problems

43.6.1 Earthing practices

A large number of reported power quality problems are caused by incorrect earthing practices. Verification of earthing arrangements, particularly when harmonics problems are reported, should always be conducted early in a power quality investigation.

43.6.2 Standby UPS

Consisting of a rectifier, battery, inverter and static switches, the standby UPS is the most popularly used UPS available today. The static transfer switches will be controlled to allow the load to be fed from the mains supply under normal operation, when there is a mains disturbance leading to a reduction in the mains voltage below some predetermined level the switches will open and close respectively. The load will then be fed from the battery, via the inverter ensuring continuation of supply to the load. The inverter output of a standby UPS must always operate in

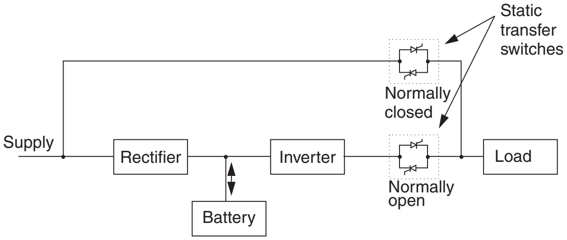


Figure 43.6 Standby UPS schematic arrangement

synchronism with the supply frequency to ensure a smooth transition from one supply to the other (*Figure 43.6*).

43.6.3 On-line UPS

An on-line UPS is configured such that the load is always fed from the UPS, in this way the load is isolated from the mains supply at all times. These systems are in general expensive and have high operating losses. Very similar to a standby system to view schematically, but with a manual transfer switch in place of the static transfer switches.

43.6.4 Hybrid UPS

The hybrid UPS system has a configuration similar to standby UPS systems, with the exception that some form of voltage regulator, such as a ferro-resonant transformer, is used in place of the static switch device(s). The transformer provides regulation to the load and momentary ride-through when the transfer from mains supply to standby UPS is made.

43.6.5 Local or embedded generation

A form of local generation, such as a diesel generator, can be connected to allow for any shortfall in the mains capacity and also to provide ride-through for power quality disturbances. This will in most circumstances be viewed as an expensive solution, as the cost to keep a diesel generator running on-line indefinitely would be a high price to pay for improved power quality. However, some forms of embedded generation, such as micro-turbines, fuel cells and Stirling engines, are likely to have increased domestic usage in the near future.

43.6.6 Transfer switches

Transfer switches are used to transfer a load connection from one supply to another, allowing the choice of two supplies for the load (or sub network), should one supply suffer power disturbances then the other supply will be automatically switched in reducing the possibility of supply disruption to the load.

43.6.7 Static breakers

The power electronic equivalent of a circuit-breaker with a sub cyclic-response time. The static breaker will allow the isolation of faulted circuits in the shortest possible time frame, other nearby loads will therefore have improved power quality.

43.6.8 Active filters and SVCs

The control of reactive power, and therefore harmonics, can be achieved by controlling a proportion of the power systems current through a reactive element. Conventionally this is achieved by switching inductors and capacitors in shunt with the power system, using thyristors. With the SVC the control of the current is achieved by controlling the output voltage magnitude of an inverter. SVCs are used to absorb or inject reactive currents to eliminate the harmonic distorting currents drawn by non-linear loads.

Unified power flow controllers (UPFCs) are similar to SVCs but allow both series and shunt compensation.

43.6.9 Passive filters

Passive filters or power line filters are simple filters consisting of discrete capacitors and/or inductors. Normally designed to attenuate high frequencies (low pass filters), fitted to equipment to remove higher order harmonic frequencies from the supply.

43.6.10 Energy storage system

All electrical energy storage systems have the same basic components, interface with power system, power conditioning system, charge/discharge control and the energy storage medium itself. Each storage medium has different characteristics, energy density, charge/discharge time, effect of repeated cycling on performance and life, cost, maintenance requirements etc. These characteristics help to make the decision of what storage medium is best suited to which application, each medium having merits that make it the most suitable in different circumstances. Energy storage systems available:

(1) Super-conducting magnetic energy storage.
(2) Flywheel energy storage.
(3) Battery/advanced battery energy storage.
(4) Capacitor or ultra-capacitor storage.

43.6.11 Ferro-resonant transformers

A constant voltage, or ferro-resonant, transformer is normally a transformer with a 1:1 turns ratio and with a core that is highly magnetised close to saturation under normal operation. The variation of primary voltage has a much-reduced effect on the secondary voltage, hence the output is not significantly effected by voltage sags.

43.7 Power quality codes and standards

Below is a list of the most commonly used standards and recommendations used in the field of power quality within the European community and USA.

(i) **EN 50 160 (2000)** '*Voltage characteristics of electricity supplied by public distribution systems*'.
(ii) **EN 61000-2-2 (1993)** *EMC Environment section*. '*Compatibility levels for low-frequency conducted disturbances and signalling in public LV power supply systems*'.
(iii) **EN 61000-3-2 (1999)** *EMC Limits*. '*Limits for harmonic current emissions (Equipment input current less than or equal to 16 A per phase)*'.
(iv) **EN 61000-3-3 (1998)** *EMC Limits*. '*Limitation of voltage fluctuations and flicker in low voltage supply systems for equipment with rated current less than or equal to 16 A*'.
(v) **EN 61000-3-4 (1998)** *EMC Limits*. '*Limits for harmonic current emissions (Equipment input current greater than 16 A per phase)*'.
(vi) **EN 61000-3-5 (1994)** *EMC Limits*. '*Limitation of voltage fluctuations and flicker in low voltage supply systems for equipment with rated current greater than 16 A*'.
(vii) **EN 61000-3-6 (1996)** *EMC Limits*. '*Assessment of emission limits for distorting loads in MV and HV power systems*'.
(viii) **EN 61000-4-7 (1995)** '*EMC testing and measurement techniques—General guide on harmonics and inter-*

harmonics measurements and instrumentation, for power supply systems and equipment connected thereto'.

(ix) **EN 61000-4-11 (1994)** '*EMC testing and measurement techniques—Voltage dips, short interruptions and voltage variation immunity tests*'.

(x) **Engineering recommendation G.5/4 (2001)** '*Planning levels for voltage harmonic distortion and the connection of nonlinear equipment to transmission systems and distribution networks in the UK.*'

(xi) **Engineering recommendation P28 (1989)** '*Planning limits for voltage fluctuations caused by industrial, commercial and domestic equipment in the United Kingdom*'.

(xii) **Engineering recommendation P29 (1990)** '*Planning limits for voltage unbalance in the United Kingdom for 132 kV and below*'.

(xiii) **ITIC (Information Technology Industry Council) Curve**—*the ITI (CBEMA) curve, published by the technical committee 3 (TC3), of the information technology industry council Supersedes CBEMA* (Computer and Business Equipment Manufacturers Association) curve.

(xiv) **IEEE Std 1159 (1995)** '*Recommended practice for monitoring electric power quality*'.

(xv) **IEEE 446 (1995)** '*Recommended practice for emergency and standby power systems for industrial and commercial applications—IEEE orange book*'.

(xvi) **IEEE 519 (1992)** '*Recommended practice and requirements for harmonic control in electrical power systems*'.

(xvii) **IEEE 1100 (1993)** '*Recommended practice for powering and grounding electronic equipment—IEEE Emerald Book*'.

(xviii) **IEEE 1159 (1995)** '*Recommended practice for monitoring electric power quality*'.

(xix) **IEEE 1250 (1995)** '*Guide for service to equipment sensitive to momentary voltage disturbances*'

(xx) **IEEE 1346 (1998)** '*Recommended practice for evaluating electric power system compatibility with electronic process equipment*'.

Bibliography

DUGAN, R. C., McGRANAGHAN, M. F. and BEATY, H. W., *Electrical Power Systems Power Quality*, McGraw-Hill (1996)

Section I

Sectors of Electricity Use

44

Road Transport

J S Davenport
Lucas Automotive Ltd
(Section 44.1)

M G Howard BSc, CEng, MRAeS, MACM (USA)
Kennedy and Donkin Transportation Ltd
(Section 44.2)

R H Busby ERD, MSc, CEng, EICE, MIHT, MConsE
Kennedy and Donkin Transportation Ltd
(Section 44.2)

S B Preston
Silent Power Ltd
(Section 44.3)

M J H Chandler CEng, FIEE, FICE, FIHT
GEC Traffic Automation Ltd
(Section 44.4)

Contents

44.1 Electrical equipment of road transport vehicles 44/3
  44.1.1 Batteries 44/3
  44.1.2 Charging systems 44/3
  44.1.3 A.c. generator 44/5
  44.1.4 Ignition systems 44/7
  44.1.5 Starter motor 44/9
  44.1.6 Electronic fuel-injection systems 44/9

44.2 Light rail transit 44/13
  44.2.1 Introduction 44/13
  44.2.2 Definition 44/13
  44.2.3 Rolling stock 44/14
  44.2.4 Trackwork 44/16
  44.2.5 Power supply 44/17
  44.2.6 Signalling and control 44/17
  44.2.7 Stops 44/17
  44.2.8 Depot 44/17
  44.2.9 Economics and investment 44/17

44.3 Battery vehicles 44/17
  44.3.1 Low-speed battery vehicles 44/18
  44.3.2 Advanced electric vehicles 44/22

44.4 Road traffic control and information systems 44/22
  44.4.1 Traffic signalling 44/22
  44.4.2 Traffic control systems 44/25
  44.4.3 Driver information systems 44/26

44.1 Electrical equipment of road transport vehicles

Within the framework of the modern automobile there is an elaborate mechanism to enable energy to be changed readily from one form to another. Although the electrical equipment is only a small proportion of the whole, its satisfactory operation is essential for the normal running of the automobile.

The chemical energy in the fuel is changed within the engine to provide primarily the propulsion of the automobile. A small portion of the transformed energy is taken to operate the electrical equipment. In its turn the electrical apparatus may have to convert this energy into heat (demister), light (legal requirements) and sound (horn), or back to mechanical energy (starter motor).

In addition, the electrical equipment may be called upon to supply energy when the automobile engine is at rest; so an energy storage system is necessary. To store this energy an accumulator or storage battery is used. Electrical energy is converted into chemical energy and is stored in this form in the battery. *Figure 44.1* shows the various energy conversions concerned.

44.1.1 Batteries

The type of battery used in automobile work has been designed to withstand severe vibration and give maximum capacity for minimum weight. A large discharge current for a short duration is required by the starter motor and the battery must be able to meet this requirement.

The two main types of batteries used in automobile work are the lead–acid and the nickel–alkaline.

The former has been more extensively used for private cars. The type developed has a large number of thin plates of suitable mechanical strength giving a large surface area for a given weight and volume. This construction enables the heavy current demands of the starter motor to be met. The magnitude of the discharge current is a function of the plate area exposed to the electrolyte. The specific gravity of the sulphuric acid used in these batteries is higher than that for stationary batteries: it lies between 1.150 and 1.300. Between these density limits the electrolyte has minimum resistivity—a desirable condition in view of the occasional heavy discharge rates. Too high a density may cause damage to the plates and separators. If the discharged battery is in a healthy condition, it is permissible, in urgent cases, to recharge in less than an hour to about 75% of its fully charged state by giving a boosting charge or a fast rate of charge, provided the battery temperature is not allowed to exceed 80°C. Typical voltage/time relations for a lead–acid cell are shown in *Figure 44.2*.

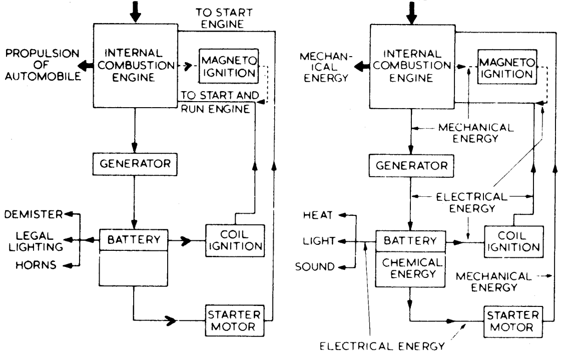


Figure 44.1 Energy transformations in automobile equipment

The nickel–alkaline battery is of much more robust construction than the lead–acid battery. The plates do not buckle when short circuited. The electrolyte is a solution of caustic potash, and unlike the lead–acid battery, no indication of the state of charge is given by the specific gravity. The normal working value is 1.200. After about 2 years service the gravity will fall to 1.160, owing to absorption of impurities from the atmosphere. The action of the battery will become sluggish, and complete renewal of the electrolyte is required.

44.1.2 Charging systems

A generator driven by the automobile engine supplies the energy to charge the battery. If a commutator-type generator is used, an automatic switching device (the cut-out relay) is embodied in the circuit to connect the generator and battery when the generator voltage is of the correct value for charging purposes, and to disconnect the generator when the speed of the automobile engine drops below a certain figure or when the engine is at rest. No cut-out is required in an alternator charging system, since the semiconductor diodes built into the alternator for converting a.c. to d.c. output prevent current reversal.

The automobile driving the generator has a wide speed range, so that the electromotive force produced by the generator, which is proportional to the product of the flux and the speed, would vary considerably unless otherwise prevented. If the required charging current was obtained at low speeds, then the battery would receive a charging current in excess of requirements at the normal driving speeds on the open road.

To keep the generator voltage within the limits set by the battery, a reduction in field current is required for increasing speed. The following systems of control have been adopted for d.c. generators: (1) compensated voltage control; and (2) current voltage control.

44.1.2.1 Compensated voltage control

The compensated voltage control system has a single vibrating-contact regulator which carries both a shunt and a series winding. The shunt winding is connected across the generator output and controls a pair of contacts which, in the normally closed position, short out a high-value resistor in the generator field circuit. Ignoring for the moment the compensating action of the regulator series winding, the shunt winding enables a constant voltage to be obtained

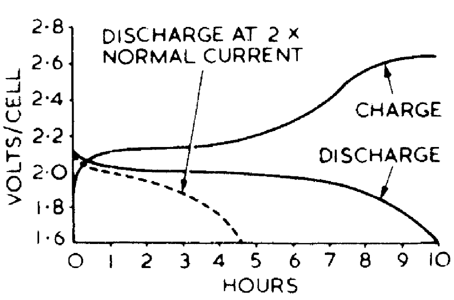


Figure 44.2 Charge and discharge curves of a lead–acid cell

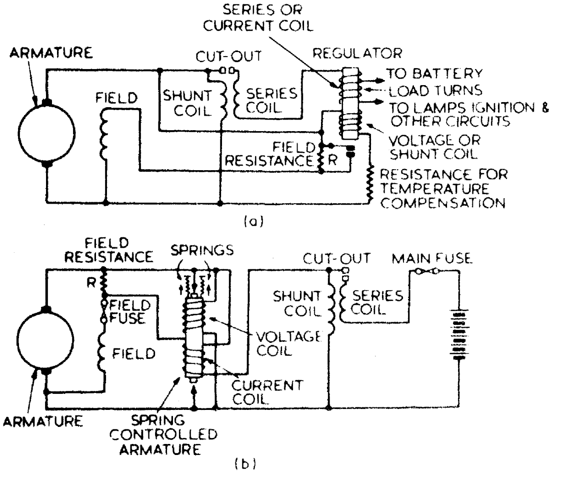


Figure 44.3 (a) Voltage regulator with single contact: minimum field current with R in series with the field winding; maximum field current with no resistance in field circuit; (b) Voltage regulator with double contact: minimum field current when the field winding is short circuited; intermediate field current when R is in series with the field winding; maximum field current when there is no resistance in the field circuit

from the generator. At low speeds the contacts are held together by a spring and generated voltage is applied to the field windings. As speed and voltage rise, the pull of the shunt wound electromagnet overcomes that of the spring and the resistor is inserted into the field circuit, causing the voltage to drop (*Figure 44.3*(*a*)), beginning the next contact opening and closing cycle. This vibratory movement of the contacts takes place some 50 times per second and results in a practically constant voltage, the contacts remaining closed for a shorter period of time as speed rises. For applications where the insertion of resistance does not cover requirements over the whole range of speed, a second pair of contacts, which short circuit the field winding, are arranged to close at the higher speeds, which enables constant voltage to be maintained in two stages. *Figure 44.3*(*b*) shows a barrel-type double-contact voltage regulator.

Constant voltage regulation has to be modified in practice to ensure that current demand by battery charging and other loads is not so high as to overload the generator. In the compensated voltage control system, current control is effected by adding a series winding to the regulator to assist the shunt winding. The voltage characteristic of such a regulator will be a falling one—the voltage will drop with increase of load and thus prevent overloading of the generator. Variation of voltage setting with rising temperature of the voltage coil is prevented by having an armature spring of bimetallic strip type. This temperature compensating device is arranged to give a somewhat higher voltage when cold, to assist in replacing the energy taken from the battery by the starter motor.

The tapering charge characteristic obtained (*Figure 44.4*) means that the compensated voltage control system is unable to utilise those occasions when spare generator capacity is available for increasing the charging rate. This drawback is overcome by current voltage control.

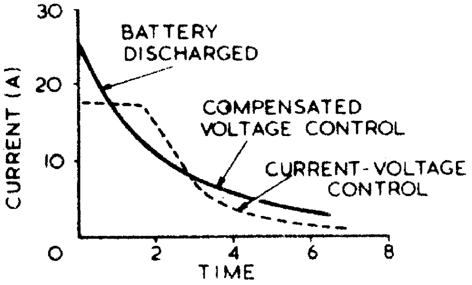


Figure 44.4 Comparison of charging characteristics

44.1.2.2 *Current–voltage control*

A regulator of the current–voltage type (*Figure 44.5*) allows the generator to give its safe maximum output into a discharged battery and to continue to do so until the battery approaches the fully charged condition, when the charging current is reduced to a trickle charge. A constant charging current is maintained until a certain battery voltage is reached; then the voltage regulator operates to give constant voltage control. The charging current will decrease until finally a trickle charge is passing through the battery. This type of regulator is used when the vehicle has a heavy electrical load and a more definite regulation is required than that given by the compensated voltage regulator.

In this type of control two separate regulators are used. These are mechanically separate but electrically interlocked. One regulator is shunt wound and so responsive to voltage conditions; the other is series wound and responsive only to current. Their contacts are connected in series with the generator field circuit.

When current from the generator flows to the battery, it passes through the current regulator winding, and on reaching the maximum rated value, it operates the armature of the current regulator, thus inserting resistance into the field circuit of the generator and decreasing the output current. The vibrations of the armature of the current regulator will prevent the rated output current being exceeded. As the battery voltage rises, the voltage regulator takes over, causing the second pair of contacts to operate, thus controlling the output according to the load and the state of the battery.

If, however, the battery is fully charged with little or no load switched on, then as the speed of the dynamo rises, the terminal voltage will increase to the operating value for the voltage regulator, the resistance will be inserted into the generator field circuit by the operation of the voltage regulator armature and the output current will be a function of the potential difference between generator and battery.

44.1.2.3 *Cut-out or reverse-current relay*

The cut-out relay is a simple automatic relay and is used to connect the battery to the generator when the latter's voltage is just in excess of the battery's. It also disconnects the battery from the generator when a discharge current flows from the battery and the generator tends to run as a 'motor'.

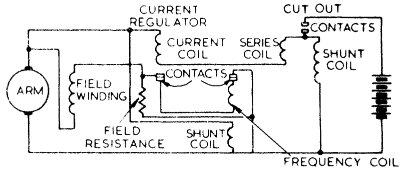


Figure 44.5 Current–voltage regulator

Two coils are provided on the relay: the first, a voltage or shunt coil connected across the generator terminals; and the second, a series coil carrying the generator output current. The fine-gauge wire coil has a large number of turns, but the current is small. This produces the required magnetic effect. When the current flowing in the voltage coil reaches a predetermined value proportional to the generator voltage, the armature is pulled down and the contacts close, allowing current to flow to the battery. This current flows through the series coil and the combined magnetic effect produced by the two coils pulls the contacts close together. The current coil assists the shunt coil as long as a changing current flows. When the generator speed and voltage fall, the current in the series coil decreases, then reverses, so that the resulting magnetic flux produced by the two coils together decreases. With a further fall in generator voltage the magnetic pull exerted on the armature will be overcome by the pull of the spring and the contacts will open, allowing no further discharge current to flow to the generator.

44.1.3 A.c. generator

The continuing increase in the use of electrically operated components on automobiles calls for higher generator outputs. Higher outputs are also required for vehicles which have long periods of idling or frequent stopping and starting.

The increased currents to meet these needs give rise to commutation difficulties if the size, weight and speed of the dynamo are to be consistent with a reasonable service life. An alternative to increasing the size of the d.c. generator is to produce an a.c. in the generating machine and convert to direct current by means of rectifiers. The a.c. machine can have a higher maximum operating speed, so a higher drive ratio can be chosen which will allow a greater output at lower road speeds.

The alternator has a stationary three-phase output winding wound in slots in the stator so that the generated current can be taken directly from the terminals of the stator winding. There is, therefore, no problem of collecting the heavy current from a commutator. The rotating portion of the alternator carries the field winding, which is connected to two slip-rings with associated brush-gear mounted on the end bracket. The field current is small and can be introduced into the windings without arcing at the brushes. The alternator is not affected by the polarity of the magnetic field or the direction of rotation of the rotor, as in the case of the d.c. machine.

A field pole system of the imbricated type is preferred to salient poles by most alternator manufacturers, as it enables a relatively large number of poles (typically 8–12) to be energised by a single field winding.

The alternator cutting-in speed is about the same as that of the d.c. generator; consequently, it provides a useful charge at engine idling speed. The advantage for the alternator is derived from the higher drive ratio.

Rectifier Rectification of the alternating current output is by means of six silicon diodes in a full-wave three-phase bridge connection. The diodes are usually built into the bracket at the slip-ring end of the machine.

Control Since the alternator is designed to be inherently self-regulating as to maximum current output and since the rectifiers eliminate the need for a cut-out relay, the only form of output control necessary is a voltage regulator in the field circuit.

The conventional type of electromagnetic vibrating-contact voltage regulator, with either single or double contacts, has been used. Also employed to enable higher field current to be used and to give longer contact life has been a vibrating-contact regulator with the contacts connected in the base circuit of a transistor, the transistor being protected by a field discharge diode against inductive voltage surges.

To eliminate all maintenance and wear problems, it is now general practice to use solid state components in all parts of the voltage control unit.

44.1.3.1 Separately excited alternator systems

Fitted to a number of passenger cars and light commercial vehicles are Lucas 10AC and 11AC alternators, which have field systems energised from the battery as opposed to directly from the machines. The fully transistorised control employed for these alternators is shown in *Figure 44.6*. Used in place of the voltage coil and tension spring of a vibrating-contact regulator are two silicon transistors for field switching and a Zener diode (voltage control diode) for voltage reference. The Zener diode is a device that opposes the passage of current until a certain voltage is reached, known as the breakdown voltage. When the ignition is switched on, the base current required to render the power transistor T2 conducting is provided through resistor R1; as a consequence, current flows in the collector–emitter portion of T2, which acts as a closed switch in the field circuit and applies battery voltage to the field winding. Rising voltage generated across the stator output winding is applied to the potential divider (R3, R2 and R4) and, according to the position of the tapping point on R2, a proportion of this potential is applied to the Zener diode ZD. When the breakdown point of the Zener diode is reached, the diode conducts and current flows in the base circuit of a driver transistor T1. The base current of T2 is reduced and so is the alternator field excitation. To limit power dissipation, it is desirable to switch the voltage across the field winding rapidly on and off instead of using the transistors to provide continuous regulation; this oscillation is achieved by the positive feedback circuit comprising resistor R5 and capacitor C2. Transistor T2 is protected from very high inductive voltage surges by the surge quench diode D connected across the field winding, which also serves to provide a measure of field current smoothing. Radio interference is eliminated by negative feedback provided by capacitor C1. Resistor R6 provides a path for any small leakage currents through the Zener diode that would otherwise flow through the T1 base circuit at high temperatures and adversely affect regulator action. Automatic compensation for changes in ambient temperature is provided by the thermistor connected in parallel with resistor R3.

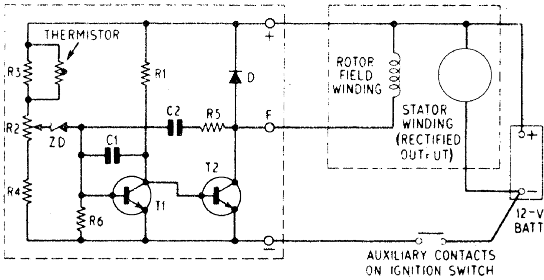


Figure 44.6 Fully transistorised control for a Lucas 10AC or 11AC alternator (positive earth circuit)

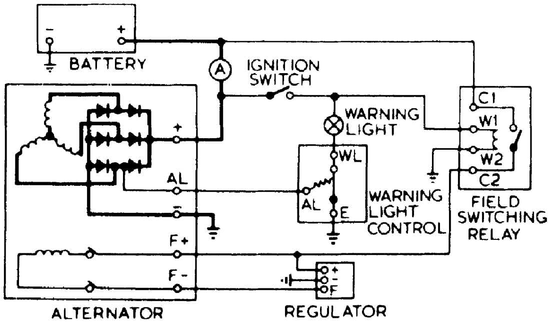


Figure 44.7 A Lucas 10AC or 11AC alternator charging system

Figure 44.7 shows a typical Lucas 10AC or 11AC alternator charging system diagram, including a field switching relay and warning light control unit of 'hot wire' type. The hot-wire resistor is connected to the centre point of one of the pairs of diodes in the alternator; when hot, it lengthens and allows the contacts to open under spring tension, extinguishing the warning lamp.

44.1.3.2 Self-excited alternator systems

The majority of present-day alternators employ what is known as the nine-diode system, and are self-excited at normal operating speeds. The nine-diode arrangement has become popular owing to the simplicity of the warning lamp circuit, which, if combined with an in-built electronic regulator, provides a machine with the minimum of external wiring (*Figure 44.8*). The three field excitation diodes for supplying the field with rectified a.c. are included in the rectifier pack. Operating voltage is built up at starting by energisation of the field winding from the battery through an ignition (or start) switch, a non-charge warning lamp and the voltage regulator.

Lucas manufacture a series of alternators of this type—all with built-in electronic control units—having outputs ranging from 28 to 75 A. An exploded view of a typical alternator is shown in *Figure 44.9*.

Control In the simple self-excited system shown in *Figure 44.8* the regulator controls the voltage at the field diodes' output terminal and, therefore, controls battery voltage only indirectly. For systems employing high-output alternators or having significant cable resistance between battery and alternator—for example, because the battery is remotely sited—indirect control of battery voltage becomes unacceptable. For such applications, progressive levels of control can be provided as follows: (1) the regulator can be designed to sense and control the voltage at the alternator output terminal; or (2) the sensing lead of the regulator can be brought out of the machine to sense and control battery voltage directly. Method (2) is the most appropriate for remote-battery/high-power systems.

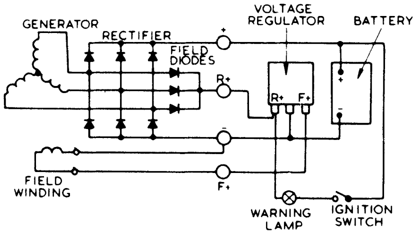


Figure 44.8 Charging system with self-excited alternator

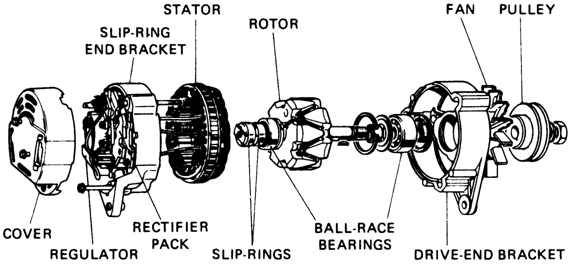


Figure 44.9 Exploded view of a typical nine-diode alternator with a built-in regulator

Figure 44.10 shows a Lucas regulator designed to sense alternator output terminal voltage. Field drive transistor T_2 is an integrated Darlington-pair device giving increased gain. This enables the circuit resistor values to be substantially increased, thereby reducing the permanent battery drain (taken through the sensing lead) to an acceptable level. Base drive resistor R_4 is fed from the sensing lead so that, in the event of the latter being disconnected or broken, T_2 is turned off, shutting down the alternator. If R_4 were fed from the field diodes, an open-circuit sensing lead would result in continuous maximum output from the alternator and destruction of the battery.

Figure 44.11 is an example of a regulator designed for remote voltage sensing. It includes the safety feature which affords protection against an open-circuit sensing lead (described above) and, in addition, protection against damage from disconnection of the alternator output lead. In the latter event, since the battery is no longer being charged, the battery/sensing lead voltage falls and the regulator therefore increases the alternator field excitation. This increases the alternator output voltage, which further increases field excitation. Without protection this cumulative effect can result in the destruction of the field drive transistor T_2. Protection is provided by way of resistor R_6 and diode D_2. R_6 is chosen so that under normal regulating conditions insufficient current flows through the potentiometer chain to reverse-bias diode D_2, which is therefore conducting: hence, the regulator controls the sensing lead voltage. The rise in alternator output terminal voltage, which occurs with a break in the output lead, increases the current through R_6, causing D_2 to become reverse-biased. Regulation of the output terminal voltage now takes place at a safe level dictated by R_6, R_1, R_2, R_3 and ZD_1, and T_1

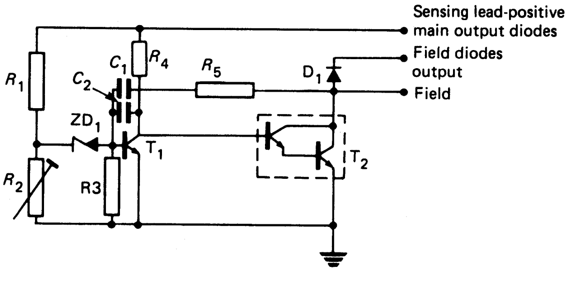


Figure 44.10 Regulator used to sense alternator output terminal voltage

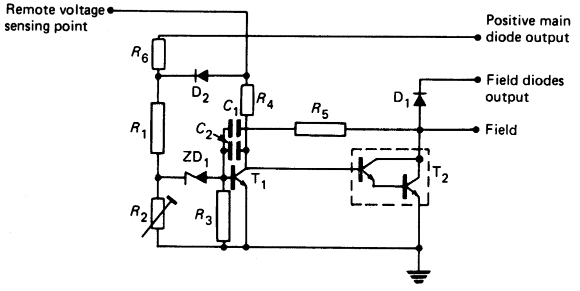


Figure 44.11 Regulator for remote voltage sensing

base–emitter voltage. This secondary regulation level is typically 4 V above the normal regulating voltage.

This regulator can be designed to work with a remote network providing battery temperature compensation. The network, which includes a thermistor, is connected in series with the voltage sensing lead and is mounted to sense battery temperature directly.

Lucas alternators incorporating nine-diode systems are also protected against surge voltages arising from disconnection of the alternator at high output currents, and from switching transients resulting from other equipment (e.g. the ignition system). The protection takes the form of a high-power Zener diode connected between the field diodes' output terminal and earth. Surge voltages are thereby limited to a safe level of typically 35 V. If the Zener diode is overloaded, its normal failure mode is short-circuit. This cuts off the field excitation and switches on the warning lamp. No other damage occurs and the normal charging function of the alternator is restored on the Zener diode's being renewed. A detailed discussion of transient causes and protection is given in SAE paper number 730043, 'Transient overvoltages in alternator systems'. Later alternators incorporate regulators whose semiconductor devices can withstand overvoltage surges. On these alternators the separate overload diode is deleted.

44.1.4 Ignition systems

The ignition of the compressed charge in a cylinder of an internal combustion engine with petrol as the fuel is brought about by an electric discharge. This discharge takes place across the points of a sparking plug. The voltage required to produce this spark will vary between 7000 and 25 000 V. The equipment may be self-generating, as in the magneto (seldom used on automobiles), or battery powered, as with coil ignition. The high-voltage impulses must be delivered to each cylinder as required and at a precisely controlled time in accordance with engine speed and load. There are several systems now in use to satisfy these requirements.

44.1.4.1 Contact breaker ignition

The major components of a modern coil ignition system are shown in *Figure 44.12.*

Storage battery The source of low-tension current.

Primary winding A coil of a few hundred turns of relatively heavy wire connected to the battery through a contact-breaker.

Secondary winding A coil of many thousand turns of fine wire in which the high-tension voltage is induced. One end is usually connected to the more positive end of the primary winding, the other to the sparking plugs through the distributor.

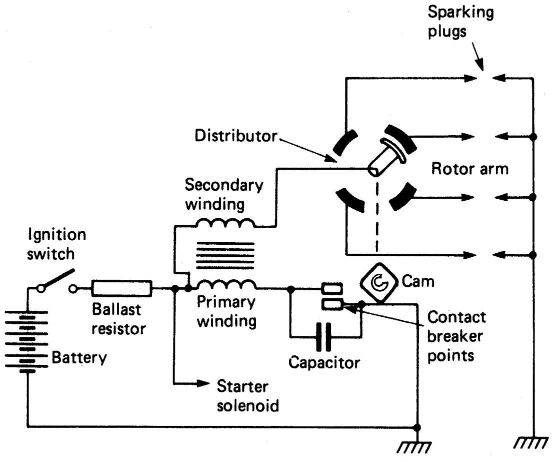


Figure 44.12 Elements of a coil ignition system

Contact-breaker A switch for opening and closing the primary circuit at the required instants. It is operated by a cam driven at half engine speed.

Capacitor Connected across the contact-breaker points to suppress arcing.

Rotor arm Together with the distributor head, it forms a rotary switch to distribute the high voltage to each of the cylinders in correct order of firing. The rotor arm may incorporate a resistor to aid radio interference suppression.

Ballast resistor A device to aid starting. It is a resistor of 1–2 Ω connected in series with the primary winding and arranged to be short-circuited when the starter motor is operated.

Sparking plug The high voltage is led to the insulated central electrode of the plug (one per cylinder) and the spark passes across the gap at the required instant. The gap must be maintained within certain limits, usually 0.3–1.0 mm. In selecting a sparking plug for a given engine, attention must be given to the engine manufacturer's recommendations. The reach of the plug must be such that the correct amount of projection into the combustion chamber is achieved. The correct heat grade is also important: otherwise pre-ignition (overheating) or misfiring (too cold) may occur.

Distributor 'Distributor' is a generic term used to describe the assembly of contact-breaker, a capacitor, rotor arm and distributor head. It contains a shaft driven at half engine speed on which is mounted a centrifugally operated mechanism coupled to the cam. The mechanism is restrained by two springs and is so designed as to maintain the relationships between spark timing and speed. The cam has as many lobes as there are cylinders.

Usually incorporated is a vacuum control which automatically advances the ignition timing as the engine load falls. Control is achieved by a spring loaded diaphragm which moves in response to changes in depression in the carburettor venturi. As the depression increases, the diaphragm moves the contact-breaker mounting plate contrary to cam rotation and so advances the ignition. Occasionally, more complex vacuum control devices are employed having facilities to retard the ignition timing under certain engine operating modes (usually closed throttle deceleration) to reduce hydrocarbon exhaust emissions.

44.1.4.2 Electronic (contactless) ignition

In contact-breaker ignition systems, it is necessary for consistent performance to adjust and, if necessary, to replace the contact points at regular intervals. Failure to do so may result in misfiring and/or excessive levels of exhaust emissions, which in most countries are now controlled by legislation. As a result, electronic (contactless) ignition is used on many vehicles—especially in countries with severe exhaust emission regulations. A general circuit is shown in *Figure 44.13*. The current interrupting function of the contact breaker is now performed by a power transistor, whereas the timing requirement is handled by a sensor in conjunction with a control module.

Many forms of timing transducers are in use. That in *Figure 44.13* is of variable reluctance type. A permanent magnet drives a flux through a magnetic circuit which includes a pickup coil and the distributor shaft. As the shaft revolves at half engine speed, the flux is modulated by a reluctor (replacing the cam of conventional ignition), which is shaped like a cog with as many teeth as there are cylinders. The varying flux generates in the pickup coil a voltage waveform which, after modification by the control module, switches off the power transistor (and, therefore, the coil primary current), to produce a spark. Timing and distribution of the spark are identical with a conventional contact breaker system.

Other forms of timing transducers are:

(1) a photoelectric sensor which is switched by an infra-red light beam interrupted by a shutter;
(2) a Hall-effect device in which a magnetic field is interrupted by a steel shutter—the varying flux in the Hall device causes a proportional voltage change across it; and
(3) radiofrequency differential transformers and eddy-current transducers in which the output voltage is varied by rotating magnetic devices.

In its simplest form the electronic control module may be an amplification circuit to drive the switching transistor, and the duty ratio of the sensor signal—equivalent to dwell angle in conventional systems—remains unchanged. More complex systems are in use in which the control module calculates the point of current switch-on in the primary winding according to engine speed. This allows use of special ignition coils having a very short time constant. The module also has a current-limiting facility of usually 5–7 A and a device which switches off the current when the engine is stalled. The advantage of such systems is that the ignition spark energy is constant over wide speed and battery voltage ranges. Hence, they are usually referred to as 'constant-energy systems'.

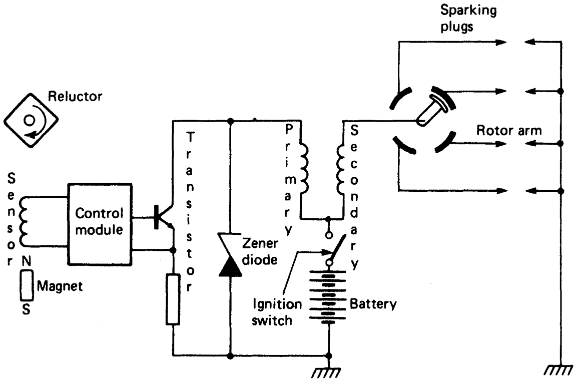


Figure 44.13 Electronic (contactless) ignition system

44.1.4.3 Electronically timed ignition systems

Contactless ignition systems overcome the problems of contact point degradation but the limitations of spark timing by centrifugal and vacuum operated mechanisms remain. As legislation on exhaust emission levels and energy conservation increases in severity, there is an urgent need to improve the efficiency of internal combustion engines, and a greater accuracy and flexibility in spark timing is required. This is possible with computer control of spark timing. A typical arrangement is shown in *Figure 44.14*.

The timing information for the engine is contained in a semiconductor memory unit. Typically the timing values for over 500 points of the engine speed/load characteristic may be stored permanently in the memory. Sensors placed at the engine crankshaft and in the inlet manifold generate signals in accordance with engine speed, crankshaft position and engine load. These signals are processed by the central processor unit (microprocessor) into a binary form suitable for addressing the memory. The nearest timing value in the memory for the particular speed and load of the engine is read into the spark generator unit, where, by means of a count-down procedure, the current in an ignition coil primary winding is interrupted at precisely the instant to produce a spark in the correct cylinder.

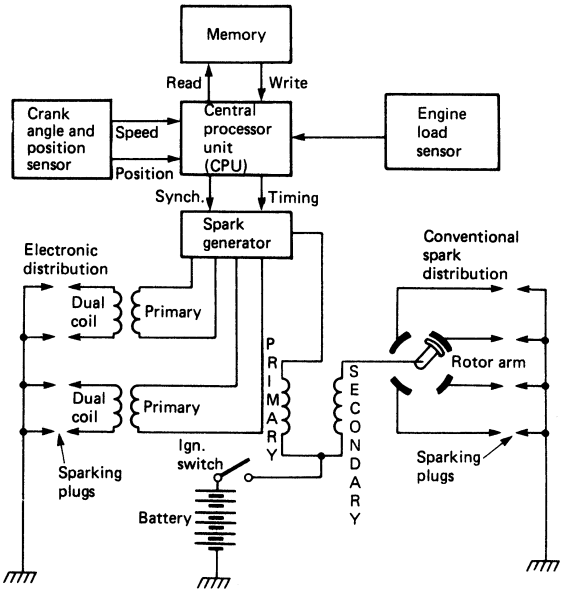


Figure 44.14 Digitally timed ignition (mechanical and electronic distribution systems are shown)

The spark generator unit may include a constant energy circuit as previously described, although other more complex energy control circuits may be employed and can be programmed by instructions also held in the memory.

Distribution of the high voltage to the sparking plugs may be achieved by a distributor head and rotor arm, or obtained electronically by sequential switching of the primary windings of dual output ignition coils (two for a four-cylinder engine).

The crankshaft sensors are usually of the eddy current type but occasionally variable reluctance devices are used. Because of their exposed position, a robust construction is essential. A segmented disc fastened either to the rear face of the flywheel or at the crankshaft pulley is used to generate the signals. Where electronic distribution is used, two crankshaft sensors are necessary, the second being required to generate a suitable synchronising signal.

Load signals may be derived from a simple diaphragm operated potentiometer or by a more sophisticated silicon strain-gauge device. Usually the inlet manifold depression is measured rather than that in the carburettor venturi.

More ambitious systems may include programmed timing offsets for ambient temperature, engine coolant temperature, atmospheric pressure and exhaust-gas circulation.

44.1.5 Starter motor

A starter motor is required to run the internal combustion engine up to a speed sufficient to produce satisfactory carburation.

The starter motor is mounted on the engine casing and a pinion on the end of the starter motor shaft engages with the flywheel teeth. The gear ratio between pinion and flywheel is about 10:1. A machine capable of developing its maximum torque at zero speed is required. The series wound motor has speed and torque characteristics ideal for this purpose.

The engagement of the pinion with the flywheel is effected in different ways. Perhaps the two most commonly used are the inertia engaged pinion and the pre-engaged pinion methods.

In inertia engagement the drive pinion is mounted freely on a helically threaded sleeve on the armature motor shaft. When the starter switch is operated, the armature shaft revolves, causing the pinion, owing to its inertia, to revolve more slowly than the shaft. Consequently, the pinion is propelled along the shaft by the thread into mesh with the flywheel ring gear. Torque is then transmitted from the shaft to the sleeve and pinion through a heavy torsion spring, which takes up the initial shock of engagement. As soon as the engine fires, the load on the pinion teeth is reversed and the pinion tends to be thrown out of engagement. Inertia drives are usually inboard, i.e. the pinion moves inward towards the starter motor to engage with the ring gear; an inboard is lighter and cheaper than an outboard starter.

To obtain maximum lock torque (i.e. turning effort at zero speed), the flux and armature current must be at a maximum, so resistance in the starter circuit (windings, cables, switch and all connections) must be a minimum; any additional resistance will reduce the starting torque. Generally, the inertia engaged starter motor is energised via a solenoid switch, permitting the use of a shorter starter cable and assuring firm closing of the main starter-switch contacts, with consequent reduction in voltage drop. The use of graphite brushes with a high metallic content also assists in minimising loss of voltage.

While inertia drive has been the most popular method of pinion engagement for British petrol-engined vehicles, the use of outboard pre-engaged drive is increasing. The pre-engaged starter is essential on all vehicles exported to cold climates and for compression ignition engines which need a prolonged starting period.

The simplest pre-engaged type of drive is the overrunning clutch type. In this drive, the pinion is pushed into mesh by a forked lever when the starter switch is operated, the lever often being operated by the plunger of a solenoid switch mounted on the motor casing. Motor current is automatically switched on after a set distance of lever movement. The pinion is retained in mesh until the starter switch is released, when a spring returns it. To overcome edge-to-edge tooth contact and ensure meshing, spring pressure or a rotating motion is applied to the pinion. An overrunning clutch carried by the pinion prevents the motor armature from being driven by the flywheel after the engine has fired. Various refinements may be incorporated, especially in heavy-duty starters. Among these are: a slip device in the overrunning clutch to protect the motor against overload; a solenoid switch carrying a series closing coil and a shunt hold-on coil; an armature braking or other device to reduce the possibility of re-engagement while the armature and drive are still rotating; a two-stage solenoid switch to ensure full engagement of the starter pinion into the flywheel teeth before maximum torque is developed (*Figure 44.15*).

Two other pre-engaged types of starter are used for heavy compression ignition engines—the coaxial and axial types.

The compact size of the coaxial starter is achieved by mounting a two-stage operating solenoid and switching mechanism inside the yoke, coaxial with the armature shaft. When the starter solenoid is energised, the plunger is attracted into the solenoid, which causes the pinion sleeve and integral pinion to move axially along the armature shaft. At the same time the first-stage contacts close, to energise the starter windings through a built-in resistor (*Figure 44.16*). The armature rotates under reduced power and the pinion is driven into engagement by means of the armature shaft helix. When the pinion is almost fully engaged, the second-stage contacts close, to cut out the resistor, which enables full power to be developed.

The axial starter employs a sliding armature, which is moved axially against spring pressure to bring the pinion into mesh. The starter also has a two-stage switching arrangement (*Figure 44.17*) to ensure that pinion/ring gear engagement occurs before maximum torque is developed.

44.1.6 Electronic fuel-injection systems

To comply with exhaust regulations and to optimise fuel consumption, the modern petrol engine requires a fuel system of extreme accuracy, reliability and flexibility. To meet this need various electronic fuel injection systems have been developed, the one shown in *Figure 44.18* being typical. The system consists principally of an air-flow meter, engine speed sensor, throttle switch, air and coolant temperature sensors, control unit and fuel injectors. Fuel is delivered to the injectors at constant pressure, so that the amount of fuel to be injected is determined solely by the time for which the injectors are held open. Both the time of opening of the injectors and the period for which they are held open are determined by the electronic control unit from the various sensor signals it receives.

Fuel injectors There is one injector per cylinder with each injector clamped between the fuel rail and the inlet manifold. An injector consists of a solenoid-operated needle valve (*Figure 44.19*). The movable plunger is attached to

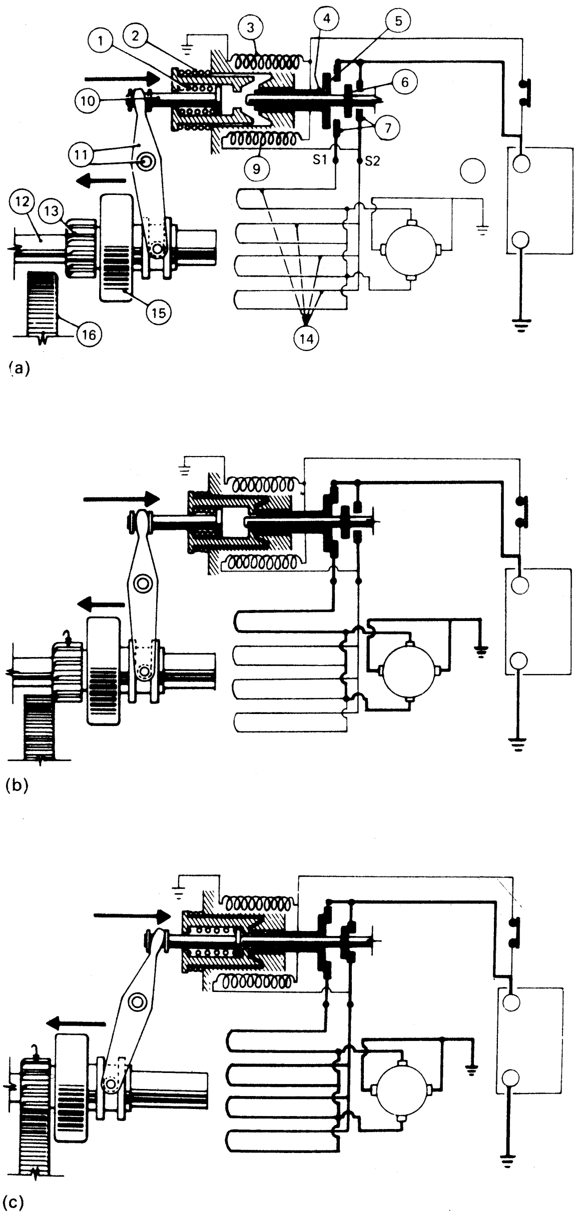

1
2
3
4
5
6
7
9
10
11
12
13
14
15
16
S1
S2
(a)
(b)
(c)

Figure 44.15 Operation of a two-stage switching solenoid: (a) The solenoid is energised in the conventional manner to move the pinion towards the gear ring on the vehicle flywheel: 1, engagement spring; 2, return spring; 3, solenoid hold-on winding; 4, switch-operating spindles (concentric); 5, first set of contacts; 6, second set of contacts; 7, fixed contacts; 8, battery; 9, solenoid operating winder; 10, plunger; 11, operating level and pivot; 12, armature shaft; 13, pinion; 14, field system (four field coils in parallel); 15, roller clutch; 16, gear ring; (b) If tooth-to-tooth abutment occurs, the first set of solenoid contacts close and energise one field coil only, thus giving low power indexing to move the pinion teeth into a meshing position; (c) On full drive engagement, the second set of solenoid contacts close, giving full cranking power. If the pinion teeth, on moving forward, can mesh immediately with the gear ring, full drive engagement takes place with the simultaneous closing of both contacts in the final stage

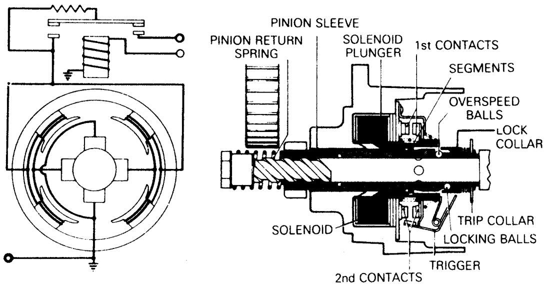

PINION SLEEVE
PINION RETURN SPRING
SOLENOID PLUNGER
1st CONTACTS
SEGMENTS
OVERSPEED BALLS
LOCK COLLAR
TRIP COLLAR
LOCKING BALLS
TRIGGER
SOLENOID
2nd CONTACTS

Figure 44.16 Internal wiring and construction of the two-stage switching mechanism of a coaxial starter

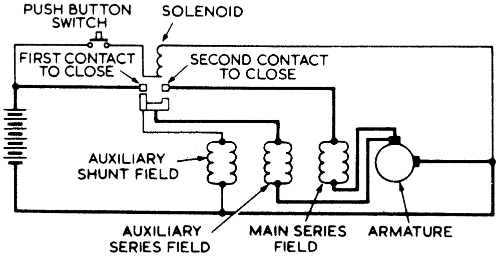

PUSH BUTTON SWITCH
SOLENOID
FIRST CONTACT TO CLOSE
SECOND CONTACT TO CLOSE
AUXILIARY SHUNT FIELD
AUXILIARY SERIES FIELD
MAIN SERIES FIELD
ARMATURE

Figure 44.17 Sliding armature or axial type starter

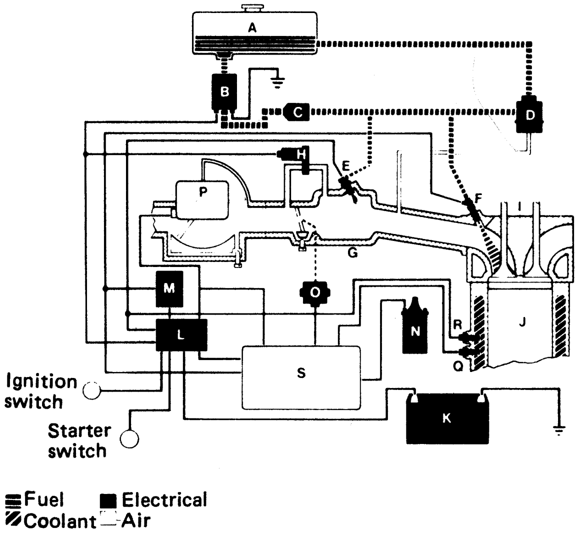

A
B
C
D
E
F
G
H
I
J
K
L
M
N
O
P
Q
R
S
Ignition switch
Starter switch
Fuel
Electrical
Coolant
Air

Figure 44.18 Lucas air-flow meter electronic fuel injection system: A, fuel tank; B, fuel pump; C, fuel filter; D, fuel-pressure regulation; E, cold-start fuel injector; F, fuel injector; G, intake manifold; H, extra-air valve; I, cylinder head; J, piston; K, battery; L, relay; M, power resistor; N, ignition coil; O, throttle switch; P, air-flow sensor; Q, coolant-temperature sensor; R, thermo-time switch; S, electronic control limit

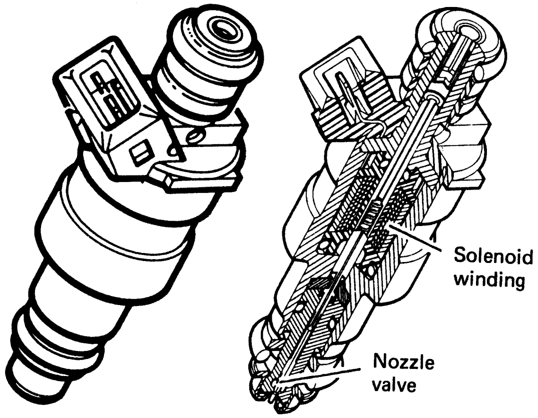


Figure 44.19 Model 8NJ fuel injector

the needle which is held in the closed position by a helical spring. The solenoid winding is located in the rear of the injector body and may be either 16 or 2.5 Ω resistance, depending upon engine design, size and number of cylinders.

The electronic control unit energises the injector solenoid winding, creating a magnetic field which attracts the plunger away from the nozzle seat. This allows pressurised fuel to flow through the injector via an inbuilt filter, into the inlet manifold.

Injectors should always be replaced at the intervals specified by the vehicle manufacturer.

Air-flow meters The basic fuel requirement of the engine is determined from engine load data supplied by the air-flow meter, which is situated between the air filter and the induction manifold. In *moving flap air-flow meters* air flow through the meter deflects a movable flap, which takes up a defined angular position depending upon the force exerted on it by the incoming air. A potentiometer operated by the flap converts the angular position to a corresponding voltage. In the control unit this voltage is divided by engine speed to give the air intake per stroke, from which the basic fuel requirement is then derived.

Since the air-flow meter is used to determine the total mass of air drawn into the engine, adjustments to the basic fuel requirement have to be made for variations in air density (which is temperature dependent). An air temperature sensor is incorporated within the air-flow meter and is connected to the control unit for this purpose.

The *hot-wire air-flow meter* (*Figure 44.20*) has a cast alloy body with an electronic module mounted on the top. Air from the air cleaner is drawn through the body of the air-flow meter and into the engine. Some of this intake air flows through a small bypass in which two wires are mounted, a sensing wire and a compensating wire. A heating current is passed through the sensing wire via the electronic module. The compensating wire is also connected to the module, but is unheated.

As air is drawn over the wires, a cooling effect occurs which alters the value of the resistance and current in the sensing wire. The compensating wire reacts only to the temperature of the intake air. The electronic module monitors the reaction of these wires, thus measuring the air-flow rate, and sends details of air-flow rate to the electronic control unit.

Note: each air-flow meter module is matched to its sensor and body during manufacture; therefore repair of the unit is not possible if a fault occurs.

Adjustments to the basic fuel quantity are also necessary during engine cranking, cold starting and warm-up, idling, full-load operation and acceleration.

Cranking During cranking at any engine temperature, a signal from the starter switching circuit is provided to the control unit in order to increase the ‘open’ time of all

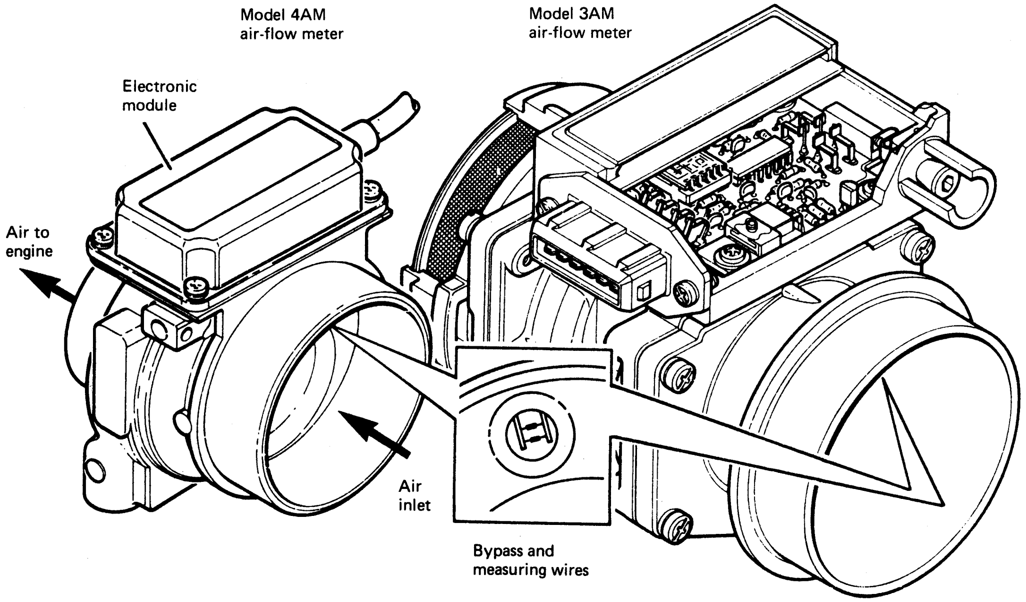


Figure 44.20 Hot-wire air-flow meters

the injectors above that required to supply the basic fuel quantity.

Cold starting and warm-up During cold starting a greatly enriched mixture is required to offset the effect of fuel condensing on the walls of the inlet port and cylinders. This extra fuel is provided by a cold-start injector which delivers a finely atomised spray into the inlet manifold. The cold-start injector operates only when the starter is energised and the thermotime (bimetallic) switch—which is sensitive to coolant temperature—completes the electrical circuit. Additional air required during cold starting and warm-up is controlled by an extra-air valve which bypasses the throttle butterfly. The valve aperture is adjusted by the action of a bimetal strip, which is responsive to the combined temperatures of the engine and an internal heater. The valve becomes fully closed when normal running temperatures are reached. During engine warm-up the control unit steadily decreases fuel enrichment in accordance with signals received from the coolant temperature sensor.

Idling, full-load and accelerating modes In *Figure 44.18* a throttle position switch with two sets of contacts is used to signal engine idling or full-load operating conditions to the control unit and thereby obtain the necessary fuel enrichment. Some applications use a potentiometer instead of a throttle switch. Opening of the throttle is then detected by the control unit as an increasing voltage from the potentiometer, causing acceleration enrichment circuits to be triggered.

44.1.6.1 *Closed-loop electronic fuel injection systems*

To obtain the very low exhaust emissions required by stringent legislation, a closed-loop electronic fuel injection system may be employed in conjunction with a three-way exhaust catalyst (*Figure 44.21*). The catalyst works at optimum efficiency in converting carbon monoxide and hydrocarbon emissions into carbon dioxide and water, and nitrogen oxides into oxygen and nitrogen when the exhaust gases are from an engine operating near to the stoichiometric air/fuel ratio. The operating condition will be indicated by the amount of oxygen present in the exhaust gases, and is monitored in the closed-loop system by an oxygen (lambda) sensor mounted in the exhaust manifold. The sensor provides a feedback signal to the control unit which continuously adjusts the fuelling level to maintain engine operation at the stoichiometric air/fuel ratio.

44.1.6.2 *Digital electronic fuel-injection system*

Until recently, all electronic fuel injection systems have employed analogue computing techniques. This has meant that for reasons of unit size and cost the complexity of the fuelling schedule stored in the analogue control unit has had to be limited and significant compromises made (*Figure 44.22*). In contrast, the latest Lucas electronic fuel injection system employs a digital control unit (incorporating large-scale integrated circuits) in which it has been possible to match very accurately the fuel requirements of an engine under all operating conditions: moreover, this has been achieved in a control unit which has a size readily accommodated on the vehicle.

A key part of the control-unit information-processing capability is a digital read-only memory (1024 bit) which contains the fuel schedule. The latter is stored as a function of 16 discrete values of engine speed and 8 of load. The fuel requirement at each of these memory sites is identified by an 8-bit number. The fuelling characteristic is smoothed between the points of inflexion (i.e. the discrete value stored at each memory site) by a 32-point interpolation procedure which operates on the load and speed signals to effectively increase the memory size by a factor of 16 ×~~16~~.

The basic steps in the processing of information within the control unit are shown in *Figure 44.23*. The engine-speed signal (obtained from the ignition system) and the load signal are each converted into a digital word which is modulated by the interpolation function. The modified numbers representing speed and load are then used to select the site in the memory that stores the fuel requirements for these operating conditions. The memory output number (fuel quantity) is fed into a number-to-time counter, where it is stepped to zero by the fuel trim oscillator. The time taken to count down to zero will therefore be proportional to the memory output and will determine the 'open' period of an injector. Signals of air and coolant temperatures and acceleration adjust the fuel quantity read-out of the memory by modifying the frequency of the fuel trim oscillator. The solenoid operated injectors are energised (usually in groups) through power circuits for the countdown period of the number-to-time counter, when fuel is delivered to the engine.

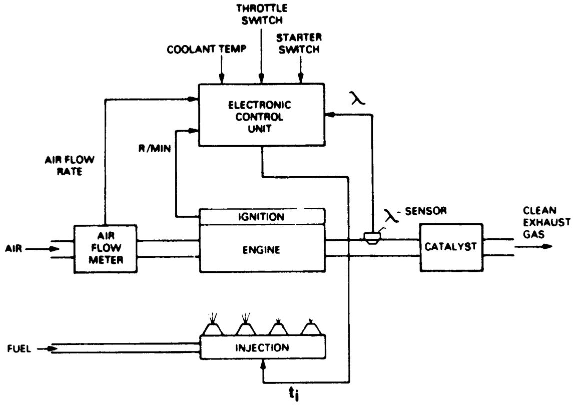


Figure 44.21 Closed-loop electronic fuel injection system

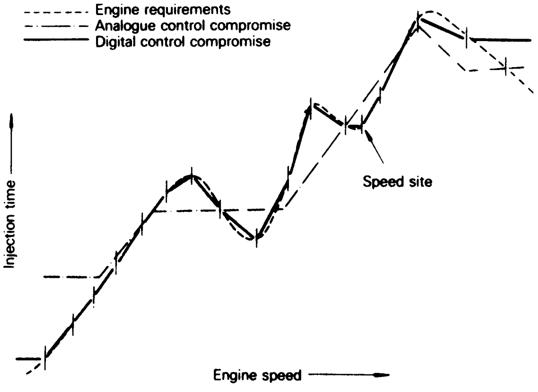


Figure 44.22 Comparison of the fuelling characteristics of analogue and digital systems for a given engine load

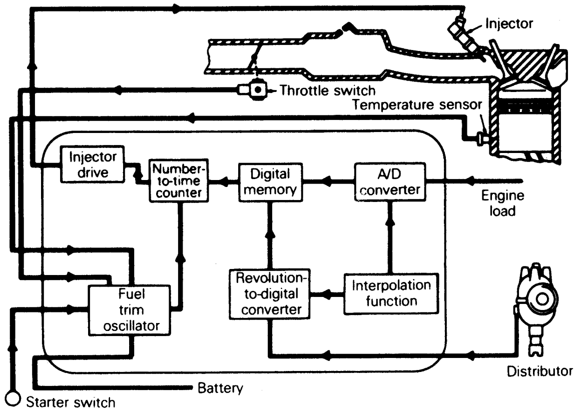


Figure 44.23 Signal processing within the digital electronic fuel injection control unit

44.2 Light rail transit

44.2.1 Introduction

Such has been the rapid development of light rail technology that inevitably it has become subject to many misconceptions. It has also been heavily oversubscribed in written terms and has therefore been the subject of transport 'hype'. It has varyingly been referred to as light weight railway suburban stock at one end of the scale and as an up-market successor to the tram on the other. Whilst in the formative years there is an element of truth in this sort of statement, nevertheless the developing forms of transport have crystallised into its recognisably present status of light rail transit.

Present day light rail transit systems fall into a number of defined categories—the older systems used in many parts of the world (particularly Europe), and the new systems to be found in the UK and the USA, where the public transport renaissance is most marked. These new systems have greatly benefited from the new technologies manifest in traction power supply and collection, suspension and running gear, and control and traction equipment, but in many respects more importantly in the public recognition of the advantages of light rail transit in a country where urban and peri-urban congestion with its environmental pollution, noise, congestion, and attendant loss of working hours has taken dramatic toll of our lifestyles. The new system with light-bodied energy-efficient vehicles and effective operating specification is now firmly established.

The new systems originate from the 1920s and 1930s where, in the UK, design and construction was fragmented and standards were at variance. This situation was rectified in the early 1930s by introducing standard types of tram, mainly as a standardised product in the USA arising from the President's Conference Committee. The present light rail transit systems are designed to form an integral link within an overall transport system. They will support the distributive role which the motor bus still fulfils and provide a fast and efficient carrier system between centres of population, as well as an interface with heavy rail systems.

44.2.2 Definition

Light rail transit is derived from the old tramway or streetcar systems. These operate mostly on the street in mixed traffic but sometimes also have limited separation from other traffic by preferential treatment at junctions or separate rights of way. Although tramways were almost totally abandoned in the UK, they continued to be developed elsewhere in the world into the higher quality light rail systems. These used higher capacity vehicles which made greater use of separated rights of way outside the central urban areas to avoid road congestion and thus offer shorter journey times. In some instances, separate rights of way were created even within the central areas.

Light rail transit applications range from on-street operation using small vehicles (100 passengers) through to fully segregated operation with large vehicles (200 passengers or more). Light rail transit uses steel-wheeled articulated vehicles propelled by rotary traction motors which normally collect power from an overhead wire at up to 750 V d.c.

The vehicles may have single or double articulated bodies and may operate in trains of up to four vehicles, limited by station platform length, and by traffic management restrictions on street-running sections. Vehicles are manually driven and may have driving positions at each end, and doors on each side, to permit bidirectional working. Some systems use single-ended vehicles with driving positions at one end and doors on one side only. These offer greater passenger capacity for a given size of vehicle, but require space for turning loops at the ends of the line. Vehicles are usually manually driven, although automatic driving is possible. Vehicles run on conventional railway track of either 1000 or 1435 mm gauge, although new systems almost invariably use 1435 mm (standard gauge). For segregated operation, conventional railway rails and turnouts may be used, although grooved rail and special turnouts are need for street running.

Light rail transit is used in many European cities and elsewhere in the world and is growing in popularity in the UK. At present, for example, there are over 100 systems in the CIS, 58 systems in Germany, 19 in Japan, 14 in the USA, 14 in Romania, 13 in Poland, 10 in Czechoslovakia, 6 in Switzerland, and 5 in each of Austria, France and Italy.

In the UK, Blackpool operates the oldest system on the mainland, being the only remaining tramway system. The Tyne and Wear Metro uses vehicles based on light rail technology, but in most other respects this system is operated like a heavy rail system, being segregated and fully signalled. The original Docklands Light Railway also used vehicles based on light rail technology but was also fully segregated, taking advantage of this segregation to use automatically controlled, driverless vehicles and third rail, low level conductor rails. Following its planned expansion and extension it will become even more atypical.

A scheme has been opened in Manchester; the Midlands Metro has an Act of Parliament for its first line, and has Bills for two further lines currently before Parliament. A system for Sheffield is under construction with a grant from the Department of Transport under Section 56 of the Transport Act. Advanced Transport for Avon has obtained one private Act of Parliament and deposited two further Bills in November 1989. At present in the UK more than 40 other schemes are being considered.

Light rail transit can operate on elevated or in tunnelled sections to provide segregation. However, constructing a new segregated alignment would involve almost as much demolition and land acquisition as building a new road. Elevating the route would add some £7 million to £11 million per double-track kilometre to the capital costs of the system as well as posing severe environmental problems in installing the guideway over existing roads. Tunnelling would cost between £20 million and £30 million per double-track kilometre and, in addition to this civil cost, the system would have to be designed to the same standards as an underground railway from the point of view of safety, which would increase the cost further.

In the UK, HM Railway Inspectorate has defined three types of operation:

(1) *LRT1* for street running where light rail vehicles share the carriageway with other road users;
(2) *LRT2* for street running where the light rail vehicles run in a road where the track is not shared with other traffic, but is available for use by other road traffic in an emergency; and
(3) *LRT3* where the light rail tracks are fully segregated from all other road users.

The main advantages include an ability to move a large volume of people quickly, quietly and comfortably whilst being pollution free at the point of operation. It fills the gap in the transport spectrum between buses and heavy rail, i.e. between about 4000 to 12 000 passengers an hour. Its main disadvantage is its inflexibility as it runs on set track, although the tracks themselves do not create a physical barrier to pedestrians crossing the road in street running sections.

44.2.3 Rolling stock

44.2.3.1 Examples

Two typical light rail vehicles are shown in *Figure 44.24*. *Figure 44.24*(*a*) shows a vehicle which is widely used in Germany on systems with extensive street running. It is available in either a single or double articulated form. It is 2.3 m wide and in its single articulated form it is 20 m long overall and accommodates 101 passengers of whom 36 are seated. *Figure 44.24*(*b*) shows a larger vehicle which is more widely used on systems which have a higher degree of segregation from other traffic. It has a single articulation and is 2.65 m wide and 27 m long overall. It accommodates 156 passengers of whom 70 are seated. The performance of a typical modern light rail vehicle is summarised in *Table 44.1*.

44.2.3.2 Dimensions

Light rail vehicles may be up to 30 m in length overall and the recommended widths are 2.2, 2.4 and 2.65 m. The largest width of 2.65 m permits seating to be arranged two seats either side of central gangway (2 + 2). Narrower

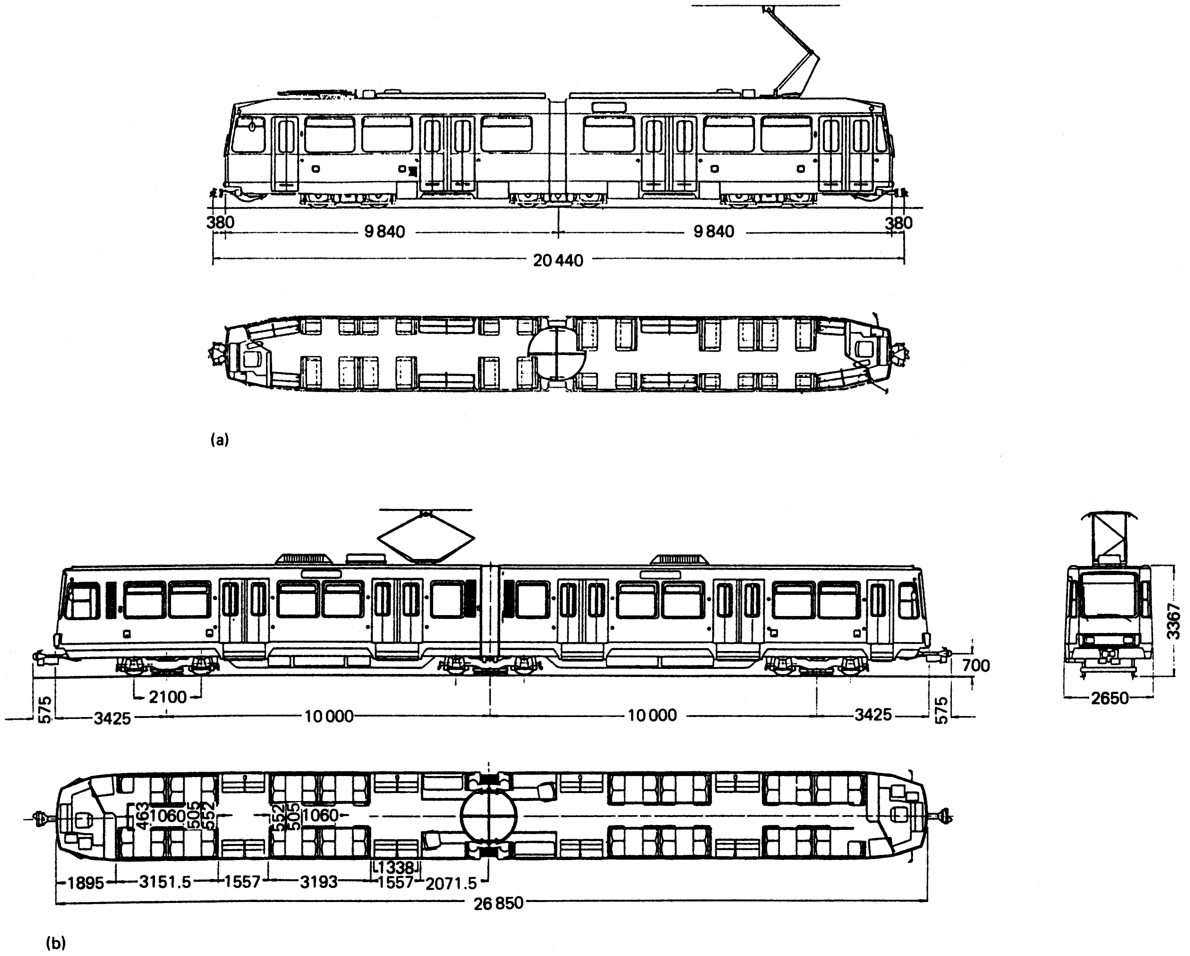


Figure 44.24 (a) Duewag N6 light rail vehicle; (b) Duewag B80 light rail vehicle. Dimensions are in millimetres

Table 44.1 Typical light rail vehicle performance

Maximum speed	80 km/h
Service speed	70 km/h
Initial acceleration	1.1 m/s^2
Service braking	1.2 m/s^2
Emergency braking	$\simeq$3 m/s^2
Minimum curve radius in service	20 m
Minimum curve radius in depot	15 m
Maximum gradient	$\simeq$6%
Maximum gradient (all axles motored)	$\simeq$10%

widths limit the seating arrangement to 2 + 1. Although the greatest width of 2.65 m is larger than that normally permitted for buses (2.5 m), the fact that the light rail vehicle is guided means that the swept-path needed is no greater than for a bus.

The height from rail to roof ranges between 3.0 and 3.3 m, although in some designs roof-mounted equipment may increase this to over 3.6 m.

Empty weights range from under 30 t to between 40 and 50 t.

44.2.3.3 Passenger capacity

Modern light rail vehicles have nominal capacities ranging from 100 to 250 passengers of whom about one-third would be seated. Traditional light rail vehicles have floor levels about 800 mm above the rail which require the use of high platforms or retractable steps to permit access from street level. High platforms can be difficult to incorporate in existing urban environments and steps present difficulties to the mobility impaired. In the Manchester system light rail vehicles stop with one door alongside a short length of raised platform to provide this access in the street running sections.

The interior of modern vehicles is laid out to accommodate the needs of the mobility impaired with provision for wheelchairs and carefully positioned handles and grabrails.

Several manufacturers, including BN, Breda, Duewag and MAN, have developed low-floor vehicles with floor heights between 300 and 350 mm above the rail. These range from conventional vehicles with one section of floor dropped to those providing low floors over their entire length. The latter is generally only achieved at the expense of significant mechanical complication in the design of the bogies and positioning of traction motors, often requiring additional drive shafts and gears. *Figure 44.25* illustrates the prototype vehicle developed by MAN for Bremen.

44.2.3.4 Structure

Traditionally, body shells are of light-weight welded steel design in which cable and air ducts, equipment boxes, longitudinal, articulation and end members are welded together to form an underframe. Roof and side-wall frames, fabricated of rolled and bent profiles, are welded together with side-wall sheeting and the underframe to form an integrated structure.

Light alloy construction is also used employing large size extruded sections with pressed on supports for equipment for the side walls, roofs and floors to form an integrated structural unit.

44.2.3.5 Traction and braking equipment

Vehicles are propelled by electric motors which obtain power from an overhead conductor wire via a roof-mounted pantograph.

Most modern light rail vehicles employ d.c. traction motors with d.c.-chopper or a.c.-inverter control both with microprocessor control, although some installations use a.c. induction motors with three-phase inverters.

The motors may be self-ventilated or force ventilated with armatures and field windings insulated with class F or H materials.

Braking is achieved by a combination of rheostatic or regenerative braking by the traction motors combined with friction brakes. For street running, electromagnetic track brakes are fitted to provide an emergency braking rate of about 3 m/s^2. Parking brakes are also fitted. Traction and braking systems are integrated to provide:

(1) slip/slide protection;
(2) blending of regenerative or rheostatic brake and friction brake;
(3) control and protection of power electronics; and
(4) the facility to operate coupled vehicles.

44.2.3.6 Bogies

Two axle bogies are used. The bogie frames are usually fabricated from steel in a box section design. The frame is supported on the axle boxes by the primary suspension which may consist of chevron-type rubber springs which allow radial adjustment of the wheel sets or of leaf-guided helical or rubber spring elements. Secondary suspension, between the bogie frame and the vehicle body may be by helical, rubber or air springs or by a combination of these. Although the most complex, air springs offer a high ride quality and can maintain the vehicle floor at a constant height above rail level at stops which facilitates level access for mobility-impaired passengers.

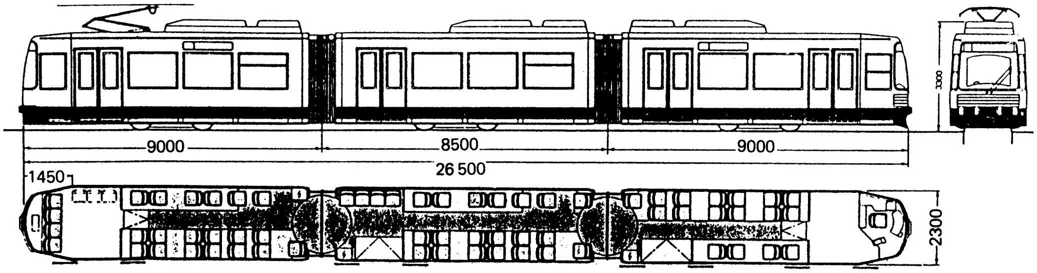


Figure 44.25 MAN low-floor light rail vehicle. Dimensions are in millimetres

There are two general arrangements for motor bogies. In the mono-motor layout a single motor is positioned in the fore-and-aft direction with drive shafts at both ends connected by right-angle gears to the axles. The traction motor and gears form a single unit which is suspended either on the bogie frame or, by means of rubber couplings, on the wheelsets. In the bimotor layout a separate motor is provided for each wheel set and is either hung from the axle or suspended from the bogie frame.

The mono-motor design offers the advantages of lower weight, simpler maintenance and lower cost because of its greater simplicity. A further advantage is that because both axles are connected together wheels cannot begin to spin or slide until all four wheels have lost adhesion. A disadvantage has been that because of restrictions on available space the size, and hence power, of the motor may be limited although recent advances in the design of compact motors has alleviated this to some extent. There is also a suspicion that mono-motor bogies may contribute the phenomenon of rail corrugation in certain circumstances because all wheels are mechanically connected, although the causes of rail corrugation are not well understood.

In addition to rheostatic or generative braking using the traction motor, mechanical brakes are fitted. These may take the form of clasp brakes operating on the wheel treads or disk brakes. Both types are of the spring-applied, power-released type.

Figure 44.26 shows the bogie design used on the MAN low-floor vehicle. The motor is body-mounted and connected to the bogie by a shaft drive with universal joints. Another shaft transmits the power to the two driven wheels because axles are omitted in order to lower the floor.

44.2.3.7 Auxiliary equipment

Folding doors are commonly used which are operated electrically or pneumatically. Retractable steps provide access from street level. Their operation is interlinked with the doors and it is possible to arrange for the steps to operate in different combinations to deal with varying platform heights on different parts of a network. Heating and ventilation are normally fitted.

Automatic couplings can be fitted which will make all necessary mechanical and electrical connections automatically on buffering up two vehicles. Pneumatic uncoupling and central locking can be provided. Where vehicles are not normally operated coupled, provision is made for manual coupling to enable a failed vehicle to be towed or pushed away.

Pneumatic sanders can be fitted for street operation or where steep gradients are encountered.

Auxiliary electrical power supplies are normally fitted for controls, lighting, battery, charging and other purposes. In addition a 24 V d.c. supply is available for control circuits, public address and radio systems, etc. The battery is capable of maintaining emergency lights, public address and radio systems, and marker lights together with local control equipment for a period not less than 1 h.

Miscellaneous equipment normally fitted includes: a public address system to enable the driver to speak to passengers; passenger alarm to allow passengers to attract the driver's attention; radio equipment to enable the driver to communicate with the controller; route and destination displays; and ticket cancelling or validating machines.

External lighting is fitted generally in accordance with the 'Construction and Use of Public Service Vehicles', but taking account of the particular requirements of light rail vehicles. Additional lights may be fitted for operation on fully segregated sections.

44.2.4 Trackwork

On fully segregated sections conventional railway track is generally used. This consists of rails and fixings, generally with sleepers laid on ballast, although it may be convenient in certain areas to fix the rails direct to the supporting structure.

On street-running sections grooved rail is used, set flush into the road surface. The extent of the swept path of the vehicles is indicated by using a different surface texture or colour, or by white lines to highlight the edges as necessary. There are several ways to install the rails in street-running sections. A recently developed method involves providing grooves in the surface. The rails are correctly located to line and level in the grooves before an elastomer material is poured around them. When set this material holds the rails in position yet provides a slight degree of resilience which helps reduce the noise and vibration generated as the vehicle wheels roll over the rail.

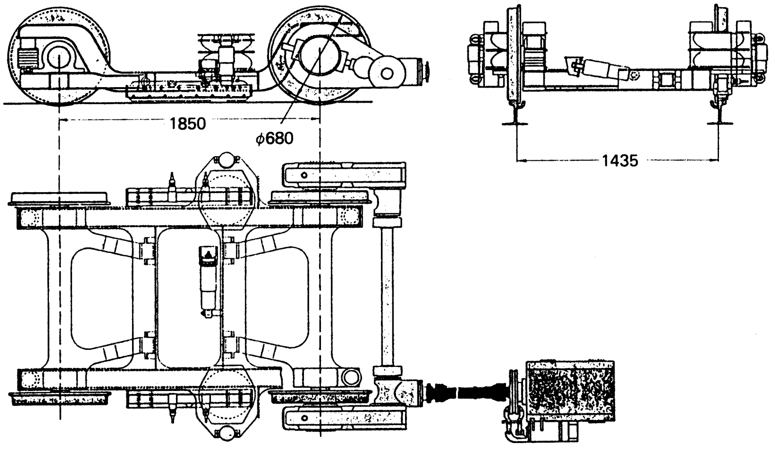


Figure 44.26 MAN low-floor bogie. Dimensions are in millimetres

Special care must be taken to minimise the effect of stray or leakage traction return current because of the use of d.c., especially in the urban areas. The rails themselves form the immediate return path from the vehicle and thus whatever method of fixing the rails is employed it must provide adequate electrical insulation. The rails themselves are often bonded at frequent intervals to a separate traction return cable. Reinforcing mesh or bars in any concrete slab under the track need to be bonded together and to the traction return cable. In some circumstances it is prudent to insert insulating sections in adjacent water or water pipes, especially those entering buildings, to prevent their being used by stray currents.

Where the line runs in a roadside or other reserve, grassed track can be used. In this, the rails are set flush with the level of the ground which is then grassed between the rails as well as outside them which greatly reduces the visual intrusion.

44.2.5 Power supply

Electrical power is normally obtained from the local electricity supply company; in the UK this is normally at 11 kV a.c. Substations comprising transformer/rectifier units convert this to a lower voltage d.c. This is supplied to the overhead system at a voltage not exceeding 750 V d.c. on the street-running sections, although 1500 V d.c. may be used on segregated sections of the line.

The visual impact of the overhead-line equipment is greatly reduced by the use of insulated support wires made from Parafil which eliminates the need for obtrusive insulators. Wires can also be attached to existing buildings to obviate the need for separate poles. The minimum wire height on street-running sections in the UK is 5.5 m in order to provide safe clearance above other road traffic, but may be reduced to about 3.8 m on fully segregated sections.

44.2.6 Signalling and control

Signalling and control systems permit safe and expeditious operation, while requiring the minimum of manual intervention in normal operation. In street-running sections separate heads are fitted to traffic signals to control the light rail vehicles. These are frequently used to give priority to the light rail vehicles as a means of encouraging the use of public transport. Apart from this, the vehicles are driven like any other road vehicle, with the driver being responsible for maintaining a safe speed.

A form of automatic identification system (such as the Phillips VETAG system) is often fitted to the vehicles. Information about the vehicle's route may be pre-set so that on approaching junctions the required route may be set automatically and the vehicle given a degree of priority over other traffic. Similarly, line-of-sight operation may be used on partially or fully segregated sections with drivers being responsible for maintaining a safe speed without the need for an elaborate signalling system. Where signalling is needed, at say termini, then a simple, local system can be used.

Mobile radio is normally fitted to the vehicles to enable drivers to remain in contact with a central control room. Staff in the central control room are responsible for the overall supervision of the operation and have good, direct communications to the police, fire and ambulance services for use in case of emergency.

44.2.7 Stops

Stops are simple and unmanned. They basically consist of a raised platform to provide stepless, level entry to the vehicles. If low-floor vehicles are used then on street-running sections the platforms need only be raised some 200–250 mm above the level of the pavement when the height of the kerb is taken into account. Simple passenger shelters only are needed which can also be used to support passenger information notice boards and ticket vending machines selling adult and reduced rate single tickets. Passengers are generally encouraged to purchase multi-journey tickets (*carnets*), fare cards or reduced-rate tickets off the system.

44.2.8 Depot

The depot provides maintenance facilities for the system and stabling siding for the vehicles. The depot would be equipped to carry out day-to-day servicing and maintenance of vehicles, power supply, signalling, trackwork, etc. Depending on the scale of operation, it may also be equipped to carry out more extensive maintenance work, although it may be found more economical to subcontract certain heavy maintenance of vehicles. One example is turning worn wheels. A wheel lathe for turning wheels *in situ* on the vehicles costs over £1 million. Facilities provided at the depot typically include: office building; control centre; staff facilities; tracked workshop; general workshop; storage areas for equipment, plant and spares; stabling sidings; washing plant; and external stores compounds.

44.2.9 Economics and investment

In many other countries, public transport is funded by special taxes. In Germany there is a tax on petrol which is designated for roads and public transport, and in France local communities may decide to introduce a payroll tax on employers to fund public transport. In the USA there are a number of State-operated taxes (such as sales taxes) which are used to generate funds. In the UK grants towards construction costs are available from public funds under Section 56 of the Transport Act 1968—half coming from central government and half from local authorities. The grant is only available to fill any gap between the capital cost and the funds available from other sources and must be justified by the economic benefits the system offers to non-users. Primarily this is in terms of time and cost saving to road users who do *not* use the light rail system.

Light rail transit systems can claim decongestion, and social and environmental benefits which, if quantified, would make the schemes economically viable in cost–benefit terms. Nevertheless, no system recovers its full cost through fares, although *direct* operating costs which exclude any element of capital charges can be covered. The capital cost of light rail transit varies widely with the type of system, but typically it can vary between £2 million and £3 million per single track kilometre. Similarly, direct operating costs can also vary widely, but in the UK have been estimated to range between £2 and £4 per vehicle kilometre.

44.3 Battery vehicles

For many years, battery-powered vehicles have been used quite extensively in the UK for low-speed, low-range, applications. These vehicles, which use well-proven technology, are described in Section 44.3.1.

Recently, there has been a surge of activity in the development of faster, longer range vehicles, with performance similar to combustion-engine-driven types, and capable of use in commercial fleets or for personal commuting. Such advanced vehicles are not yet fully developed for commercial production, but are expected to be so in the second half of the 1990s. Their current status is described in Section 44.3.2.

44.3.1 Low-speed battery vehicles

The main sphere of use is for local delivery work such as the door-to-door delivery of milk, bread, laundry, coal, mineral waters and other goods, and specialised work such as refuse collection, steel lighting maintenance and interworks transport. In such services a vehicle may be required to make 200–300 delivery stops, while the total distance is usually between 30 and 50 km daily, and rarely reaches the maximum range of about 70 km. Under these conditions the delivery speed and the road speed are almost unrelated, the chief considerations being acceleration, ease of exit and entry to the vehicle, simplicity of control and reduction of personal fatigue. Efficient service is generally obtained with maximum road speeds of the order of 30–35 km/h.

44.3.1.1 Batteries

The lead–acid battery is customary. Choice of battery depends on vehicle duty. Thin-plate automotive starter batteries have high energy capacities but short life on deep-cycling duty. Traction batteries have thicker and firmly separated plates: their weight is kept down by using lightweight plastics cases and short intercell connections. 'Light traction' batteries are supposed to withstand 500 charge/discharge cycles, whereas true traction batteries are guaranteed for 1000–1500 cycles, and may last far more. These batteries are most economic (in £/kW-h throughput) if they are used almost daily and discharged to between 70 and 90% of their capacity.

44.3.1.2 Motors

The basic design of d.c. traction motors has not been essentially changed, but the use of improved wires and strips for windings, and class H insulation, have greatly improved thermal transfer and reduced weight.

44.3.1.3 Control

Until the introduction of semiconductors with high current capacity the accepted methods of speed and torque control were (a) series resistance, (b) series/parallel battery switching, (c) series/parallel motor switching and (d) field control. Almost all controllers provided a stepped variation or, at best, a smoothly variable control over limited ranges of torque and speed. The carbon pile variable resistor was also extensively applied, with compression by hydraulic master and slave cylinders.

Series resistance On small vehicles such as milk prams a single step of resistance usually suffices before the motor is switched on to the full battery, usually 24 V. On larger vehicles and works trucks several steps of series resistance may be used and these are cut out in sequence by a drum or cam controller. The one economical running connection is usually sufficient for low-speed vehicles of the pedestrian-controlled type. In some cases a second economical connection is obtained by field diversion.

Parallel/series battery switching By dividing the battery into halves, and putting them first in parallel to give half-voltage, then in series to give full voltage, and by using two or three resistance steps for each setting, the loss of energy in the resistors may be halved, compared with the previous system. In addition, an economic half-speed running circuit is obtained. The scheme is also less severe on batteries, since the first few current peaks are shared.

A conventional battery switching circuit employs a pair of contactors interlocked (usually mechanically) so that the double-pole contactor (*Figure 44.27*(*a*)) is closed for the parallel and the single pole for the series connection: they cannot be closed at the same time. In a circuit which involves no current breaking (*Figure 44.27*(*b*)) diodes make the parallel connections, and the contactor for the series connection reverse-biases the diodes so they do not conduct.

A patented diode/contactor arrangement elaborated from the above gives a multivoltage circuit in which, by contrast with the more obvious battery tapping systems, all batteries are used (not necessarily equally) all the time.

Field control Field control is fittingly used in addition to the two previous schemes. By use of a four-pole motor with its field windings in separate pairs, full excitation is obtained with the field windings in series, and half-excitation for increased speeds with the field winding pairs in parallel. Field diversion, for even weaker field and higher speed, may be given by shunting an appropriately low resistance across the paralleled field windings.

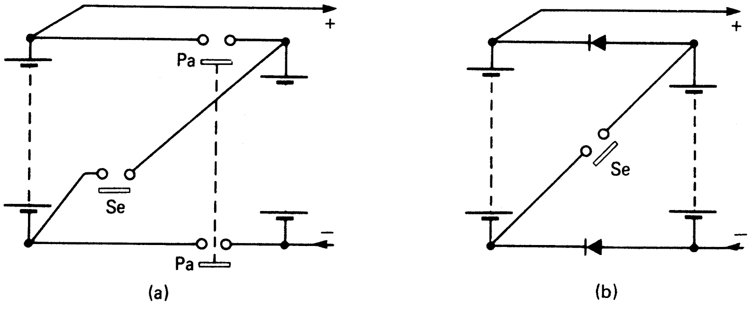


Figure 44.27 Parallel/series battery switching: (a) contactor; (b) contactor/diodes

Solid state switching With suitable commutation control equipment, thyristors can provide a substantially variable, loss-free motor power control, provided that the pulse rate is not too low. Several forms of pulsed thyristor control are available, differing only in circuit details. The variations are largely associated with the method of providing satisfactory turn-off characteristics for the main circuit thyristors, and correlation of mark/space, repetition frequency and current switching levels to obtain the required motor output characteristics.

Thyristor controllers consist of the following basic elements: (a) main power thyristors, (b) turn-off capacitor; (c) commutator thyristor, (d) function generator, (e) current limiting and fail-safe circuits, and (f) manual control device.

Chopper control In the simplified equivalent circuit shown in *Figure 44.28*, the switch (Sw) represents the main thyristor in an actual circuit arrangement. In (*a*) for normal motor operation, Sw is closed for a time t_1 and opened for t_2 (each time being of the order of a few milliseconds), the sequence being repeated with a switching time $T = t_1 + t_2$. During t_1 the current rises at a rate proportional to the difference between the battery voltage V and the motor armature e.m.f. E, such that $L\ (\mathrm{d}i/\mathrm{d}t) = V - E$, where L is the circuit inductance and resistance is neglected. With Sw now opened for time t_2, the motor current continues through the freewheel diode D_1, driven by the armature e.m.f. E and falling at a rate such that $L(\mathrm{d}i/\mathrm{d}t) = E$. Under steady load conditions the rise Δi during the 'on' time is equal to the fall Δi during the 'off'. As the times are short, then approximately

$$\Delta i = t_1(V - E)/L = t_2 \cdot (E/L) \Leftarrow$$

whence $E/V = t_1/(t_1 + t_2) = t_1/T$. The e.m.f. ratio is equal to the ratio between 'on' time and total switching period. For an average motor current i the mean battery supply current is $I_s = i(t_1/T)$ and the mean motor current is $I_m = I_s(T/t_1)$. With $E = V(t_1/T)$ it follows that

$$EI_m = V(t_1/T) \cdot I_s(T/t_1) = VI_s$$

whence the chopper acts like a 'd.c. transformer' in terms of the input and output voltage and current ratios, under steady-state conditions and with resistance and other losses neglected.

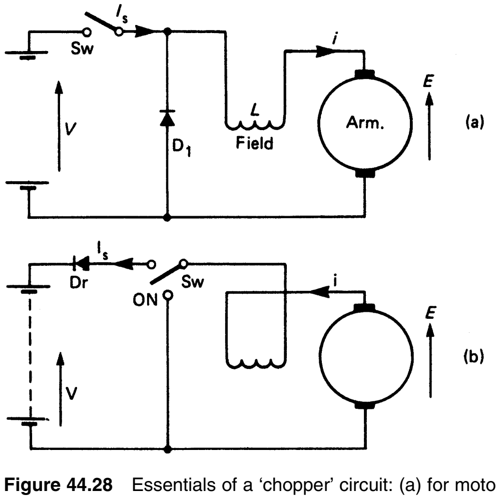


Figure 44.28 Essentials of a 'chopper' circuit: (a) for motoring; (b) for regeneration

In operation, t_1/T is increased from zero towards unity to start the motor (with control by the current or voltage limiter). With $t_1/T = 0.95$ or thereabouts, t_2 may be too short to allow the thyristor current to quench. In many equipments a contactor is closed to bypass the thyristor and connect the battery direct to the motor, thus eliminating thyristor forward loss.

If sustained low-voltage high-current operation is required (and it is in such a case that the efficiency of the system is much higher than with resistance control), the mean square motor current is higher than the square of the mean current, which raises the motor I^2R loss. Enhanced cooling or reduced rating are then necessary in chopper control.

For *regeneration* the basic circuit is rearranged as in *Figure 44.28*(*b*). Diode D_r prevents feedback of the supply to the motor. When Sw is 'on', the machine generates, converting kinetic energy into magnetic energy in the inductance; in the 'off' time the machine is connected to the supply, and E together with $L(\mathrm{d}i/\mathrm{d}t)$ drives current through D_r into the battery. Then $V/E = T/t_1$ and $I_s = I_m(t_1/T)$.

For *control* the function generator in the chopper system is combined with comparators to give current and voltage control, switching off when current or voltage attains a demand value. Both values are increased as the 'accelerator' pedal is depressed. Regeneration, if fitted, is usually current controlled by the initial movement of the brake pedal, which thereafter applies the friction brakes.

The flexibility of electronic control systems allows the use of a controlled field shunt motor, with electronically controlled field current. Full-field acceleration to the running voltage, with power thyristor control (as above) for the armature current, is followed by controlled field weakening to give, e.g. constant current motoring to a preset speed or the weakest allowable field. Regeneration by field strengthening is then easy to arrange. So far, schemes of this nature have been experimental.

As an example, *Figure 44.29* shows in more detail the scheme for chopper control of a battery electric car.

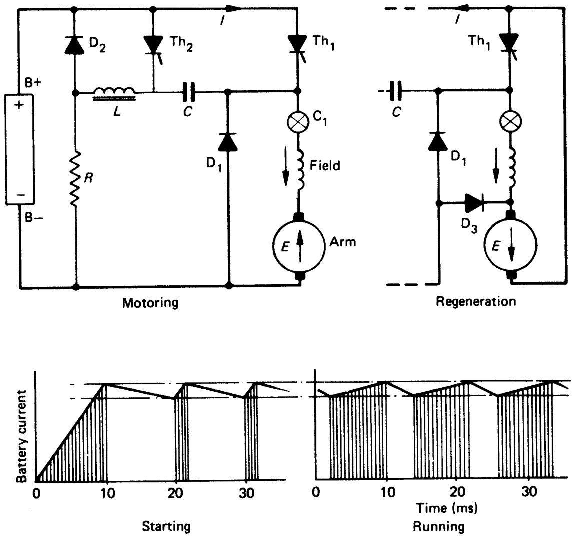


Figure 44.29 Thyristor control of an electric car

For *motoring* the battery motor circuit is completed by firing thyristor Th_1. The current rises, and at a predetermined value is cut off by the switching action of the current monitor C_1. The inductive energy maintains the motor current through the free-wheel diode D_1. The current decays, and at a predetermined minimum C_1 switches on Th_1. The motor current fluctuates between the two limits, but battery current flows only while Th_1 conducts. The rate of current fluctuation depends on the armature e.m.f. To switch off Th_1 it is necessary to reduce its current momentarily to a very small value by means of Th_2, C, L, R and D_2. Prior to any current demand, Th_2 is switched on and charges capacitor C so that the potential of the left-hand side is raised to that of the positive battery terminal B+⇐and the right-hand side to that of the negative B−: then Th_2 ceases to conduct owing to lack of 'holding' current. When Th_1 is fired, the right-hand side of C is immediately raised to B+⇐and the left-hand side to 2B+. Current starts to flow from C through inductor L and diode D_2, reducing the voltage across C and increasing the inductive energy in L. When the left-hand side of C has fallen to B+, the current in L continues, which causes a continued fall in the voltage of C. When the current in L has ceased, the left-hand side of C will be at a potential B−⇐and diode D_2 becomes reverse-biased. The potentials on C are now B−⇐left hand) and B+⇐ (right hand), and remain so until Th_2 again fires. A bleed path through R ensures that leakage currents through D_2 and Th_2 do not raise the left-hand voltage on C when Th_1 conducts for long periods. When Th_2 is fired again, the left-hand side of C is taken up to B+⇐and the right-hand side to 2B+. The cathode of Th_1 is then positive to its anode and it therefore ceases to conduct. The motor current, which is thus transferred from Th_1 to Th_2, flows through C and lowers the right-hand potential to B−. Then, as the motor current diverts through D_1, conduction through Th_2 ceases for lack of holding current. The inductive energy of the motor maintains current through D_1. The circuit is now ready for Th_1 to be fired again.

For *regeneration* the circuit is set up by changing the armature polarity with respect to the field and by adding diode D_3, accomplished by a simple changeover switch. When Th_1 conducts, the motor is effectively short-circuited and the current builds up rapidly. At maximum current level Th_1 is switched off and the motor current, maintained by the motor inductance, diverts through D_1, supplying a charging current to the battery. The cycle repeats to give a roughly constant average motor current and a constant resulting braking torque.

Conditions are complicated by the transition 'spikes' of voltage. They have a short duration but a large magnitude. The spikes have some effect on the battery, and increase motor core and I^2R losses. They may also impair commutation, particularly at low pulse rate frequency.

The cost of pulsed thyristor equipment is a disadvantage. The reduction of battery energy depends on the type of vehicle and its duty. The overall saving may be about 10% for typical delivery rounds, and rather more for industrial vehicles such as fork-lift trucks.

44.3.1.4 Vehicle and operational details

Charging Taper chargers are now used extensively for the charging of traction batteries, and are normally provided with germanium or silicon rectifiers. Voltage sensitive thermal relays in conjunction with synchronous timing motors and contactors are used for charge termination. In one form of charge control device a temperature sensitive non-linear resistor, preheated electrically, is cooled by hydrogen from pilot cells when gassing begins. The change in resistance through the cooling of the device causes transductors to regulate the output of the charger.

Improved conductor insulation and high-grade magnetic core material, together with forced cooling, have resulted in weight saving in charger equipments carried on vehicles. A typical charger for connection to a 13 A power outlet, and making full use of this current rating up to the start of the gassing period, has the performance shown in *Figure 44.30* when charging a 48 V, 260 A-h lead–acid battery.

Torque transmission Most road and industrial battery vehicles are fitted with axles of conventional differential gear design, still the most convenient form of torque transmission to a pair of driving wheels. It has been shown by test that there is scope for improvement in performance by modifying the lubricant viscosity: the axles are normally fitted with extra gear reduction nose-pieces or external chain-and-sprocket reductions, and a fully run-in differential may absorb up to 1.5 kW at maximum road speed. Oils of lower viscosity have been used without excessive gear wear, although the noise is greater.

Hydraulic transmission has found only limited application. Even simple forms cannot compete as alternatives to thyristor motor control gear for driver-type vehicles, but may be suitable for some industrial applications where the infinitely variable characteristic makes elaborate motor control gear unnecessary.

Motorised wheels (i.e. motors built into the wheels) are extensively used in fork-lift trucks, sometimes with low-speed designs of reduced length and increased diameter. The powered wheels can be used for steering, thus improving manoeuvrability.

Regeneration In general, the small amount of energy recoverable does not warrant the cost and complexity of the control gear necessary for regeneration, although it reduces brake maintenance (a major operating requirement). Certain industrial vehicles such as fork-lift trucks might benefit from regeneration if enough load lowering duty were involved; this might be possible by use of recirculating ball screws in place of the normal hydraulic lifting rams, enabling the motor to generate on lowering.

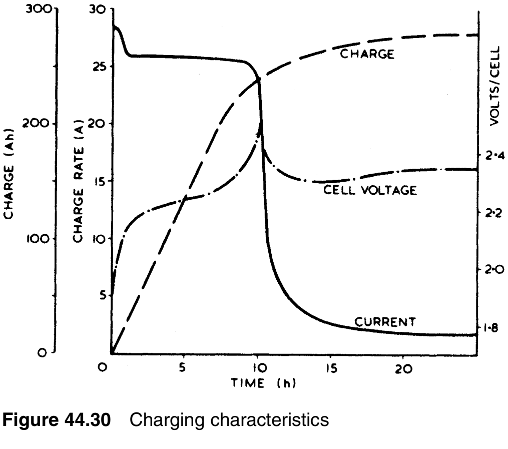


Figure 44.30 Charging characteristics

Tyres Battery vehicles are affected to a greater extent than other vehicles by the rolling resistance of tyres. Radial ply tyres provide a more supple casing and as a result the loss due to flexing of the walls is less. A saving of 13% in energy consumption (or an increase of 15% in range) can be obtained. An advantage is that radial ply tyres are less sensitive to departures from optimum inflation pressure.

Bodywork Most recent rider–driver battery vehicles use glass-fibre-resin mouldings, which permit improved appearance with much lower tooling cost. Metal inserts are readily moulded into laminates to provide anchorage points or attachments for components. Translucent panels can be moulded into otherwise opaque surfaces.

44.3.1.5 Applications

Pedestrian controlled vehicles Bakery and dairy delivery vans, 'led' by the driver, have capacities between 0.5 and 1 t and a range up to about 20 km. The steering handle carries running and brake controls and there is a separate parking brake. A typical specification includes a 2 kW motor, 24 V, 100 A-h battery, and chain or reduction gear differential drive.

Rider–driver vehicles Invalid carriages with 36 V, 75 A-h batteries have speeds up to 20 km/h and can negotiate any normal road. Municipal vehicles (refuse collection, tower wagons and tractors) are usually 4–6 t vehicles with batteries of up to 144 V, 600 A-h. The battery weight is useful in adding stability to the tower wagon. The most widely used industrial trucks are platform, elevating and fork-lift trucks, road and rail tractors, mobile cranes, and trucks for carrying hot forgings and molten metal buckets. The capacities range up to 1–2 t. Some lighter types are three-wheelers having the motor and front wheel mounted on a turntable, which makes them exceptionally manoeuvrable.

On elevating, tiering, crane and other types of trucks with power operation, it is usual to employ a separate motor to perform these tasks, power being drawn from the main battery. The location of the battery on the vehicle varies according to the design. On some they are mounted beneath the platform, but in many cases they are mounted on a separate platform or under the driving seat. It is common for these vehicles to work long hours and in such cases two or more sets of batteries are employed, a discharged battery being replaced when necessary. The voltage varies between 24 and 80 V.

When trucks are fitted with a platform for the driver to stand on, steering is usually by means of a tiller, while the controller is hand operated and interlocked with the brake pedal to give a 'dead man' effect. With this design the brakes are held off by brake pedal. When pressure on the pedal is released, the brakes are automatically applied and the electric circuit is broken.

Locomotives Battery shunting locomotives are used by many railways, factories and mines. A typical small locomotive develops a drawbar pull of 1.6 kN (360 lb-ft) at 6.5 km/h from a 4 kW, 48 V totally enclosed series motor with split field windings for control and worm drive on each of the two axles. Such a locomotive is designed for 0.46, 0.51 or 0.61 m gauge. Larger locomotives may weigh up to 8 t, operate at 120 V, and have two 12 kW series motors with series/parallel control, giving a 1 h tractive effort of 11.5 kN at 8.5 km/h.

44.3.1.6 Electric buses

Two experimental vehicles of the early 1970s demonstrated the feasibility of battery buses. One was a 50-passenger single-deck vehicle, 16 t in weight (including a 4.8 t battery) with chopper control (with regeneration of a 72 kW, 330 V series motor, 64 km range, 64 km/h speed and 1.0 m/s^2 acceleration). The other was a 'minibus' for 34 passengers in a short single-deck body, with a 100 kW motor, 150 km cruising range, 80 km/h maximum speed, 0.9 m/s^2 average acceleration and weight 9.7 t.

From the mid-1970s, several experimental fleets of electric and duo-mode buses were established in Germany as part of a programme initiated by RWE, Germany's largest privately owned electric utility company. The projects included a fleet of MAN battery electric buses, operating in Münchengladbach and Düsseldorf. The buses were originally designed for rapid battery exchange, to permit extended operation, but were later equipped with rapid recharge facilities at selected bus stops. Although the range on a single charge was only 60 km, these buses were able to achieve 360 km in a 16 h working shift due to 'opportunity charging'. Other experimental bus fleets were established in France by Renault.

More recently two prototype purpose-designed electric buses with 16-seat capacity have been built by a British company for clients in the USA and Hong Kong, and two 25-seat passenger vehicles have entered service in Santa Barbara, California, as part of an integrated pollution-free pedestrian/public transport scheme in the downtown area.

44.3.1.7 Range and power assessment

The range is limited by the energy/mass storage capability. A 1000 kg lead–acid battery operating at a load of 10–15 kW/t stores about 20 kW-h, and as the transmission efficiency from motor terminals to wheel treads is about 70%, less than two-thirds of the capability is available at the wheels.

A vehicle of weight G with a (battery/total) weight ratio f, driven against a tractive resistance R, has a range S that is roughly estimated from

$$S = 50(fG/R)$$

Let a vehicle have the following particulars:

G	total weight (kg)	k_r	rolling resistance coefficient (N/kg)
u	speed (km/h)		
R	tractive resistance (N)	r_a	air resistance (N)
A	frontal area (m^2)	k_a	drag coefficient
r_r	rolling resistance (N)		

The tractive resistance R comprises the rolling resistance r_r and the air resistance r_a. The former is $r_r = k_r G$, where k_r has a value in the range 0.1–0.2. The air resistance, proportional to the square of the speed, is

$$r_a = k_a A u^2/21$$

The drag coefficient k_a is 1.0 for bluff-fronted vehicles (e.g. vans), about 0.5 for cars and as low as 0.2 for fully streamlined bodies. The total tractive resistance is then $R = (r_r + r_a)$, and can be used to estimate the range S (in km).

Let $G = 8000$ kg, $f = 0.25$, $u = 40$ km/h, $A = 5.0$ m^2, $k_r = 0.15$ and $k_a = 0.78$. Then $r_r = 1200$ N, $r_a = 300$ N and $R = 1500$ N. Then the range $S = 67$ km.

The power at the wheels (in kW) is $P = Ru/3600$. For the parameters above, $P = 17$ kW. The input to the motor is about $17/0.7 = 24$ kW. Motor inputs of 8 kW/t may be required for a battery vehicle to emulate an internal combustion-engined vehicle.

If a vehicle is braked to rest from an appreciable speed, the range S is reduced by reason of the loss of kinetic energy in friction. Some allowance may be made by reducing S by about 0.3 km for each brake stop from 50 km/h, and less for lower speeds in proportion to the square of the speed.

44.3.2 Advanced electric vehicles

At the beginning of the 1990s there was a surge of activity world-wide in the development of advanced electric vehicles. The driving force for this activity was mainly concern over atmospheric pollution, particularly in urban areas, and the concern in Western countries about the security of oil supplies, highlighted by the Gulf War of 1991. Increasingly stringent legislation was introduced to limit exhaust-pipe emissions, particularly in the USA where the introduction of zero emission vehicles before the end of century was mandated. As a result, practically all major vehicle manufacturers started development programmes for battery operated vehicles.

44.3.2.1 Advanced batteries

Lead–acid batteries Modern lead–acid batteries specifically developed for electric vehicle application operate at defined conditions (temperature control, no acid stratification, battery management) and yield between 500 and 1500 duty cycles. Energy density is limited to about 30 W-h/kg, and at the 1–2 h rate this often falls below 25 W-h/kg. Valve-regulated lead–acid batteries employing gelled electrolyte allow total maintenance-free operation.

Nickel–cadmium batteries Nickel–cadmium batteries at the 1–2 h rate yield twice the energy density of lead–acid batteries, i.e. about 50 W-h/kg. They operate over a wide temperature range, give approximately 2000 cycles, and can be charged in less than 1 h. Nickel–cadmium batteries are used in various developmental electric vehicles. Unfortunately, the high cost of the electrode materials leads to prices 3–4 times higher than lead–acid, only partly offset by longer life.

The presence of cadmium poses a problem with disposal, and development work is in hand to replace the cadmium electrode with a hydrogen-storing electrode, to give the so-called nickel–metal hydride battery. Further developments will lead to a sealed, fully maintenance-free, version.

Sodium–sulphur batteries A host of other battery technologies are under development, including zinc–bromine, nickel–iron, and various lithium systems. However, it is the sodium sulphur system which is best developed and is now in pilot production by two major European groups. Sodium–sulphur batteries operate internally at temperatures above 300°C and require high-performance thermal insulation. However, they offer energy densities around 100 W-h/kg, fully maintenance-free operation, 100% coulombic efficiency, fast recharge times, good deep discharge performance, and the potential (yet to be proven) for long cycle life and costs similar to lead–acid batteries.

Together with a similar system, sodium–metal chloride, these batteries have been tested in a range of different vehicles in Europe and the USA, and are expected to be in volume production by the mid-1990s.

44.3.2.2 Advanced drive systems

Although d.c. traction motors are still widely used, some vehicle manufacturers are now developing their own a.c. systems which give higher power density. In some designs the motors drive the wheels directly, in others the motor is incorporated into the axle, or drives through a conventional differential.

Advanced electronic controllers are under development which interface with the battery electronics to provide overall system management.

44.3.2.3 Applications

As the technology of electric vehicles improves, the emphasis is switching from commercial delivery vehicles towards personal transport. One US vehicle manufacturer has demonstrated a battery vehicle which achieves 0–60 mile/h in 8 s and has a top speed of 110 mile/h. Several European manufacturers have established small fleets of vehicles, often using conversions of existing models as a 'mule' vehicle to test out the drive system. Unless there are unforeseen problems, it can be expected that a new generation of electric vehicles will be on general sale by the middle of the 1990s.

44.4 Road traffic control and information systems

44.4.1 Traffic signalling

44.4.1.1 Signals

Traffic signal heads usually have two aspects for pedestrians, or three for vehicles with additional aspects for turning movements. Standards exist for colour, light output (typically 400 cd over 30° for good daytime visibility), and sun phantom (the permitted re-emitted level of sunlight).

The aspects are switched on and off at the signal controller via multicore cables, usually at mains voltage. Low-voltage halogen lamps fed by transformers in the signal head are standard in the UK, but mains voltage tungsten lamps are also widely used. High-intensity signals are dimmed at night.

Possible developments include digitally controlled distributed switching having the necessary level of integrity, and solid-state light sources having acceptable colour and intensity.

Primary signals are located at the stop line, on each side of the approach if necessary. Secondary signals are located downstream of the stop line, either on the nearside or the farside of the intersection according to safety considerations.

44.4.1.2 Phasing

Where signal heads control movements which can be started and stopped at the same time they are wired in parallel to a set of controller switches, and form a unit of control known as a 'phase group' or 'signal group'. Phases may be sequential or, providing they do not control conflicting movements, overlap.

To obtain the greatest capacity from the physical intersection space available as many non-conflicting movements are run together as possible, and the sequence of phases is optimised to minimise lost time between movements, taking account of any moves which are prohibited on safety grounds. A simple intersection may have about eight phases and a complex one 24, or more where trams have to be controlled for example.

Where an overall coordination plan is to be followed, for example from a central computer, it may not be necessary to control each phase individually, and phases are grouped into sequential stages such that a new stage is defined whenever a new combination of phases is needed. A normally complete sequence of stages is called a cycle, but stages may be omitted or their order varied according to traffic conditions. Several parallel stage streams (e.g. eight) may be defined in a modern single controller.

A simple three-stage four-phase intersection is shown in *Figure 44.31.*

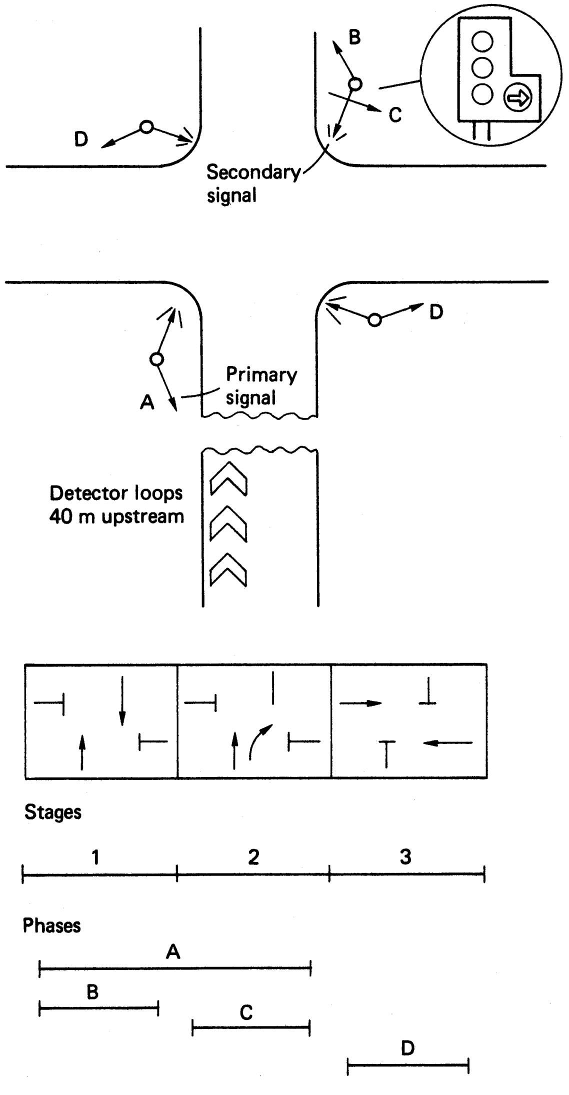


Figure 44.31 Simple three-stage four-phase intersection

44.4.1.3 Timing

Under vehicle or traffic actuated control, a phase is only called if a demand is received from a vehicle detector, unless it is a phase which must appear. For that phase to begin, conflicting phases must be safely terminated, usually via a leaving amber interval, followed by any all-red period needed to clear conflicting vehicles. In the UK a starting amber is shown before the green signal, which must then run for a minimum green period.

If traffic is still flowing over the detectors at the end of the minimum green interval, extensions of green time occur according to the speed of the vehicles. The phase is terminated when a gap in the extensions occurs or a predetermined maximum green time is reached, provided a demand for a conflicting phase exists.

Under heavier traffic the maximum green times are usually reached. The optimum values must be calculated for the cycle time, and the amount of it which is allocated to each stage in order to minimise overall delay. Although controllers are becoming 'smarter', it is usual to precalculate the values from surveyed data.

Each approach has an actual flow (q) which must be measured, and a saturation flow (s) which is the maximum that could occur if continuously fed by a queue. Saturation flow is proportional to road width, corrected for gradient.

The ratio $y = q/s$ tells us the claim that each stage will have on the available cycle time. (A narrow road with a large flow has a high y value.) Some of the cycle time is lost (L seconds) to traffic flow during the intergreen period between stages.

The optimum cycle time (c_o) for minimum delay, e.g. by Webster's method,[1] is approximately

$$\frac{1.5L + 5}{1 - Y}$$

where

$$Y = y_1 + y_2 + \cdots + y_n$$

As $Y \to 1$, $c_o \to \infty$ and a practical limit has to be set at about 120 s. Green times for the stages (the splits) can then be allocated,

$$g_1 = \frac{y_1}{Y}(c_o - L), \text{ etc.}$$

A single set of timings may not be appropriate for all peak and off-peak conditions throughout the week. Alternative timing plans (typically eight) are stored in the controller, and selected according to a time-table which is also stored and read in conjunction with a real-time clock. A controller can operate in several modes: fixed time, vehicle (traffic) actuated, police operated, emergency vehicle or rail priority (or pre-empt), or under central computer control.

44.4.1.4 Controller equipment

Historically, a very simple and reliable controller consisted of cam-operated switches driven by a synchronous motor, with dials for alternative timing plans. A modern controller consists of a microprocessor printed circuit board with programmable read only memory (PROM) and battery backed random access memory (RAM) to hold the site configuration data; a user interface to a hand-held terminal or personal computer (PC); interfaces for vehicle detectors;

a central computer connection; lamp switching modules (usually triac); and, critically, a secure device to detect the possibility of conflicting green signals.

The whole unit must be relatively cheap, and housed in a weatherproof cabinet with an internal temperature range of at least −15 to +70 °C. Additional features include signal lamp monitoring, and autodialled fault reporting to a central point, strategies which seek to provide better distribution of capacity particularly under heavy traffic conditions.

Autoscheduling is used to help the traffic engineer prepare the considerable amount of data which may be necessary for each intersection, and many features vary from site to site. Controllers such as the GEC GX therefore allow specific values to be attached to the generic values.

44.4.1.5 Vehicle detection

Vehicle detectors most commonly use shallow loops of wire in the roadway whose characteristics at frequencies of around 100 kHz are altered by the presence of vehicles. These detectors are simple, unobtrusive and precise, but demand good installation standards and relatively frequent recutting of the detector loops.

Detection of variations in the earth magnetic field, and piezo/tribo devices to measure mechanical pressure are used less commonly.

Above-surface detectors include microwave, infra-red and ultrasonic measurement of reflected energy. Structures are needed near to the point of detection at a suitable height and position to avoid vehicle obscuration. Video-image processing is used for incident detection.

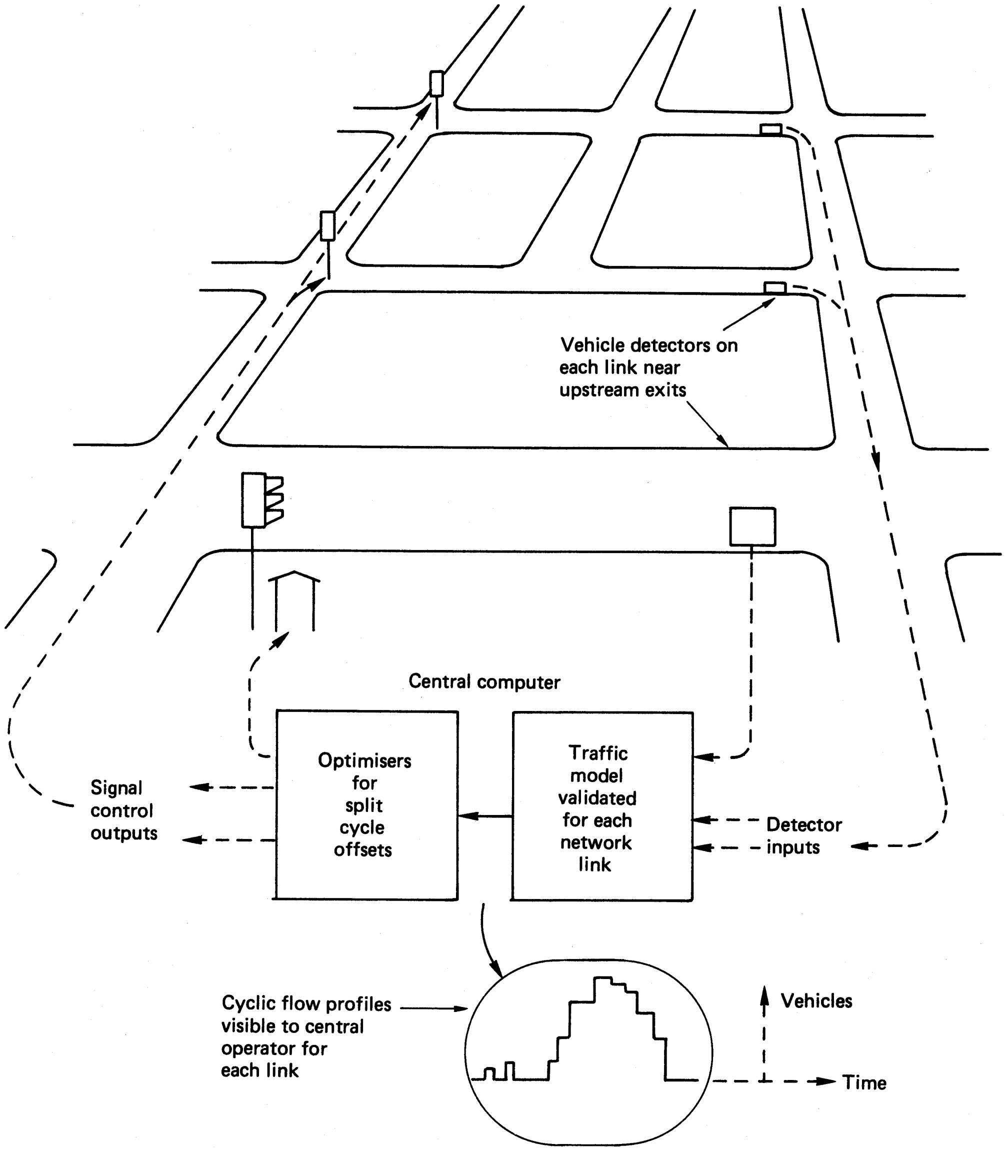

Vehicle detectors on each link near upstream exits
Central computer
Optimisers for split cycle offsets
Traffic model validated for each network link
Signal control outputs
Detector inputs
Cyclic flow profiles visible to central operator for each link
Vehicles
Time

Figure 44.32 SCOOT adaptive traffic control system

44.4.1.6 Signal coordination

Where the travel time between intersections is short enough for traffic platoons not be become unduly dispersed (e.g. about 30 s or less), it is beneficial to co-ordinate the time between the appearance of successive green signals, known as the 'offset'.

This can be done locally for small groups of intersections by synchronising the controller real-time clocks to the mains power frequency. The initial reference is set up manually, or using the dial-up remote monitoring equipment.

44.4.2 Traffic control systems

44.4.2.1 Urban traffic control

Urban traffic control (UTC) systems use central computers to co-ordinate and monitor larger numbers of intersections, from typically 15 up to thousands of intersections in a complete metropolitan area.

Optimisation of signal settings for minimum overall delay in two-dimensional networks is traditionally done off-line using surveyed traffic flows and a method such as TRANSYT.[2] Fixed time plans are then loaded into the control computers together with a time-table as described for the controller. At least 15% overall reduction in journey times should be achieved.

The computer sends stage-by-stage change messages to each controller, with typically eight on one 1200 bits per second (bps) communication channel, and checks that the controllers have responded.

Plans may be changed by operators in a control room to suit abnormal traffic conditions, using closed-circuit television to monitor key intersections. It is difficult to quantify the benefits of manual intervention in terms of traffic measurements, but the cost of surveillance equipment is often justified in terms of the overall savings from the scheme.

Fixed time plans typically require updating after about 3 years if the initial benefits are not to deteriorate. Fully traffic adaptive systems such as SCOOT are now standard which, second by second, optimise splits, cycle time and offset from real-time traffic measurements. The principle is illustrated in *Figure 44.32*.

A standard system such as the GEC Traffic Management System combines SCOOT with car-park and diversion-sign control, emergency vehicle priority, and the ability to pre-program operator actions for holidays, etc. More central computing cells and operator PCs are simply added to the central local area network in order to control larger numbers of intersections.

44.4.2.2 Motorway control

Roadside communication devices along a longitudinal cable, a central computer system and control centres are also used for motorway surveillance and control. The most basic requirement is for emergency telephones, which in some countries use solar power and radio.

Sign control is primarily for access and hazard warning. Ramp metering is a method of dynamic access control used to feed joining traffic onto the motorway. Signs may be

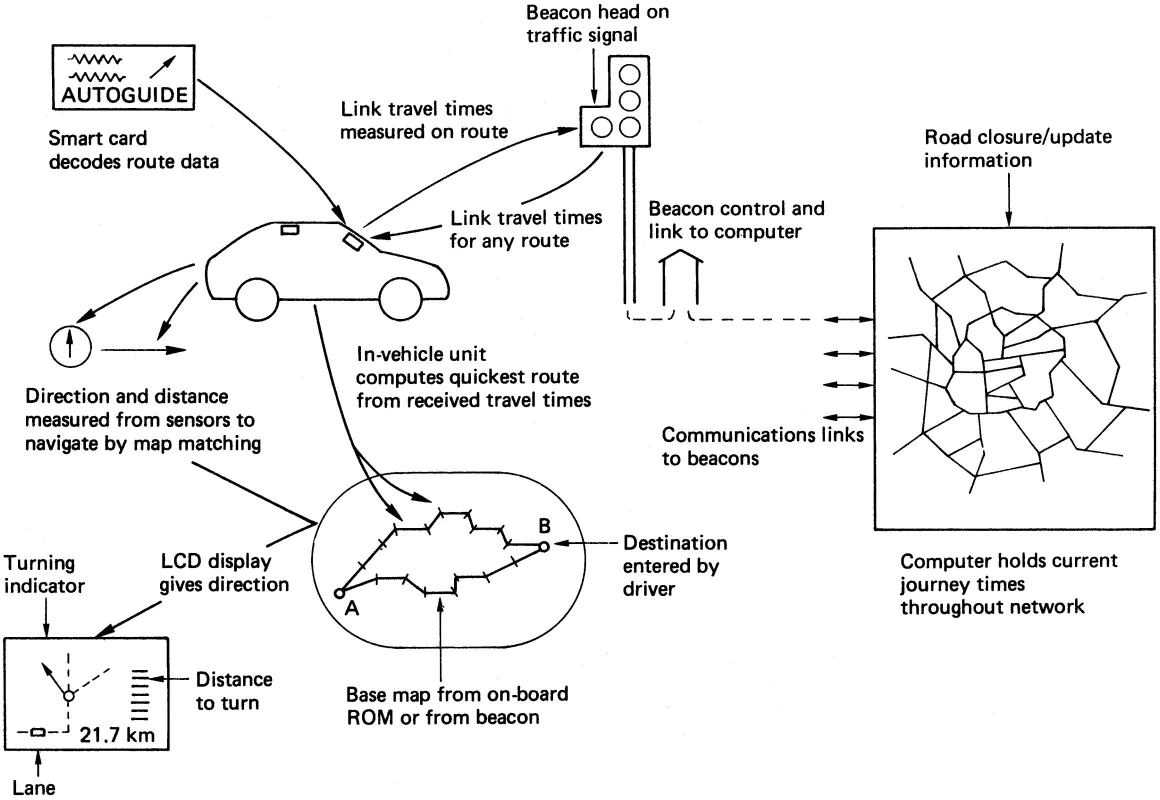

Beacon head on traffic signal
AUTOGUIDE
Smart card decodes route data
Link travel times measured on route
Road closure/update information
Link travel times for any route
Beacon control and link to computer
In-vehicle unit computes quickest route from received travel times
Direction and distance measured from sensors to navigate by map matching
Communications links to beacons
B
A
Destination entered by driver
Turning indicator
LCD display gives direction
Computer holds current journey times throughout network
Distance to turn
21.7 km
Base map from on-board ROM or from beacon
Lane

Figure 44.33 Autoguide driver-information system

used to close ramps in emergencies, and re-route traffic in conjunction with the UTC system in urban areas.

Hazard warnings on the motorway must be rapid and relevant, and are demanding in terms of closely spaced signs, vehicle and weather detection, and a reliable, fast and automatic incident detection algorithm. Their justification (by comparison with lighting, for example) must be set against relatively infrequent, but major, accidents.

44.4.3 Driver information systems

44.4.3.1 Route guidance

Traffic control schemes seek to make the best use of existing roadspace, but they can only respond to demand on the network resulting from drivers' route choice. This, in turn, is based on imprecise information about up-to-the-minute traffic conditions.

In a scheme such as GEC Autoguide, and its German equivalent Euroscout, drivers receive junction by junction routing information on a relatively cheap in-vehicle unit, based on up to the minute traffic data held in a central computer. Users help to gather this data, without revealing their identity, by giving their travel times for each network link back to the system. The principle is illustrated in *Figure 44.33*.

Other types of system provide traffic information without alternative routing advice, or provide in-car map-following navigation without live traffic data.

In Autoguide type systems information is exchanged with roadside beacons using infra-red or microwave transmission, or by cellular radio, preferably to digital standards with a broadcast channel. The vehicle is fitted with sensors to measure the direction and distance travelled, and uses map matching to correct its position. Beacon density follows network density, and is governed by the need to provide sufficiently frequent updates of journey times and map information.

44.4.3.2 Added value information

Information is purchased in a commercial system by buying a smart card which enables the in-vehicle system to decode the received data. This can include parking availability or other added value information. The roadside/vehicle communication infrastructure can also be used for vehicle location, bus information, etc.

Relatively simple in-vehicle satellite terminals are also used for location and messaging, and are well suited to providing guidance information over longer distances.

44.4.3.3 Tolls and road pricing

The combination of a vehicle-to-roadside link and smart-card technology also enables prepayment for the use of roadspace without stopping, e.g. for toll roads or road pricing in a city area.

Identification of non-paying users is achieved by detectors which can locate precisely the position of vehicles in the carriageway, and capturing video frames showing the registration numbers.

44.4.3.4 Intelligent vehicle/highway systems

Intelligent vehicle/highway systems (IVHS) are enabling transport information and control systems to become fully integrated, to include pre-trip planning by phone or teletext, route guidance, hazard warning, passenger information, parking availability, etc.

References

1 TRANSPORT AND ROAD RESEARCH LABORATORY, *Technical Paper No. 56*, London (1966)
2 TRANSPORT AND ROAD RESEARCH LABORATORY, *Transyt 9 User Manual—Applications Guide, No. 8*, London (1988)

Relevant British Standards

BS 2 (SBN 580 00500 3) *Tramway and dock rails and fishplates*
BS 173 (SBN 580 00592 5) *Rotating electrical machines for use on road and rail vehicles*
BS 1727 (SBN 580 06861 7) *Motors for battery operated vehicles* (N.B. not rail vehicles)
BS 2550 (SBN 580 06249 X) *Lead–acid traction batteries for battery electric vehicles and tracks*
BS 2618 (ISBN 0 580 08473 6) *Electric traction equipment*
BS 4846 (ISBN 0 580 07387 4) *Resistors for traction purposes* (N.B. not road vehicles)

45

Railways

D S Armstrong BSc(Eng), MIEE, MIMechE
Engineering Consultant
(Sections 45.1 and 45.2)

J D Francis MIRSE
IRO Westinghouse Signals Ltd
(Section 45.3)

Contents

45.1 Railway electrification 45/3
  45.1.1 Supply 45/3
  45.1.2 Fixed equipment 45/4
  45.1.3 Power collection 45/5
  45.1.4 Power control 45/5
  45.1.5 Drives 45/8
  45.1.6 Train resistance—motor inertia 45/9
  45.1.7 Electric braking 45/9
  45.1.8 Traction vehicles 45/9
  45.1.9 Battery locomotives 45/10
  45.1.10 Underground railways 45/10

45.2 Diesel-electric traction 45/11
  45.2.1 Locomotives 45/11
  45.2.2 Design 45/11
  45.2.3 Electrical equipment 45/12

45.3 Systems, EMC and standards 45/13
  45.3.1 Novel systems 45/13
  45.3.2 Coaching stock 45/13
  45.3.3 Electromagnetic compatibility 45/13
  45.3.4 System simulation 45/13
  45.3.5 Standards 45/13

45.4 Railway signalling and control 45/14
  45.4.1 D.c. track circuit 45/14
  45.4.2 A.c. track circuit 45/15
  45.4.3 Jointless track circuit 45/15
  45.4.4 Other forms of track circuit 45/16
  45.4.5 Other means of vehicle detection 45/16
  45.4.6 Multiple-aspect signalling 45/16
  45.4.7 Signalling for junctions 45/18
  45.4.8 Coloured light signals 45/18
  45.4.9 Signal-aspect controls 45/18
  45.4.10 Point operation 45/18
  45.4.11 The modern control centre 45/19
  45.4.12 Interlocking equipment 45/20
  45.4.13 Automatic warning system 45/22
  45.4.14 Automatic train protection 45/23
  45.4.15 Signal power supply 45/25
  45.4.16 Protecting signalling against traction currents 45/25
  45.4.17 Level crossings 45/25
  45.4.18 Train to signal box radio 45/26

45.1 Railway electrification

The benefits of using electrified lines for heavy traffic loadings are established and many countries now operate electric trains. About 180 000 route kilometres are electrified, with a variety of systems. The USSR contains about one-third of this total. Electric transmission is standard for high powered diesel locomotives.

45.1.1 Supply

Electrified-line systems fall into two broad divisions: (1) a.c. systems and (2) d.c. systems.

For city and suburban services, d.c. is most common; for main-line services there are many examples of both a.c. and d.c. systems. The use of a.c. has developed mainly in Europe and has become the standard for new schemes. In countries where the electrification of railways took place in the period 1920–1940, d.c. systems predominate and extensions maintain compatibility. For new high-power schemes, 25 or 50 kV a.c. is used for the distribution of power and sometimes as the voltage delivered to the locomotive. For lower power schemes such as new suburban lines, 25 kV a.c. is the usual voltage.

45.1.1.1 A.c. systems

A.c. systems can be divided into (1) single-phase low-frequency ($16\frac{2}{3}$ and 25 Hz) and (2) single-phase industrial frequency (50 or 60 Hz). Early three-phase systems have been replaced, although some special vehicles use three-phase collection.

Single-phase low frequency This system is widely used in Europe at $16\frac{2}{3}$ Hz and to a limited extent in the USA at 25 Hz. The low frequency was chosen to allow commutator motors to be used directly, with simple voltage control of power. Supply is taken from an overhead conductor at voltages up to 16 kV, with the rails used for return.

Figure 45.1 shows a typical power control circuit for a two-motor equipment. The supply is taken from the overhead line through a circuit-breaker to the primary of a step-down transformer, the secondary of which is tapped for low-voltage feed to the traction motors. Motor voltages of a few hundred volts are typical.

Single-phase industrial frequency The advantage here is the use of main national grid systems as the power source, instead of special low-frequency generation or conversion stations. Pioneering work was done by the French in 1951, with several experimental locomotives subsequently being built. Although 50 Hz commutator motors were feasible, the use of d.c. motors from a rectified supply was preferred. Power control is either by control of the a.c. voltage prior to full-wave rectification, using a tap-changer, or control of the d.c. voltage by phase-angle control of a thyristor bridge. British Railways adopted the 25 kV, 50 Hz system in 1956; an early traction-control circuit in which a tap-changed voltage was used to buck or boost a fixed voltage in the feed to the rectifier is shown in *Figure 45.2*.

Industrial frequency is now used in many countries, with about 20 000 route kilometres in operation. Voltages of 50 kV are used, in some cases with locomotive equipment rated for this voltage, in other cases being used for feeding autotransformers to deliver 25 kV to the train. In the UK, 25 kV has been used for all schemes, but the Channel Tunnel route to London will use 50 kV with autotransformers. *Figure 45.3* shows the principle of autotransformer feeding with these transformers connected between catenary, rail and the auxiliary feeder which is at −25 kV. Current is forced to be equal in the two windings of these transformers and the 200 A train load is provided by addition of the components shown. The supply transformer is then feeding 100 A at 50 kV.

45.1.1.2 D.c. systems

Systems operating at 600–1000 V are used extensively for urban and suburban electrification, usually with one live insulated rail and running rail return. Some four-rail systems exist, with positive and negative insulated electrical supply rails. Systems working at more than 1000 V use overhead catenary-supported conductors; typical voltages are 1500 and 3000 V. For heavily trafficked routes, 12 000 V has been considered.

D.c. motors are customary, although the trend is now towards the use of three-phase induction motors, driven from semiconductor inverters. The use of d.c. will reduce problems of electromagnetic compatibility, since voltages are not induced in track-side conductors. High power is more difficult to provide in d.c. systems than in a.c. systems since the voltage rating of d.c. train equipment cannot approach the 50 kV of the a.c. train. Currents are therefore high, requiring the use of heavy overhead conductors and closely spaced substations. With d.c. lines there are also possible problems of corrosion due to leakage currents.

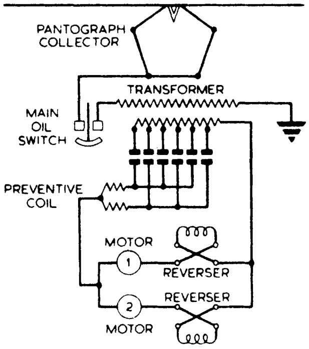


Figure 45.1 Main-circuit diagram for two-motor multiple-unit equipment with single-phase series motors

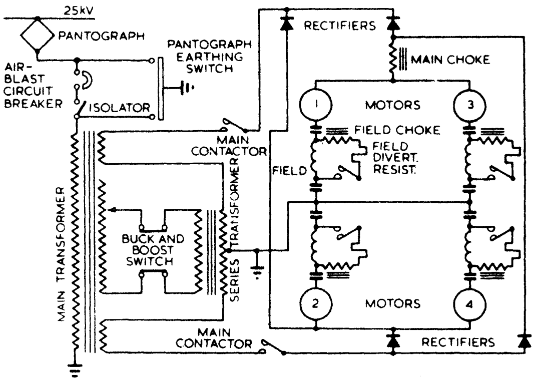


Figure 45.2 Main-circuit diagram for a 25 kV locomotive

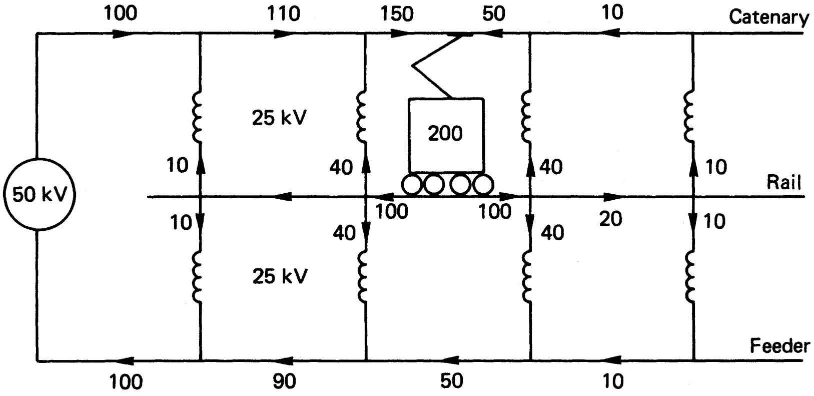


Figure 45.3 Current distribution in an autotransformer system for 200 A train load

45.1.2 Fixed equipment

Provision of the live conductor for carrying power to the trains is complex and expensive, probably costing as much as the trains which will be used.

45.1.2.1 D.c. conductor rail

The d.c. conductor rail is typically a flat-bottomed cast iron (99.75% Fe) rail of 50–75 kg/m mounted on porcelain insulators at the side of, and spaced about 40 cm from, one of the running rails. The negative return is via the running rails unless (as on London Underground lines) circumstances require an insulated return rail, usually mounted on insulators between the two track rails. A 45 kg/m rail has a resistance of about 20 $\mu\Omega$/m.

A composite conductor rail is available with an aluminium body and a stainless steel contact surface. This has the advantages of reduced electrical resistance (typically half) and mass, but at a higher cost. This may be recouped if the substations can be placed further apart for the same voltage drop. Such rail is used on the Botley to Fareham line of British Railways, on the Docklands Light Railway, London (under-side contact), and the Singapore Mass Transit.

45.1.2.2 D.c. overhead conductor

A hard-drawn copper contact wire is supported above the track by a catenary and dropper wires, in such a way that the contact wire sags by a small amount in mid-span. This sag results in an approximately level path for the pantograph at speed. On heavily loaded 1500 V lines, a twin contact wire may be used.

To maintain the wire tension in varying weather conditions, the catenary and contact wires are anchored via insulators to a structure at the mid-point of the wire length and are then stretched by weights at each outer end. The longitudinal force in catenary and contact wire is about 10 kN. The lateral position of the contact wire alternates from side to side of the track centre line, traversing the active width of the pantograph collector strip during running to equalise the wear of the contact materials.

45.1.2.3 Feeder stations (d.c.)

Multipulse rectification is used with star/delta transformer connections and the basic ripple frequency on the d.c. output is 6 or 12 times the a.c. frequency. Mercury are rectifiers are still in service, but new stations use silicon rectifiers. The d.c. rail (or overhead conductor) is connected as a continuous circuit, energised at each feeder station.

45.1.2.4 A.c. overhead conductor

This is similar to the d.c. system. For high-speed operation, a compound arrangement is used with an intermediate catenary. In another design, a partial second catenary some metres long is placed at the support, to reduce the change of mechanical impedance at the contact point. The tension may have to be increased for high-speed use and the Deutsches Bundesbahn use 15 kN for their 250 km/h design (70 mm^2 catenary and 120 mm^2 contact wires).

The Mark III design used by British Railways has a catenary of five strands of aluminium each 3.95 mm diameter with two equal strands of steel to give strength. This has an equivalent copper area of 42 mm^2. A copper contact wire of 107 mm^2 is provided. Where overhead conductors are provided to give a preferential path for return current, these have 10 strands of aluminium, each 4.22 mm in diameter.

A consideration is the avoidance of conductor clashing in the event of short circuit. This places a minimum value on the vertical separation between the catenary and contact wires.

Where vertical space is at a premium such as in tunnels, a solid bar can be used instead of a contact wire. Such an arrangement can be used at speeds up to 140 km/h.

45.1.2.5 Feeder stations (a.c.)

The $16\frac{2}{3}$ and 25 Hz systems use frequency converter stations at the points of connection to the national electrical network. Static frequency converters using thyristors are now in service at ratings of several megavolt-amperes, displacing the earlier synchronous converters.

Each feeder station at 25 kV supplies about 50 km of route and, to distribute the single-phase load on the main electrical network, different phases are used at succeeding feeder stations. The mid-point between feeders consists of an insulating neutral section through which the trains coast at zero power.

A wide spacing between feeder stations is a particular advantage in countries which do not have an extensive industrial frequency network. In extreme cases, the electrical supply for the railway has to be taken along the railway route as a three-phase high-voltage system, sometimes using structures common with those which support the overhead contact wire. One alternative feeding method is to use a 25 kV supply from overhead wire to rail, feeding this through autotransformers fed from a 50 kV longitudinal feeder.

The traction current passing through the locomotive transformer returns to the feeder station via the rails. If no special arrangements are made, the current will enter the earth as well as the rails. This current distribution can induce voltages in parallel electrical conductors such as communication cables. These voltages can reach dangerous values and it is customary to provide a return conductor in parallel with the rails to reduce induction. A further improvement is obtained if current transformers are connected in the energised and in the return conductors, forcing the current into the latter and removing the current from the track rails and earth.

The feeder station includes protection switchgear to remove the supply in the event of a short circuit. Vacuum circuit-breakers are used, giving a small switch which can be mounted on a steel mast as part of the overhead construction. Motorised or hand-operated isolators are provided to ensure safety during maintenance work.

45.1.3 Power collection

45.1.3.1 D.c. shoe

The collector shoes are of cast iron, and adequate contact between shoe and rail is effected by the mass of the shoe (10–25 kg). For high-speed running a spring-loaded shoe may be used with a contact force of about 300 N.

The use of an upwards-facing contact surface on the rail permits the use of a simple gravity shoe but the surface is vulnerable to contamination and ice formation. Alternative systems use side contact or bottom contact, with spring-loaded shoes.

45.1.3.2 Current collectors (overhead)

A mechanism is required to maintain a current collector in contact with the overhead system. This, normally a pantograph, can be of several forms: a typical version comprises a folded arm which is raised by an air cylinder, carrying a contact head on separate springs. The collector material which slides along the overhead conductor can be of various materials, but metallised carbon is used in the UK. The pantograph frame ensures that the head moves on an approximately vertical path and maintains a constant contact force of about 90 N between head and conductor.

45.1.4 Power control

45.1.4.1 A.c. supply to train

Transformer design features for traction The prime objectives in transformer design are low mass and high reliability. A modern 1.5 MVA naturally cooled transformer weighs about 3100 kg. A 7.7 MVA unit weighs about 8200 kg.

Tap-changer Older designs used mechanical switches to change the taps on a transformer, this stepped voltage being fed to rectifiers and then to the d.c. motors. On $16\frac{2}{3}$ Hz the traction motors are capable of working with the a.c. voltage source.

Thyristor phase angle Recent designs use thyristors to regulate the voltage. *Figure 45.4* shows the basic elements of the armature power control circuit of a phase-angle-controlled locomotive. The output winding of the transformer is divided into a number of parts (usually two but more can be used), which feed asymmetrically controlled thyristor/diode bridges. Each bridge is sequentially advanced to full conduction as increased power is required.

Thyristor current normally extinguishes by natural commutation as the voltage reverses. *Figure 45.5(a)* shows the voltage and current waveforms for partial conduction of a single bridge; *Figure 45.5(b)* shows full conduction. The latter shape is that which applies for tap-changer/diode circuits.

The power factor can be improved if the current is forcibly reduced to zero during the half-cycle, a system known as sector control. The drawback is an increase in harmonic content. *Figure 45.6(a)* shows part power and *Figure 45.6(b)* full power for a single sector-controlled bridge.

An improved harmonic performance is available if the firing angles of both bridges are controlled simultaneously, with different angles for each bridge. This can reduce the selected harmonic content of the primary current.

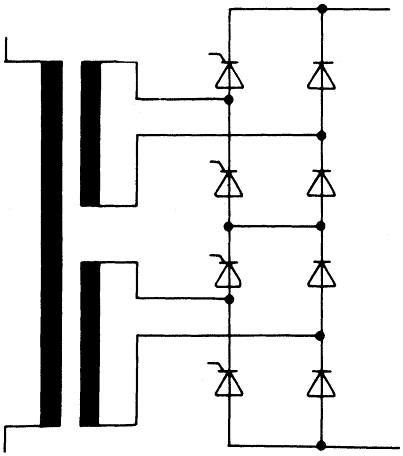


Figure 45.4 Two sequential half-controlled bridges

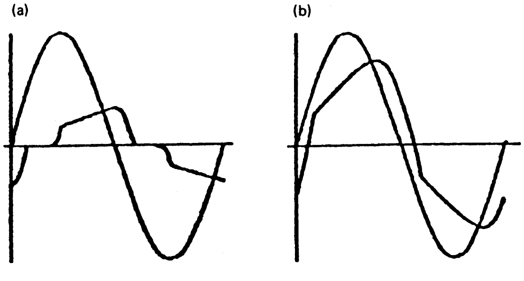


Figure 45.5 Voltage and current waveforms of a phase-angle-controlled bridge: (a) partial power; (b) full power

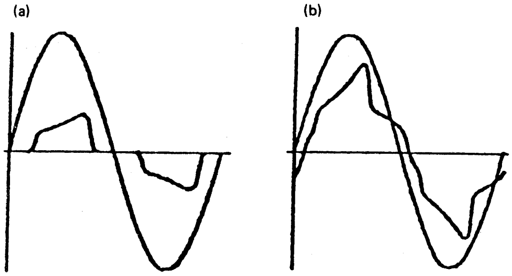


Figure 45.6 Voltage and current waveforms of a sector-controlled thyristor bridge: (a) partial power; (b) full power

The use of several secondary windings with sequential control can improve the harmonic content and power factor, but at the expense of weight and cost.

The fields of the traction motors are supplied from a separate thyristor bridge, with arrangements to link the armature and field-control circuits to maintain a satisfactory operating balance between currents in the motor.

Gate turn off The advent of the power gate turn off (g.t.o.) thyristor has changed power circuits in a dramatic way. As well as assisting the design of inverters it has allowed the use of power conditioning of the single-phase input in novel ways. Individual g.t.o.s of 4000 V and 2500 V are available, allowing single device per arm circuits to be built for traction ratings.

Four-quadrant chopper This circuit, shown in *Figure 45.7*, converts single-phase a.c. into d.c. (or vice versa) with a high power factor and low harmonic content. The essential process is the creation of energy in the inductor L and transfer of this energy to the link capacitor C. The g.t.o. thyristors T1 to T4 are switched to provide three voltages across the inductor; either (V a.c.) or (V a.c. + V d.c.) or (V a.c. – V d.c.). For positive a.c. current, the (V a.c.) condition is provided via (T2 + D4) or (T3 + D1), (V a.c. + V d.c.) via (T1 + T3), and (V a.c. – V d.c.) via (D1 + D4). Other circuit paths are switched to cope with negative a.c. current. The four-quadrant chopper circuit can operate in all four quadrants of the voltage/current vector diagram and permits inversion from d.c. to a.c. In traction use the circuit is usually controlled to give a constant d.c. link voltage. Switching operations take place at 300–500 Hz and pairs of four-quadrant circuits can be interleaved to give a.c.-side harmonics at 600 or 1000 Hz. On/off control can be regulated to give a unity power factor fundamental current with no harmonics below switching frequency. The switching produces an output component at twice the supply frequency and the filter LF and CF is provided to absorb this current.

Although using more components than phase-angle control and requiring a complex firing control, the benefits of unity power factor, low harmonic content, and simple regeneration make this the preferred control circuit for single-phase traction.

Figure 45.8(*a*) shows part power and *Figure 45.8*(*b*) full power voltage and current waveforms for a four-quadrant chopper circuit. The in-phase nature of the current is apparent, as is the low harmonic content.

Three-phase inverter Early inverters required the use of many components and were complex, heavy and subject to failure. Developments in controlled semiconductors and

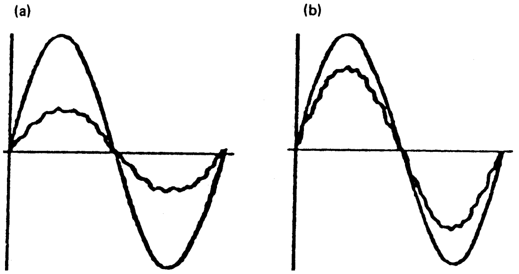


Figure 45.8 Voltage and current waveforms of a four-quadrant chopper circuit: (a) partial power; (b) full power

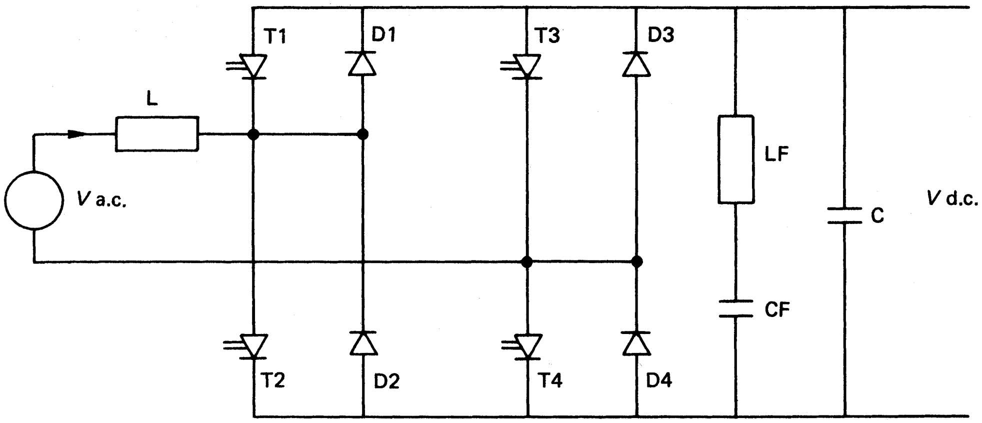


Figure 45.7 Four-quadrant chopper circuit

improved cooling techniques have now produced small compact designs more suited for traction. Freon cooled inverter modules are used on some railways with a very low specific mass. One example, handling 1.1 MW, weighed only 140 kg.

The switching pattern of the power semiconductors is usually regulated to minimise the generation of harmonic currents at signalling frequences.

45.1.4.2 D.c. supply to train

Electropneumatic camshaft control The most noteworthy feature of this control unit is that the camshaft which controls the operating sequence of the accelerating contactors makes a complete revolution in one direction for the series notches, and a revolution in the reverse direction for the parallel notches, the transfer from series to parallel motor connection being made on a separate pneumatic switch unit. The initial and final positions of the camshaft are identical, so that in the case of a power interruption the equipment is ready for an immediate restart.

It will be seen that this arrangement, which is made possible by a special system of motor and resistance circuits (*Figure 45.9*) introduces each contactor in circuit twice during the accelerating period, once during series and once during parallel notching. Thus, when compared with control units employing separate contact systems for the series and parallel conditions, the number of accelerating notches available in a contactor group of given size and weight is almost doubled, while the simplicity of subsidiary control circuits and of mechanical construction, which are features of the pneumatic camshaft principle, are retained, with low power consumption for control apparatus.

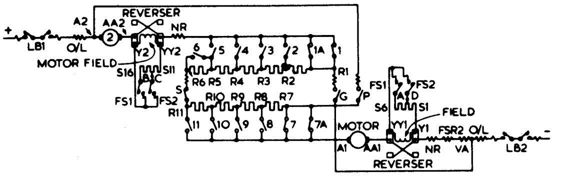
Some equipments have 10 series and 10 parallel steps (in contrast to the five and four common with simple resistance control), making possible higher average rates of acceleration without wheel slip.

The driver's controller will have a small number of positions, and transitions between them will be regulated by the current limit relay. Typical positions are 'off', 'shunt' (motors in series with all resistors in circuit), 'series' (motors in series with resistors out of circuit), 'parallel' (motors in parallel with resistors out of circuit), plus a number of 'weak field' positions. In urban units the control sequence lasts only for a short time and may be completed before the train has left the station.

Thyristor regulator (chopper) control Regulation of the average voltage applied to the traction motor is available if a thyristor switch is used to connect the supply cyclically to the motor. A thyristor carrying direct current will only switch off if the current is reduced to zero and the thyristor has a reverse voltage applied for a short time. Many circuits are in use to achieve this, the basic concept being that a charged capacitor is used as a temporary source of current, reverse biasing the thyristor to turn it off. The capacitor is switched into circuit using an auxiliary thyristor. A typical circuit is shown in *Figure 45.10*.

When the supply is connected, thyristor T2 is turned on to charge capacitor C via the motor path. When C is charged, the current in T2 falls and it extinguishes. Turning on thyristor T1 applies the full supply voltage to the motor and also allows the charge on C to oscillate (via L and D1) for a half-period, thereby reversing the polarity of the potential across C. When a sufficient current is flowing through the motor, thyristor T2 is turned on. Capacitor C acts as a current source for the motor and applies a reverse voltage to T1. If C has a sufficient charge, T1 will be reverse biased for long enough to be extinguished, thereby blocking the flow of forward current; C will then be charged as in the initial operation. The motor current will decay via the free-wheel diode D2. When the motor current has fallen to a selected value, T1 is again fired and the cycle repeated.

By control of the firing pulses to T1 and T2 the mean voltage applied to the motor can be regulated as the train speed changes, while maintaining motor current and tractive effort. It is customary to fire T1 at a fixed frequency, so that the input filter L_fC_f can be designed to limit the ripple currents



SEQUENCE TABLE

	STEP	PC POS'N	LB1	LB2	P	S	G	1	1A	2	3	4	5	6	7	7A	8	9	10	11	FS1	FS2
SERIES	OFF	1				●		●	●						●	●						
	1	1	●	●		●		●	●						●	●						
	2	2	●	●		●		●						●	●	●						
	3	3	●	●		●				●				●	●		●					
	4	4	●	●		●				●	●			●			●					
	5	5	●	●		●					●			●			●	●				
	6	6	●	●		●					●	●		●				●				
	7	7	●	●		●						●		●				●	●			
	8	8	●	●		●						●	●	●					●			
	9	9	●	●		●							●	●					●	●		
	10	10	●	●		●							●	●						●		
	T	10	●	●	●	●	●						●	●						●		
PARALLEL	10	10	●	●	●		●						●	●						●		
	11	9	●	●	●		●						●	●					●	●		
	12	8	●	●	●		●					●	●	●					●			
	13	7	●	●	●		●					●		●				●	●			
	14	6	●	●	●		●				●	●		●				●				
	15	5	●	●	●		●				●			●			●	●				
	16	4	●	●	●		●			●	●			●			●					
	17	3	●	●	●		●			●				●	●		●					
	18	2	●	●	●		●	●						●	●	●						
	19	1	●	●	●		●	●	●						●	●						
WEAK FIELD	20	1	●	●	●		●	●	●						●	●					●	
	21	1	●	●	●		●	●	●						●	●					●	●

Figure 45.9 Main-circuit diagram of a 600 V, two-motor control equipment with two field-shunt positions

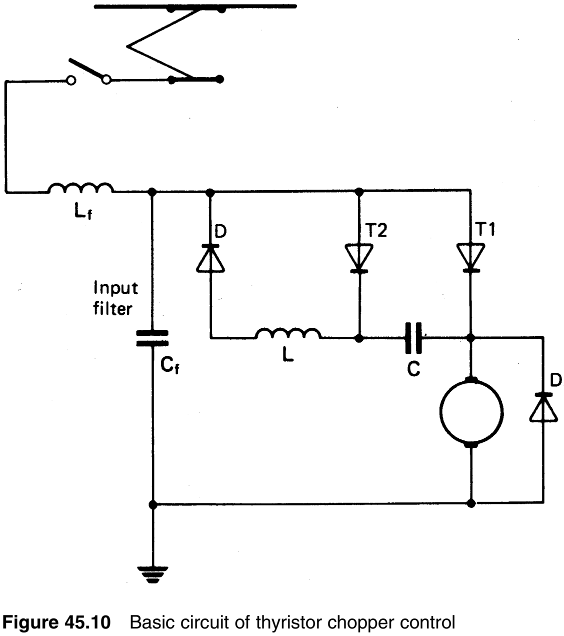


Figure 45.10 Basic circuit of thyristor chopper control

drawn from the traction supply. A fixed frequency also avoids the generation of current at track circuit frequency.

This form of control avoids the energy losses associated with rheostatic control and is particularly valuable where frequent starting is required. It is therefore used widely for urban rapid-transit systems. By reconnecting the elements of the circuit it can use the motor as a generator and convert the mechanical energy of the train during braking into electrical energy and return it to the traction supply. If the supply is not receptive, braking rheostats on the vehicle can be used.

Recent equipments have used gate turn-off thyristors which avoid the need to provide a separate thyristor-switched turnoff circuit. The chopping frequency can be varied, usually in multiple steps, to allow the wide range of output voltage to be provided. Starting is at low frequency since, at high frequency, the minimum on-time of the thyristor would deliver too high a voltage to the motor.

Inverter As with a.c. supply, inverters are used for power control. A simple inductor/capacitor input filter is used to isolate the inverter from transients on the power supply and to prevent switching frequency currents from flowing from the inverter into the supply.

45.1.4.3 Operation of multiple unit trains, train lines

Operation of several power units from one driving cab can be provided by the use of a multi-core cable which runs the length of the train. Known systems use 27 or 36 cores to control the essential functions and return indications of response.

45.1.4.4 Train line multiplex for locomotive push or pull

Where a train is to be driven from both ends but train lines have not been installed, or are expensive to provide in new vehicles, power control can be achieved using two wires only. A series of coded signals can be sent to intelligent decoders, using techniques such as time division multiplex. Actuators can be addressed individually and sent binary messages which give the necessary commands. Frequency shift keying, typically at 90/120 kHz can be used for binary codes.

45.1.4.5 Driving cab

Where possible, it is desirable to use standard cab layouts for all forms and ratings of traction. Power and brake handles should be in similar places and the flexibility with which electric controls can be positioned is then an advantage.

Display of information to the driver is essential for safe and reliable operation of the train. Unambiguous displays with good lighting are required, whilst ensuring that the driver is not distracted from outside signals. Train status can be displayed on touch-sensitive screens, with selectable menus for different functions.

The increasing use of track-based signalling systems, which deliver permitted speed values to the cab, are reducing the need to observe track-side signals. They demand, however, an increase in the amount of information which has to be presented to the driver and the ergonomic design of cabs is receiving increased attention.

Closed-circuit television with visual display units can be used to monitor the surroundings, such as passenger behaviour on platforms.

45.1.5 Drives

The mechanical means by which the drive is taken from the motor to the axle are worthy of mention since they have important implications for the type of motor which can be used and the environment in which the motor has to operate. A simple drive places half of the motor mass on the axle, with high consequential mechanical loading on the motor. Complex drives can isolate the motor from track stresses.

Axle-hung geared motor The motor is suspended partly by bearings around the driving axle and partly by a bracket resting on the bogie frame. The motor output is taken to the driving axle via a single reduction gear. This is a simple form of construction but the non-spring-borne part of the motor mass results in increased dynamic loads on the track. If high-speed motors are used an idler gear can be interposed, or a double reduction train used.

Axle-hung direct drive The axle can form a driven part of the motor with the electrical armature placed inside (held through the axle bearings) or outside the axle. In one version, a three-phase induction motor is formed from a wound stator held within a large-diameter tubular axle, the cage winding of the motor being fixed to the inner surface of the axle tube.

Frame-mounted geared motor For express locomotives and for other cases in which riding qualities are important, drives have been devised to reduce the non-spring-borne mass, to raise the mass centre and to allow for relative movement between frame-borne and axle-borne parts.

Spring drive This consists of a hollow shaft surrounding the driving axle and having sufficient clearance therefrom to permit the necessary relative movement between the spring-borne quill and the axle. The quill carries the gearwheel (or a gearwheel at each end) engaging with a pinion on the motor shaft. A twin- or double-armature motor is frequently employed, and to secure flexibility the pinions may be spring cushioned. The chief feature of the drive is the method of connecting the quill to the driving axle, which is accomplished by an arrangement of circumferential springs acting between a spider on the quill and a special spoke arrangement on the driving wheel.

Link drive To avoid the use of springs, forms of quill drive using links (such as the Buchli drive) have been designed. The locomotive frames are within the wheel space and the motor shaft extends over the frame to a pinion engaging with a gearwheel carried in a frame-mounted bearing. The gearwheel is arranged nearly concentric with the driving wheel and outside it; the connection between the two is made by an ingenious gear-link arrangement working on to pins fixed to the driving wheel.

Flexible-disc drive Here the motor is mounted on the bogie frame. The armature shaft is hollow and the drive is taken through a flexible disc coupling at one end by means of a shaft passing through the armature and connecting through another flexible disc to the pinion which is carried in bearings in the gearcase.

Alsthom drive The drive to the wheels is by means of a hollow quill shaft and flexible links, using a ‘dancing member’ and rubber-bushed bearings in the links. The quill is carried in the motor frame by large diameter taper roller bearings.

Longitudinal monomotor The motor is arranged along the longitudinal axis of the bogie frame, driving both axles from gear outputs at both ends of the motor shaft. Bevel gears are provided on each axle, with an intermediate flexible coupling. This is used on the GEC drives for the

London Docklands light railway, with flexible couplings to allow angular movement of the axles.

Bogie frame direct The motor is supported from the bogie frame, with a hollow output shaft. This shaft has a sufficient internal clear diameter to accommodate the axle and the relative axle to bogie movements. Drive is taken to both the wheel discs via links. An example of this arrangement is the 650 kW, 600 rev/min induction motor drive of the Skoda prototype 85 EO locomotive.

Bogie frame monomotor A large single motor is mounted on the bogie frame, occupying part of the body space. Drive to the axles is via gear trains. This couples the axles together for traction purposes. It gives a low yaw inertia to the bogie. French National Railways are the main users of this system.

Body mounted motors Where the bogie design has to meet special targets, for example giving high speed with low track loading, the motor can be attached to the body of the vehicle. This requires a drive which can tolerate the substantial vertical and lateral relative movements between body and axle. The motor can be arranged with the drive shaft parallel to, or at right angles to, the axle. In the former case the output can be taken back across the vehicle; in the latter a splined cardan shaft can be used with universal end couplings, to drive via a bevel gear box.

Individual wheel drives Suspension designs have been reviewed and separate wheels, not linked via common axles, are used for special cases. The Italian AVRIL train uses a pair of 200 kW motors mounted longitudinally side by side on the body, to drive two wheels on one bogie via a cardan and bevel gear transmission. Each vehicle has two pairs of motors.

45.1.6 Train resistance—motor inertia

When calculating train resistance for run timings, note that the inertia of the motors is to be added to that of the other rotating parts, and that the windage losses of the motors may be significant.

45.1.7 Electric braking

The ability to use motors for braking allows a reduction in brake wear and offers the possibility of converting brake energy into useful traction energy for use by other vehicles. Electric braking has to be blended with the mechanical braking system of the train and the mechanical brake must be arranged to take over the full braking duty if the electrical mode fails.

45.1.7.1 Regenerative braking

The braking of heavy trains on long down gradients is materially facilitated by the use of regenerative braking, whereby some of the mechanical energy released by the train in its downward progress can be reconverted into electrical energy and returned to the supply system. For this purpose it must be possible to reconnect the traction motors as generators and control the speed by the braking torque developed. Regeneration is possible with any form of electric traction at the expense of additional weight, cost and complication. Regeneration has been most successfully employed on locomotives which have to negotiate heavy grades. Regenerative braking is also feasible in heavily loaded urban railways. If the railway system is not receptive (detected typically by a rise of voltage above a given limit), the excess energy can either be returned to the national electrical supply or dissipated in rheostats at the railway substations. Regenerative braking can give significant reduction in overall energy consumption on urban railways, and typical figures of 15–20% are achieved on densely loaded systems. Cost savings due to reduced brake wear can be equal to energy savings.

45.1.7.2 Rheostatic braking

An alternative form of electric braking, referred to as rheostatic or electrodynamic braking, is available for electric locomotives and is particularly useful for a.c. locomotives, which cannot be designed for regenerative braking as readily as those operated on d.c. systems. With this type of braking, the energy generated by the motors is dissipated in resistors. Control is usually exercised by separately exciting the motor fields, but some systems use self-excitation of the motors with main resistance switching. Owing to the large amount of energy which may have to be dissipated for considerable periods if descending a steep incline, the braking resistor is usually forced-air cooled, and special types of strip resistor units have been developed for this purpose. Where substantial rheostatic braking is required, the roof of the vehicle can sometimes be used to accommodate the rheostats.

45.1.7.3 Braking using the rail

Use of the rail for braking has attractions since the thermal mass is large and new rail enters the brake zone as the train proceeds. Care needs to be taken if frequent braking takes place on the same rail since thermal stresses may become excessive, increasing the tendency of the track to buckle.

Eddy on rail Non-contact braking can be provided by the use of an eddy current brake acting on the rail head. A multipole array is suspended close to the rail head and a 2 m long device, with a 7 mm air gap, has developed a brake force of 14 kN at 250 km/h. The excitation power was 42 kW.

Contact on rail Alternatively, magnets can be arranged to make contact with the rail head when energised. Brake force depends on the coefficient of friction and forces of 5–15 kN have been found from a 2 m long unit. Power consumption is about 2 kW. Permanent magnet (neodymium) excitation is also used.

Both these methods involve substantial vertical forces.

45.1.7.4 Electric actuator for mechanical brakes, motor/screw

Although brakes are normally operated by air cylinders, a motor-driven screw can be used to apply and release the brakes. This gives a substantial reduction in mass by removing air pipes and reservoirs.

45.1.8 Traction vehicles

45.1.8.1 D.c. locomotives

On d.c. locomotives there is almost invariably an even number of traction motors. With two motors (there are usually at least four) the simple well-known series/parallel control can be used.

In this, the motors are first connected in series with starting rheostats across the contact line and rails; the rheostats are then cut out in steps, keeping roughly constant current, until the motors are running in full series; next the motors are rearranged in parallel, again with rheostats; the rheostats are cut out in steps, leaving the motors in full parallel. Some stages of field weakening are generally included. The power input remains roughly constant during the series notching, then jumps to twice this value during the parallel notching. With a four-motor equipment, a series/series–parallel/parallel connection can be used, giving three economical speeds (i.e. running without resistance) unless the line voltage is too high to be applied direct to a motor. In modern locomotives, chopper control of motor power is used and some locomotives have three-phase inverter/motor systems.

Italian Railways use high power 3000 V locomotives and the E 632 uses a three-bogie design to provide six driven axles. Intended for passenger service, with a top speed of 160 km/h, it weighs 105 t, is 17.8 m long, 4.3 m high and 3 m wide. Three large motors, each weighing 5400 kg and of 2000 V, 1635 kW, 1700 rev/min maximum rating, are frame mounted and drive two axles each via link drives.

Power control is by d.c. chopper, with the basic chopping frequency of 390 Hz converted to 2340 Hz at the supply by interleaving the six choppers. Harmonic current taken from the supply does not exceed 0.3 A up to 1300 Hz.

The main circuit contactors performing the functions of motor grouping, field weakening, and cutting out rheostats are operated either electromagnetically (e.g. by solenoid) or electropneumatically (electrically controlled compressed air cylinders); further, they may be grouped together and interlocked by a camshaft, or separated and individually controlled.

When a locomotive is fed from a third rail system, the inevitable gaps in the conductor rail at points and crossings (up to 150 m) cause momentary interruptions to the supply, resulting in snatching and surging of the vehicles, particularly with goods trains. This difficulty has been overcome by the Southern Region of British Rail, which has put into service 1100 kW, 600 V d.c. locomotives using a booster motor-generator in the traction motor circuit; a flywheel on the motor-generator set provides sufficient stored energy to enable the locomotive to run over a gap without perceptible reduction in torque.

Proper distribution of the weight of the apparatus in the locomotive body is an important design feature and strongly affects running qualities.

45.1.8.2 A.c. locomotives

Low frequency Systems operating at $16\frac{2}{3}$ or 25 Hz normally use one-phase commutator motors. These require low armature voltages to achieve good efficiency and high power factor. The motors are generally connected in parallel and, except for smaller locomotives, primary tap-changing on the main transformer is used to provide a variable voltage supply to the motors. With this type of control, each accelerating notch is an economical running point. This is in contrast to the d.c. locomotive, which cannot run continuously on the resistance starting notches. To avoid jolts and momentary short circuits, preventive coils connected across successive tappings are used. Various elaborations of the simple transformer tapping system have been devised to reduce the number of tappings without reducing the number of available motor voltages. Thus, by use of a small auxiliary transformer a tapping voltage can be increased or reduced by a required amount, so that six tappings will give 18 running voltages.

Industrial frequency D.c. series motors are fed through a rectifier (or controlled rectifier) which is, in turn, supplied by variable voltage from a tap-changer (or fixed voltage) on the main transformer. Semiconductor rectifiers are normally used and give reliable service when suitable protection is provided against the voltage transients which occur due to switching and other causes. Primary protection is provided by an air-blast circuit-breaker mounted on the roof, and in the case of solid state rectifiers, it is usual to protect each string of rectifiers by a fuse. A smoothing inductor is incorporated in the rectifier circuit in order to limit the ripple in the d.c. supply to the motors.

Figure 45.11 shows the layout of a tap-changer control a.c. locomotive. Modern locomotives for 160 km/h have similar layouts but for higher speeds the motors are normally body-mounted.

A specially prepared short train of French Railways stock (two traction vehicles and three coaches of the TGV-A) attained a speed of 515 km/h in May 1990.

45.1.8.3 Multiple unit trains

For suburban and stopping services, passenger coaches are equipped with traction apparatus placed below the floor. This leaves the body free for passengers. Both d.c. and a.c. traction supplies are used, with the 25 kV cable passing sometimes vertically through the passenger compartment from roof to transformer. Rheostat, chopper d.c. and inverter drives are used with d.c. supply and tap-changer/rectifier, controlled phase angle and inverter drives with a.c. supply. Power ratings of up to 400 kW per coach are used. These trains form the greater number of electric traction vehicles on railways.

45.1.9 Battery locomotives

These are still built and recent examples are those for the Hong Kong Mass Transit Corporation. They can operate from the 1500 V d.c. overhead or from the 750 V battery. Two 140 kW motors are installed, with three-thyristor resonant chopper control. The battery capacity is 360 A-h at the 5 h rate. Such locomotives are valuable for night-time use when the power supply is switched off for maintenance, and for shunting duties in yards which can then be left without overhead conductors.

45.1.10 Underground railways

In city centres where a high-capacity transport system is required but where surface routes are not available, railways can be built underground to provide a rapid passenger transport. The cost of tunnelling is very high and this method is only economic for large concentrated passenger flows. The electrical supply is normally by third-rail system at about 750 V d.c.; but where tunnelling costs permit, overhead electrification at higher voltages is provided. Underground railways use the same range of electrical propulsion and control techniques as on surface railways. The acceleration of underground trains is rapid and the separation between stations is small, perhaps 800 m. Under these conditions, if rheostatic camshaft control is used, the starting sequence may be completed before the train leaves the station. Subsequent running is on the natural characteristic of the motors, with weak-field operation to extend the speed range.

In France some underground lines use rubber tyres for propulsion and guidance, running on concrete tracks. This provides high acceleration and braking, contributing to passenger capacity. The rubber tyres have a higher energy loss than steel wheels, and additional ventilation may be necessary if a line is converted to rubber-tyred operation.

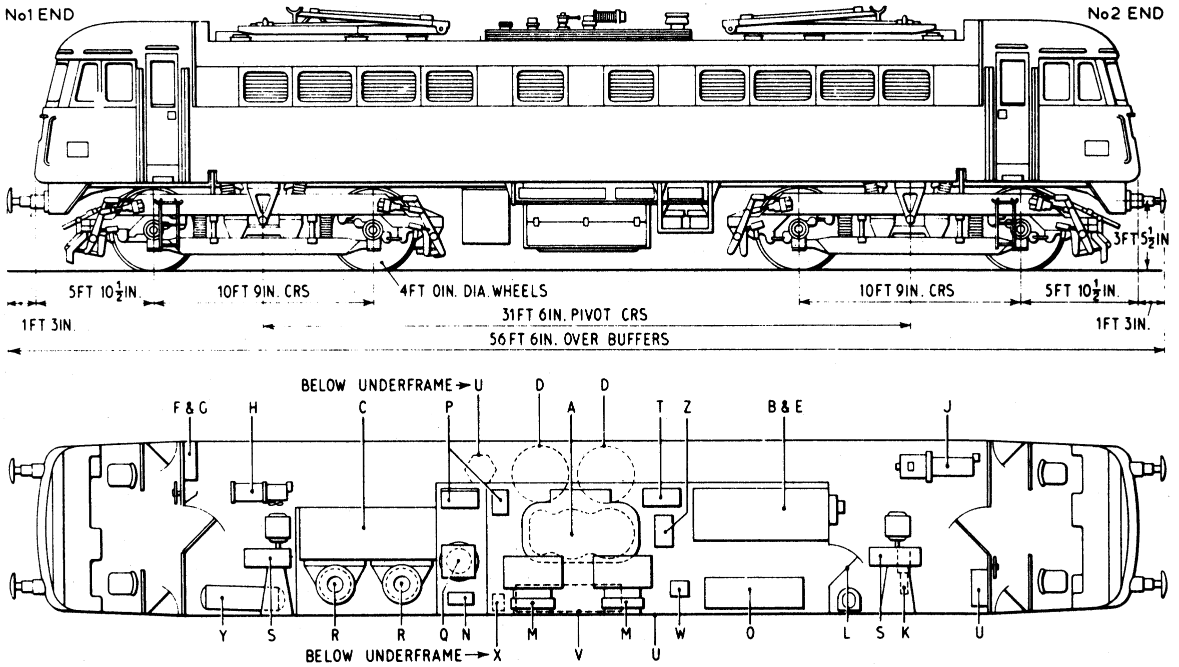


Figure 45.11 British Rail 2500 kW, 25 kV a.c. electric locomotive: A, main transformer; B, tap-changer; C, rectifier; D, smoothing inductors; E, control equipment; F, low-voltage fuses; G, fault indicator; H, main compressor; J, exhausters; K, auxiliary compressor; L, high-voltage compartment door; M, transformer coolers; N, hole storage capacitor; O, motor contactors; P, braking exciters; Q, braking resistors; R, rectifier coolers; S, motor coolers; T, auxiliary transformer; U, battery charger; V, battery; W, train heating panel; X, field divert resistor; Y, main reservoir; Z, tap-change inductor

45.2 Diesel-electric traction

Petrol and diesel engines have been used for traction purposes, both road and rail, for a number of years, but the extensive application of diesel-electric locomotives to railway service has only occurred since about 1935.

45.2.1 Locomotives

Diesel-electric locomotives may be divided into two main groups: main line and shunting. Railcars form a third group which is quite small, and in which, on the lighter cars which predominate, mechanical or hydraulic drive is more often employed.

Main-line locomotives may again be subdivided: the smaller units ranging from 450 to 750 kW, used for mixed traffic duty; and the larger units ranging from 1000 to 4000 kW for main-line service. The larger locomotives are usually arranged for multiple-unit operation and several units can be used together to haul a large train. The units are sometimes distributed throughout the train to reduce drawbar stresses and braking transients, and radio control is sometimes used to give balanced power distribution.

Diesel-electric shunting locomotives have been used for a number of years on British Rail and are mostly in the 250–350 kW range and weigh about 50 t.

45.2.2 Design

The diesel engine is primarily a constant-speed, constant-power unit, and it is necessary to convert this to a variable speed, constant power at the wheels of the locomotive. The characteristic of the series d.c. motor is ideal for traction, since it provides a falling torque with increasing speed; by varying the voltage applied to the motor the torque–speed product can be made to match the constant power of the engine. The generator, which becomes in effect a flexible power link between the engine and the motors, provides the necessary current at varying voltage.

The essence of the problem of diesel-electric equipment design is to match the generator output curve to the engine power curve. The typical generator characteristic at constant (maximum) engine speed in *Figure 45.12*(*a*) is shown by the curve ABCDE.

The voltage–current curve for constant input power at this speed is BGD. Between B and D the two do not match and engine speed falls to give the relation BFD, with loss of power. In a shunting locomotive the speed drop produces output–input balance. Alternatively, the voltage–current characteristic of the generator may be made to follow the curve AGE, but full engine power can then only be developed at G, with loss of tractive effort for other operating conditions (*Figure 45.12*(*b*)).

For main line locomotives it is desirable that full engine power be maintained over a wide range of train speeds. The diesel engine has an optimum operating zone in which the fuel efficiency is highest. For any desired output power there is therefore an optimum diesel engine rotational speed. The load regulation system normally responds to the setting by the driver of a desired engine speed. The excitation of the generator is then adjusted so that the optimum output power is taken from the diesel engine. If the train load

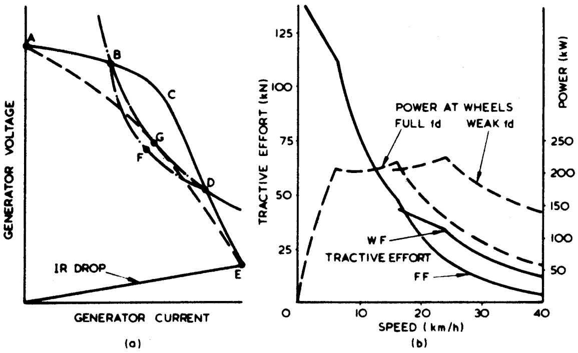


Figure 45.12 Diesel-electric characteristics: (a) voltage–current characteristics for: ABCDE, generator at constant maximum speed; BDG, constant generator input power at constant maximum speed; BFD, reduced speed. (b) Performance of 300 kW, 50 ton shunting locomotive (traction 280 kW, auxiliaries 20 kW, two motors with double reduction gearing)

reduces, the generator will run at a high-voltage, low-current condition delivering part power.

45.2.3 Electrical equipment

Electrical equipment consists essentially of a main traction generator, one or two auxiliary machines, a number of traction motors and the control equipment necessary for both main and auxiliary circuits. On large main line locomotives the auxiliary supply has to provide for compressors, traction motor blowers, lighting, heating, control and similar requirements. A battery is usually included to provide stand-by power and starting.

45.2.3.1 Generator or alternator

The generator armature is usually coupled direct to the engine crankshaft with a single bearing in the generator end housing. The auxiliary generator may be overhung from the main generator or in some cases driven together with an exciter by belts. Since the generator is a variable-voltage machine, the maximum voltage is arranged so as to obtain the most economical use of materials and the lowest I^2R losses. The maximum voltage is usually limited to 1000 V and normal operation is around 600 V.

45.2.3.2 Motors

Motors are of series type and are generally similar to those used on d.c. electric trains and locomotives. They are usually built for operation at the full generator voltage but in some cases may be connected permanently two in series. The size of motor is determined largely by the maximum tractive effort required and bears little relation to the power capacity of the diesel engine.

45.2.3.3 Control

The control equipment on a diesel-electric locomotive is concerned mainly with the following functions.

Engine-speed control The varying requirements of railway service make it necessary for the engine to have some speed control so that it can operate when required at reduced speed and power. In its simplest form a hand-operated lever, which alters the governor setting, provides a range of speeds between idling and full speed. No more is required on locomotives of the shunting type.

Generator-field control On large locomotives a speed-sensitive device is used to adjust the generator field to match the engine power.

Motor control Motors are normally connected in parallel and this provides the simplest arrangement. In certain cases, however, series/parallel control may be necessary to give the required locomotive characteristics and provision must be made for changing over the motor connections. A reverser is required for changing direction of running, and diverter contactors may be provided if the fields have to be shunted to give top-speed running.

Driver's controls On a typical shunting locomotive three control levels are provided on each side of the cab so that the driver has alternative driving positions. A time delay in the operation of the dead-man pedal is introduced so that the driver can change over from one operating position to the other without shutting off power. The engine, or master, control is used to start and stop the engine, and provides for engine testing with the locomotive at rest. A reversing lever with an off position determines the direction of movement of the locomotive. Finally, a power controller gives the driver complete control over all locomotive movement. Initial movement of this control lever prepares the power circuits and brings on generator excitation with the engine running at minimum operating speed. Further movement strengthens the generator field and finally brings the engine up to full speed.

The arrangement for main line locomotives is similar, except that a single driving position at each end of the locomotives is provided, with facilities for multiple-unit operation when double heading is required.

45.2.3.4 Starting

The usual method is to start the engine from the battery by motoring the main generator, using a special starting winding. Compressed air starting is occasionally employed, but it involves the use of a compressor for charging the air cylinders. Bendix starters, similar to the type used on motor-car engines and operated from a 24 V battery, are used on some American locomotives. In the case of most shunting locomotives it is usual to make provision for tow starting in an emergency if the battery is too far discharged to motor the generator for normal starting. This involves connecting one or more of the motors so that they act as generators to motor the main generator and thus start the engine. Although emergency starting is seldom required, it is a useful provision, since another locomotive is usually available and starting can be effected with very little delay.

45.2.3.5 Mechanical design

The design of diesel-electric locomotives follows very much the same lines as that of normal electric locomotives; main line locomotives, particularly, use axle-hung motors and bogies which are generally similar.

45.3 Systems, EMC and standards

45.3.1 Novel systems

45.3.1.1 Magnetic levitation

Research work on high-speed maglev trains has continued for several years but no public service is in operation. Two types have been taken to full-scale tests, one using controlled magnets working in attraction, the other using fixed magnets working in repulsion. Attraction systems such as the TransRapid 07, built in Germany, use a series of steel-cored magnets placed below a laminated suspension rail. The TR07 has a three-phase winding embedded in the track rail and this interacts with the vehicle magnets to provide propulsion and braking. Separate magnets give lateral guidance. The vehicle sits on a T-shaped beam. No sliding contacts are used, the traction power being supplied to the track and the vehicle auxiliaries being powered via induction pick-up from the track.

The MLU 002 of Japan National Railways uses superconducting magnets interacting in repulsion with wound coils fixed to the U-shaped track, to provide lift, guidance and traction/braking. The vehicle has wheels since levitation in repulsion is not available at low speed. No sliding contacts are required.

One maglev system is in public service at Birmingham International Airport, UK, offering a 600 m link from airport to railway station. The 6 t vehicle uses attraction magnets (which also provide guidance) with linear motor propulsion and braking. Power is supplied via sliding contacts to on-board power control for this 50 km/h vehicle.

An M-Bahn system is in use, in which permanent magnets exert an upwards force greater than the mass of the vehicle, the surplus force being taken on rollers. Other rollers provide guidance and others carry the vehicle across gaps in the magnets. Windings in the track give three-phase synchronous motor drive and braking.

The US Government has launched a maglev Initiative to produce a 480 km/h vehicle using American technology. The first full sized components are due for testing in late 1993.

45.3.1.2 Monorail

Systems using a single beam for the track, termed monorails, are in use. The Alweg design is most common, with rubber tyres used for support and guidance. Power is collected by sliding contact and the support wheels are powered.

45.3.2 Coaching stock

Modern passenger coaches use significant amounts of electrical energy. Environmental comfort requires heating and air conditioning, and many other amenities and essential services have to be provided. Power is supplied via train lines from the locomotive or from the local multiple-unit power source. Train-line voltages are typically 1000 V single phase $16\frac{2}{3}$ or 50 Hz, or 750 V d.c. On multisystem vehicles, d.c. voltages of 1500 or 3000 V are used. Depending on whether a.c. or d.c. supply is used, a transformer and/or inverter is used to supply those auxiliaries which require a.c. feeds.

Heating Resistor heating banks are provided, with thermostat control of air temperature. A coach may have 40 kW installed for this duty.

Air conditioning and pressure ventilation Power for the refrigerant compressor is required and an inverter is usually used to convert the train-line supply to three phase for this motor. A rating of 26 kVA may be required. The fans for air circulation could consume 6 kVA.

Lighting Fluorescent lighting is standard, supplied from inverters which produce high-frequency a.c. to reduce mass and avoid flicker effects.

Toilet wash water and hand dryer Hot water is required for toilets and a 1.5 kW heater is usual. If a hot air hand dryer is installed, the electrical load may be 4 kW.

Shavers, public address, telephone and audio visual service A number of small amenity loads may need to be supplied.

Battery and battery charger To ensure lighting in emergencies and sustain ventilation under fault conditions, a vehicle battery is fitted. This is typically 24 V, 440 A-h, and requires a battery charger. This charger is usually fed from a high voltage train line.

Power doors It is advantageous for train operation if doors are power operated. Electric or air motors can be used for doors. Audible warning of door closure should be provided.

Passenger information Display panels may be provided within the vehicle to announce the next station and provide route information. Train services can be publicised. Liquid crystal displays (l.c.d.), light-emitting diodes or visual-display units are available options. External displays of route information can be generated by two-state dot arrays of electromechanical discs which have black and coloured sides.

Servicing information Vehicle status can be monitored by transducers and fault condition data can be stored in a microprocessor memory, for interrogation by maintenance staff. An l.c.d. can be used to indicate areas requiring attention.

45.3.3 Electromagnetic compatibility

Electrified railways must be bonded to earth so that dangerous voltages are not produced during short-circuit conditions. The spacing and type of bond depends on the rating of the system and the distance from the feeding point.

Traction currents must not induce dangerous voltages in nearby conductors and conductor geometries are arranged as required. Leakage of d.c. current can cause corrosion to pipes and cables and bonding must be adequate to keep voltages below safe levels. In practice, when the railway has designed its systems to prevent damage to itself, nearby apparatus is normally safe. Regular inspection of safety measures is required.

Sliding current collection of power normally involves the generation of some radiofrequency noise and filters may have to be added if disturbance is caused.

45.3.4 System simulation

With the advent of fast computers, simulation of train movements on actual track layouts is feasible and can be used to study the electrical aspects of introducing a new service or rolling stock. Colour graphics are used to observe train movement and energy use.

45.3.5 Standards

Although railways tend to use their own standards and specifications when purchasing equipment, a number of standards exist. These include those listed below.

45.3.5.1 *British standards*

BS 173 Specification of rotating electric machines for traction
BS 2618 Electric traction equipment
BS 3403 Tachometer and speedometer apparatus
BS 4999 Rotating machines

Railway Industry Association 6 Buckingham Gate, London SW1 6JP, UK. Technical Specifications have been prepared in collaboration between railways and manufacturers. These include:

BRB/LUL/RIA 12 Protection of traction electronic equipment from transients in d.c. control systems
BRB/LUL/RIA 13 General specification for electrical equipment used on traction and rolling stock
BRB/LUL/RIA 18 Interference testing for electronic equipment used on traction and rolling stock
BRB/LUL/TG 22 Technical guide on e.m.c. for electronic equipment used on traction and rolling stock

45.3.5.2 *International electrotechnical commission*

A number of leaflets have been issued, giving guidance for the design of traction apparatus. These specify limits such as working voltages and permitted harmonic currents.

IEC 77 Rules for electric traction equipment
IEC 165 Rules for testing electric rolling stock
IEC 310 Traction transformers and reactors
IEC 322 Ohmic resistors
IEC 349 Rotating electrical machines
IEC 411 Power converters for electric traction
IEC 494 Pantographs of electric rolling stock
IEC 563 Limiting temperatures of electrical equipment
IEC 571 Electronic equipment on rail vehicles
IEC 631 Characteristics and tests of electrodynamic and electromagnetic braking systems
IEC 638 Criteria for assessing commutation of traction machines
IEC 737 Measures for limiting disturbance of light current installations by electric traction
IEC 801 Immunity to electromagnetic disturbance

45.4 Railway signalling and control

The purpose of railway signalling is to control the passage of trains such that they may run at speed by maintaining a safe adequate distance between following trains and safeguarding their movement at junctions. Inherent in any system, therefore, must be the ability to detect the presence of all trains and vehicles and a means to ensure separation between them.

The basis of modern signalling is the electrical track circuit to detect the presence not only of trains but also of single vehicles. A track-circuit equipment must detect the presence of a vehicle presenting a maximum resistance of 0.5 Ω between rails, a figure taking account of contact effects due to rusty rail surfaces and other features. There are several forms of track circuit, and their principles are considered first in relation to the simplest d.c. track circuit.

The fundamental philosophy of all signalling systems is that they should *fail safe*, i.e. fail to a more restrictive condition. Solid-state devices are used extensively for supervisory and information purposes, and are being introduced for safety functions such as track circuits, the transmission of vital controls and interlocking.

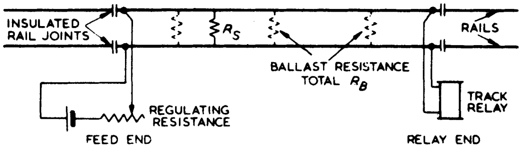


Figure 45.13 D.c. track circuit

45.4.1 D.c. track circuit

The track circuit (*Figure 45.13*) is formed by a section of the track normally isolated electrically from adjacent sections by insulated rail joints. Ordinary joints within the section are bridged by metallic bonds. Feed-end and relay-end connections are made to the rails close to the insulated joints. The rail resistance is small enough for the rails to be considered as equipotential bus-bars between which is a distributed ballast resistance R_B. The ballast resistance is the critical variable in the track-circuit performance: it varies from day to day, being low in wet conditions, high in dry and totally frozen conditions. It varies also with the type of track. Modern track is laid on reinforced-concrete sleepers, and it is necessary to ensure that the rail is isolated electrically from the sleeper. The fixings for the rail are metallic and thus the rail has to rest on an insulating pad and the clips holding down the rail are separated from the rail by insulating material. Insulation break down can cause a failure, the source of which is difficult to locate. The type of ballast may be significant. Stone ballast (preferably granite) has good resistivity; but in some areas ash ballast is found, and this can make reliable track-circuit conditions difficult to achieve. Low-resistance ballast conditions occur where the track is adjacent to the sea or is carried through undersea tunnels. A good track under the worst conditions may have a ballast resistance of 14–20 Ω per 1000 m; but on poor track it may be as low as 0.5 Ω per 1000 m.

Track-circuit adjustment must take account of variation in the battery voltage. The cell shown in *Figure 45.13* can be a trickle-charged cell which, if of the lead–acid type, has a voltage range of 1.8–2.5 V. If a guaranteed mains supply is available, the battery can be replaced by a transformer and rectifier.

45.4.1.1 *Track relay*

Standard d.c. track relays have a resistance of 4 Ω (although there are many of 2.25 or 9 Ω in service). The significant characteristics are: (a) *drop-away voltage*, at which the energised contacts ‘break’, when, after the relay has been energised, the voltage is reduced; and (b) *pick-up voltage*, at which the energised contacts ‘make’ when, after the relay has been de-energised, the voltage is increased. The drop-away voltage is a minimum of 68% of the pick-up voltage. Associated with these voltages is the train-shunt resistance R_S: (1) the *drop shunt* is the maximum resistance that produces track-relay drop-away conditions; and (2) the *pick-up shunt* is the minimum resistance such that the relay just picks up.

In practice the pick-up shunt is greater than the drop shunt and as long as this is a practical resistance the track circuit will work. If not, the track circuit will fail to re-energise.

In setting up the track circuit, the regulating resistance is adjusted for maximum battery voltage and maximum ballast resistance such that with a drop shunt of, say, 0.5 Ω the track relay is de-energised. Then it is necessary to check that the track relay will re-energise with minimum battery voltage and minimum ballast resistance. If the pick-up shunt is relatively high, say 10 Ω, then the conditions only need to become a little more critical for the track circuit to fail to re-energise. In this case, assuming that no fault exists, the selected length of track circuit is too great. On the other hand, if the pick-up shunt is relatively low under these conditions, increasing the regulating resistance will give economies of power, which is important where primary cells are used, as it gives them longer operating life.

It is normally possible to work track circuits successfully with ballast resistances down to 2 Ω. The length of track to which this corresponds depends on the ballast conductivity: on good ballast a length of 1–2 km is typical.

45.4.1.2 Fail-safe conditions

To prove the track clear, the track relay must be energised. It will de-energise if: (a) the power supply fails; (b) a disconnection occurs in the feed-end or relay-end connections, or in the rails (e.g. a rail removed); or (c) a short-circuit occurs on the track. As the de-energised state is taken to mean that a train is present, the fail-safe principle is achieved.

A wrong-side fail condition could occur if the voltage were excessive. The possibility that this condition could occur owing to a voltage from an adjacent track is eliminated where possible by reversing the supply polarity at each pair of insulated rail joints: then, should the joint insulation fail, both track circuits will fail safe. The staggered-polarity principle is applied to pointwork, where the requirements can be complicated.

45.4.1.3 Operation

The d.c. track-circuit system is reliable and economical. In certain circumstances relay timings of successive track circuits may become critical. When the supply voltage is high, the track relay may pick up faster than it drops away (which may take as long as 2 s), and in consequence a short train or single locomotive may be 'lost' for a brief time.

The d.c. track circuit will work satisfactorily in a.c. traction areas. It is necessary to ensure that the track relay is a.c. immune, to guard against false energisation from traction currents. For the same reason it is not possible to use d.c. track circuits in d.c. electrified areas and, hence, a.c. track circuits are used.

45.4.2 A.c. track circuit

The track circuit is fed by a transformer which isolates the power supply from the track. The track relay, a double-element device, has a 'local' winding connection to the same supply as the feed-end equipment (normally at 110 V), and a 'control' winding fed through the track circuit. This makes possible reliable operation with less track power.

45.4.2.1 Track relay

The two-position a.c. relay is an induction device, which is in principle the same as an induction-disc wattmeter. An aluminium vane lies in the magnetic fields produced by two electromagnets (*Figure 45.14*); one, the 'local' magnet (Q), is energised at voltage V_q from the main transformer, and the other, the 'control' magnet (R), at voltage V_r from the track circuit. The torque is proportional to $V_q V_r \sin\theta$, where θς is the angle between the voltage phasors. If V_q is large, a relatively small V_r can ensure operation of the relay, the angle θςbeing adjusted by a variation in the current-limiting devices.

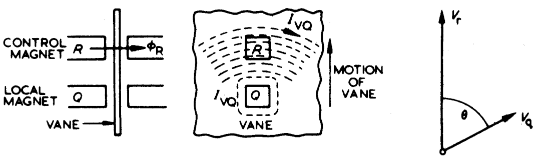


Figure 45.14 Principle of a.c. track relay

45.4.2.2 Capacitor feed

In a capacitor-fed track circuit a 1:1 isolating transformer feeds the track through a variable capacitor (*Figure 45.15*). The advantage is that there is little loss in the feed and no path for d.c. traction currents.

45.4.2.3 Impedance bonds

On electrified lines where both rails are used for traction return currents, impedance bonds are used to provide a low-resistance path for the traction current (d.c.) but a high impedance to track-circuit currents (a.c.). A simple bond is shown in *Figure 45.16*(*a*): it consists of a few turns, centre-tapped, on a laminated ferromagnetic core. Balanced traction currents give zero effective magnetomotive force (m.m.f.); but unbalance up to 20% is mitigated by the provision of an air gap in the core. The d.c. resistance of an impedance bond may be about 0.4 mΩ and its impedance of the order of 0.5 Ω at 50 Hz.

45.4.3 Jointless track circuit

In modern rail track designed with continuous welded rail, there are no joints: concrete sleepers and careful installation allow thermal expansion and contraction stresses to be accommodated.

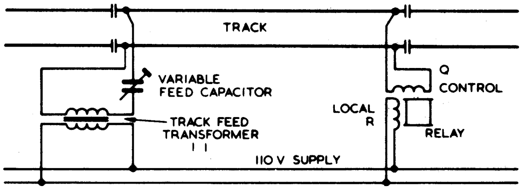


Figure 45.15 Capacitor-fed track circuit

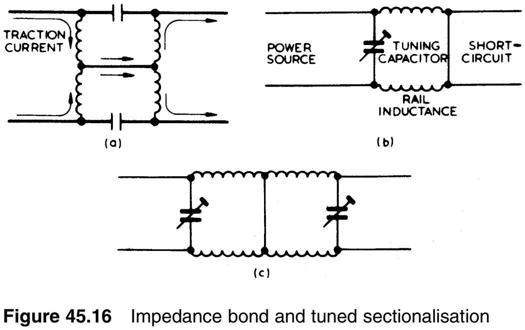


Figure 45.16 Impedance bond and tuned sectionalisation

45.4.3.1 *Audio-frequency circuits*

One method of achieving electrical separation between adjacent track circuits on jointless track relies upon tuning a short length of track using two series-resonant tuning circuits. The short circuit and rails have inductance that can produce a resonant circuit with a parallel-connected capacitor, as shown in *Figure 45.16*(*b*). With this arrangement no power from the source on the left can reach the receiver on the right. If two adjacent track circuits are required, it is possible to tune from both sides as in (*c*). A tuned length of typically 20 m results with this arrangement. The obvious disadvantage is that vehicles near to the short circuit may not be detected.

One form (*Figure 45.17*) uses a series circuit in resonance at a given frequency. Adjacent track circuits operate at different frequencies, and their limits are achieved as follows. At a frequency f_1, components C_2 and L_6 are in series resonance, short circuiting C_3 and presenting (ideally) zero impedance across EF, while C_1 is such that the whole circuit is in parallel resonance and of high impedance: thus, the end of track circuit f_1 is defined. At a different frequency f_2 (with C_2 and L_6 now off-resonance and of high impedance), C_1 and L_5 are series resonant short circuiting AB, and C_3 is such that the whole circuit is antiresonant, so defining the limit of track circuit f_2. The rails contribute inductances L_1 to L_4 to the network. Transmitter and receiver units are connected to the network through the track transformers.

Operating frequencies are a compromise between capacitor size and maximum track circuit length. With one transmitter and one receiver at opposite ends, track circuits between 50 and 1000 m in length can be operated at frequencies in the range 1.5–2.6 kHz. If a track transformer coupled with a transmitter is placed at the centre and a receiver at each end, the length of the track circuit can be doubled. Adjacent track circuits are allocated different frequencies, as are parallel track circuits on multitracked lines.

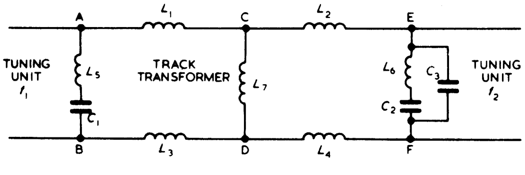


Figure 45.17 Jointless track circuits for adjacent sections

45.4.4 Other forms of track circuit

45.4.4.1 *Coded circuits*

Coded pulses (on–off or frequency–shift) may be superimposed on track circuits designed for the purpose, and detected by the receivers or by pick-up coils at the front of a train. The system can be used for train control.

45.4.4.2 *Impulse circuits*

Where wheel contact is bad, as on rusted rails or on gradients and curves where locomotives use sanding to improve adhesion, non-detection can be overcome by using high-voltage pulses detectable by the receiver.

45.4.4.3 *Reed circuits*

Mechanically coupled reed filters at each end of the track circuit ensure that only the desired frequency is fed to the track circuit and that the relay responds only to the correct frequency. The range of frequencies used is 360–400 Hz (but avoiding multiples of 50 Hz) allowing the circuit to be suitable for use in areas where both a.c. and d.c. traction are employed. By using different frequencies for successive tracks and for each parallel track on multiple lines, false operation due to failure of insulated rail joints and diverse feed paths through the traction system is avoided.

45.4.4.4 *Overlay circuits*

In all cases the relay is de-energised in the presence of a train. Sometimes interlocking controls require a train to be detected at a given spot to operate some function. This may be done by providing an ‘overlay’ circuit which energises when a train is present, additional to the normal track circuit. Frequencies in the range 20–40 kHz can be used, but the attenuation is such that detection is possible only for distances up to about 50 m.

45.4.5 Other means of vehicle detection

In addition to track circuits, various methods are available to detect the presence of a vehicle. The mechanical action of a wheel flange on the arm of a treadle is used to close electrical contacts and register that a train has passed a particular location. This method is used in association with track circuits as a diverse means of safely operating automatic level crossings. To save track-circuiting long stretches of lightly used line a system of axle counting is used. The passing of a vehicle is detected by the change in magnetic flux around a transducer caused by a vehicle wheel as it passes. When connected to an evaluator unit two such transducers can be used to count axles in and out of a section of line. Besides long stretches of single line, axle counters find use in special circumstances where track circuits are not suited, such as on lines prone to flooding and slab track formations.

45.4.6 Multiple-aspect signalling

Colour light signals are used to display a simple and logical set of indications to drivers. The displays are the same by day and night and the lens system is designed to give long-range viewing and good penetration in fog. Two groups of signal exist; main signals which are used for main line and high-speed moves and position light signals designed for controlling shunting and low-speed moves (*Figure 45.18*).

45.4.6.1 *Main signals*

Main signals are arranged to show four basic indications or aspects:

Red: stop, danger
Yellow: caution, be prepared to stop at next signal
Double yellow: preliminary caution, be prepared to find next signal exhibiting a single yellow light
Green: clear, next signal exhibiting a proceed aspect

These various aspects are used according to the density of traffic and are arranged to form two-, three- or four-aspect systems.

45.4.6.2 *Terminology*

Braking distance The service braking distance is the minimum spacing between the first caution signal that a driver sees and the red aspect to which it refers. From this distance

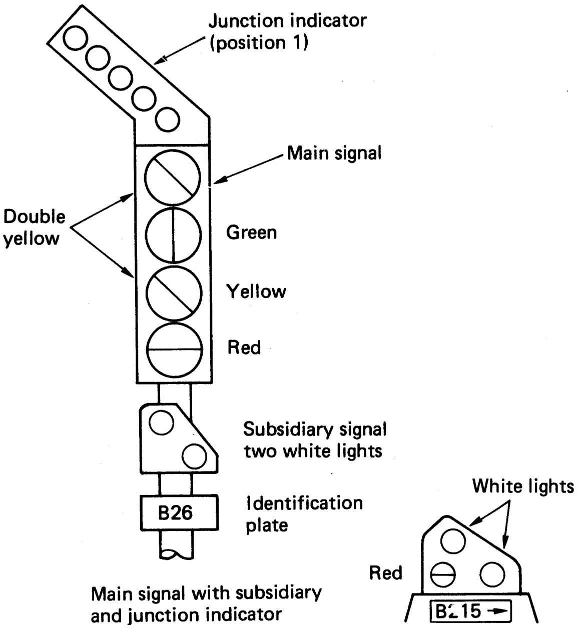

Junction indicator (position 1)
Main signal
Double yellow
Green
Yellow
Red
Subsidiary signal two white lights
Identification plate
B26
Main signal with subsidiary and junction indicator
White lights
Red
B-15

Figure 45.18 Signals

trains running at their maximum permissible speed can stop by normal service (not emergency) braking before reaching a signal at danger. The distance is derived according to maximum line speed, gradient and deceleration rate of the trains.

Sighting distance A driver on approaching a caution signal will act to reduce the speed of his train at a sighting distance from that signal.

Overlap It is practice in Britain to provide a short length of line beyond a signal to act as an 'overlap' in the event of a train overrunning a signal at danger. Normally a distance of 183 m is provided at multiple-aspect signals.

Headway This is the separation (in time or distance) between successive trains travelling at the same speed such that the following train can maintain speed.

45.4.6.3 Two-aspect signalling

Used on lines with low traffic density the two-aspect system divides the line into sections (*Figure 45.19*), each of which is protected by a red/green stop signal (B and D). At braking distance to the rear of each stop signal is a yellow/green repeater (A and C) to advise drivers in advance of the state of the stop signal to which it refers. Before a stop signal is allowed to show clear (B) the line ahead must be clear to the overlap beyond the next stop signal (D). The repeater (A) changes from yellow to green at the same time.

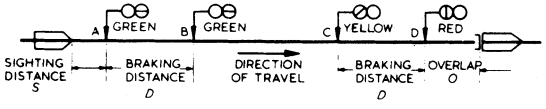

A GREEN
B GREEN
C YELLOW
D RED
SIGHTING DISTANCE S
BRAKING DISTANCE D
DIRECTION OF TRAVEL
BRAKING DISTANCE D
OVERLAP O

Figure 45.19 Two-aspect signalling

45.4.6.4 Three-aspect signalling

Where greater traffic density is required the line must be divided into smaller sections. A stage is reached where the repeater for one stop signal falls close to the stop signal in rear. The two functions are then combined to give one signal capable of displaying stop, caution and clear (*Figure 45.20*). Each signal is at least full service braking distance from the next. Signal B changes from red to yellow when the rear of the leading train passes the overlap of signal C. Signal A changes from yellow to green at the same time. The headway can now be defined: if the following train is to run at normal speed it must not approach closer to A than its sighting distance before A clears to green. For a train of length L running at speed V, the headway distance H and time T are

$$H = S + 2D + O + L \quad \text{and} \quad T = H/V \qquad (45.1)\Leftarrow$$

where S is the sighting distance, D the braking distance and O the overlap. Three-aspect signalling is usual on lines carrying medium-density traffic and can cope with trains at 5 min intervals.

45.4.6.5 Four-aspect signalling

When a three-aspect system is unable to give adequate headways in an area of dense traffic, a fourth aspect (double yellow) is introduced (*Figure 45.21*). This allows signals to be placed at half braking distance, a driver being warned initially of a red signal ahead by the double yellow aspect. The headway can now be defined as:

$$H = S + 1\tfrac{1}{2}D + O + L \quad \text{or} \quad T = H/V \qquad (45.2)$$

In practice, a four-aspect system gives around 30% greater capacity than a three-aspect system. As $D \propto V^2$ approximately, then a train travelling at 0.7 of the maximum line speed requires only one-half of the braking distance afforded by the system, thereby allowing the driver to effectively ignore the double yellow. This means that

$$H = (S + D + O + L) \qquad (45.3)\Leftarrow$$

and, therefore, greater capacity.

If one lamp of a double yellow fails, it leaves the greater restriction of single yellow, failing safe.

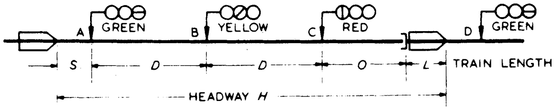

A GREEN
B YELLOW
C RED
D GREEN
S
D
D
O
L
TRAIN LENGTH
HEADWAY H

Figure 45.20 Three-aspect signalling

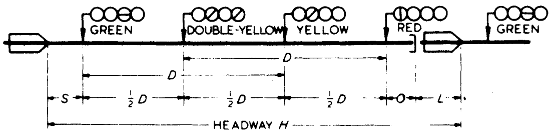

GREEN
DOUBLE-YELLOW
YELLOW
RED
GREEN
D
D
S
½D
½D
½D
O
L
HEADWAY H

Figure 45.21 Four-aspect signalling

45.4.6.6 Position light signals

These signals are used to control movements into and out of sidings, between running lines and into occupied platforms. The 'stop' indication is given by a red and white light displayed horizontally, the 'proceed' by two white lights inclined at 45°. Where a position light signal is associated with a main signal it is positioned either beneath or to the side of the main aspects. When at 'stop' it displays no indication, the main red performing this function. When showing 'proceed' the two inclined white lights illuminate and the red of the main signal remains lit.

45.4.7 Signalling for junctions

British railway practice is to employ a route style of signalling whereby a driver is informed of the direction in which the route is set ahead. Where speeds are in excess of 64 km/h the routing is given by a junction indicator mounted above the main aspects.

The indicator shows a row of five white lights in the general direction of the route to be taken. Where multiple routes exist, up to six indicator positions are available (*Figure 45.22*). An indicator must be proved alight (at least three lights) before the main signal may show a 'proceed' aspect. The main (highest speed) route is not provided with an indication.

Most diverging junctions have speed restrictions lower than for the main route, and the usual practice is to prevent the junction indicator and signal from clearing until the train has passed the previous signal, thereby enforcing the driver to begin braking. This is known as 'approach release from red'.

In many situations this is over-restrictive and so a system known as 'approach release from yellow' is applied (*Figure 45.23*). In this instance the signal beyond the junction (D) is held at red allowing the junction signal itself to show a single yellow with junction indicator (C). The signal in rear of the junction signal (B) is then arranged to show flashing single yellow to inform the driver that the junction signal is showing a 'proceed' aspect for the divergence. The driver may then slow his train for the turnout. Once the train has passed the flashing yellow signal, signal D is allowed to show a 'proceed' aspect (provided the line ahead is clear) enabling C to improve to double yellow or green. In four-aspect territory, signal A will be fitted with a flashing double yellow.

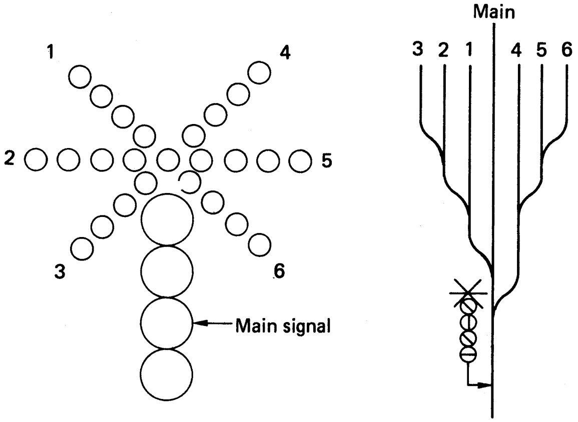


Figure 45.22 Junction indicators

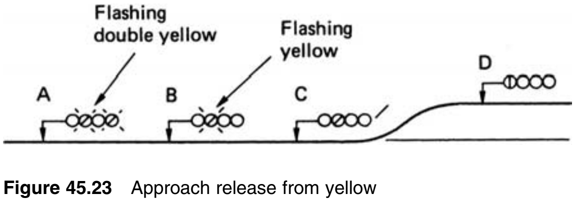


Figure 45.23 Approach release from yellow

In and around stations and other locations where the line speed is 64 km/h or less, letter or number indicators are used to identify the route with the characters formed by a selection of optical fibres lit from a single source.

45.4.8 Coloured light signals

Each aspect is provided with its own lamp and lens unit. A double-lens optical system giving an 8°⇐beam angle is employed. The 212 mm diameter outer lens is of clear glass, and has a downward-deflecting sector for close-up visibility. The 139 mm inner lens is coloured according to the aspect to be displayed. The units are set 279 mm apart in the vertical plane, are provided with hoods to screen off sunlight, and backed by matt black plates to give a dark background.

Double-filament tripole lamps are used rated at 12 V 24/24 W normally operated at 11.5 V to prolong filament life without undue loss of light flux. The main filament (with a guaranteed normal life of 1000 h) is detected by a current-sensing relay which switches the supply to the auxiliary filament if failure of the main is detected. The changeover is indicated to the control centre in order that the lamp can be replaced before complete failure.

A transformer is associated with each lamp, the feed circuit being 110 V a.c. allowing signals to be controlled from a distance. In 25 kV a.c. traction areas this is limited to 200 m to guard against false illumination brought about by induction into the cable from the traction system.

45.4.9 Signal-aspect controls

Before a signal is allowed to show a 'proceed' aspect the line ahead must be clear, including the overlap at the next signal and the next signal must be alight. The actual aspect displayed is reliant upon the aspect being shown by the next signal such that the correct sequence is always shown approaching a red signal. Therefore in four-aspect territory the sequence is green, double yellow, yellow, red.

On open track the passage of a train will replace signals to danger behind it and allow them to clear again as it progresses. At least one red indication will protect the train and be preceded by the relevant caution aspects. In areas where pointwork and diverging routes exist the signals are still replaced to danger automatically behind each train but movement of the points and selection of routes is performed by a control centre.

45.4.10 Point operation

In modern power signalling, points are moved by equipment on the ground controlled electrically. There are two forms of operation in widespread use: electric and hydraulic.

45.4.10.1 Electric

The drive machine is a d.c. split-field motor to reduce the possibility of false operation due to electromagnetic interference. The control circuits are preferably four wire.

Control is through a contactor designed to pass a control current up to 10 A at 120 V. There are separate contactor relays for the two operations: when the signalling controls require a point to be moved, then the appropriate contactor is energised first, proving that its complementary contactor is de-energised. The power is then applied to the motor. For passenger train movements it is necessary to ensure both that the point is detected in its correct position and that it is also mechanically locked. The first operation of the point machine is to unbolt this lock, then to move the point switches in the required direction and rebolt the lock. As this is completed, the point machine windings are cut off by special contacts within the machine and a special electrical snubbing circuit is brought into use by another set of contacts to prevent the motor from over-running. Also operated by the point machine contacts is a detection circuit which uses a three-position a.c. relay or two neutral polar relays: this circuit detects the point machine normal or reverse setting, including the detection of the mechanical bolt in its locked position. When the detection is in correspondence with the point machine calling, the point contactor is de-energised.

Normal operation takes about 3 s. A clutch is provided to slip if, owing to poor adjustment or an obstruction, the point machine is unable to make its full travel; and to protect the motor, a timing circuit will disconnect the machine.

45.4.10.2 Hydraulic clamp lock

The purpose of point (switch) mechanisms is to ensure that the switch mates with the stock rail without gap. A positive and definite method is to clamp the two together by an electrically controlled hydraulically operated unit. Each unit has its own electrohydraulic power pack, and control from the hydraulic unit is by means of valves which actuate two pistons, one for movement to the normal and the second for movement to the reverse position. When the movement is completed, the control circuit is de-energised. The locking of the points is achieved by the clamping mechanism, which will only make in its correct position if the switch is properly against its stock rail. Thus, in operation the clamp is unlocked, and the switches are moved and then clamped. The units include detection contacts which operate a detection circuit similar to that for a point machine.

45.4.11 The modern control centre

The control of railway signalling is today brought together into large centres responsible for the supervision of many miles of line over a wide area. The latest of these exploit microprocessor technology to its best advantage. Known as 'integrated electronic control centres' (i.e.c.c.s) they contain not only the means to manually monitor and control the railway system but also features such as automatic route setting and interfaces to adjacent control centres and information systems.

The i.e.c.c. consists basically of a number of processor modules configured on two local area networks. The networks and modules are duplicated for availability. The segregation of the system into separate networks (signalling and information) provides control over the amount of data flowing in the signalling network, in order to ensure quick response times. The response times in the information network are less critical. All of the modules are based on versa module eurocard (VME) bus architecture.

The man–machine interface is provided by the signalman's display system (s.d.s.). One is provided for each signalman up to a maximum of three per i.e.c.c. Each is housed in a workstation comprising a number of high-resolution graphic monitors, keyboard and trackerball. The graphic monitors depict the relevant control area using a special character set to display the track layout, points, signals, etc. Overview and detail screens are available plus a general-purpose display. Command input lines, output messages and alarms are handled by the general-purpose display which can also show simplified real-time movements on map displays of adjacent workstation or control areas.

Other modules include solid-state interlocking (s.s.i.) (see Section 45.4.12.1), automatic route setting (a.r.s.), timetable processor (t.t.p.), external communication system (e.c.s.), gateway system (g.w.s.), and system monitor (i.s.m.) (*Figure 45.24*).

The a.r.s. package is designed to set routes automatically in response to the approach of trains based on data available from the timetable processor. Out of sequence and late running are catered for with operating strategies designed to minimise delays and deviation whenever a conflict occurs. The signalman remains in control at all times and may set routes ahead of a.r.s. or restrict its sphere of operation. Certain shunt routes and unscheduled movements require manual route setting. This is achieved using the trackerball and associated control buttons.

The signalman uses the trackerball to position a cursor over the symbol of the signal from which the route is to be set, then presses a 'set' button. He then moves the cursor so that it is positioned over the signal at the end of the route and again presses the 'set' button. A correctly set route is shown by the normally grey track layout changing to white. The progress of the train is shown by occupation of each track circuit which is indicated by turning the relevant portion of track display red. The four-digit train description number is also displayed in line with the 'red track occupied' indication.

Actual signal indications are displayed on the graphic screens as are point positions, level-crossing and other ancillary functions. Signal, point and track-circuit names are displayed optionally upon request by the signalman.

Trackerball control is exercised over individual points and is used for many of the routine commands, aided by a set of icons depicted at the bottom of each graphic display. The keyboard is used to set up train descriptions and to interrogate the system. Additionally it is available as a fall-back mode of operation should the trackerball fail.

The timetable processor holds the various timetables appropriate to particular days of the week, downloaded from the national train service data base. Local details and short-term alterations are handled by input at a local level to ensure that the information passed to the a.r.s. processor is current. The t.t.p. also makes its contents available to the information network to assist in the running of passenger-information displays.

The gateway system is tasked with handling the interface between the signalling and information networks. It regulates the exchange of traffic between the two networks, isolating the response-critical signalling network from heavy demands of information. Incoming information from outside the i.e.c.c. area that is required by the s.d.s.s is passed into the network by the g.w.s.

System monitoring, status and fault reporting are handled by the i.s.m. which provides for a technician's interrogation facility. Time synchronisation is performed by the i.s.m. in conjunction with the Rugby radio clock.

In order to interface to other control areas or subsystems, an external communications system is provided to translate message formats to and from these connected systems.

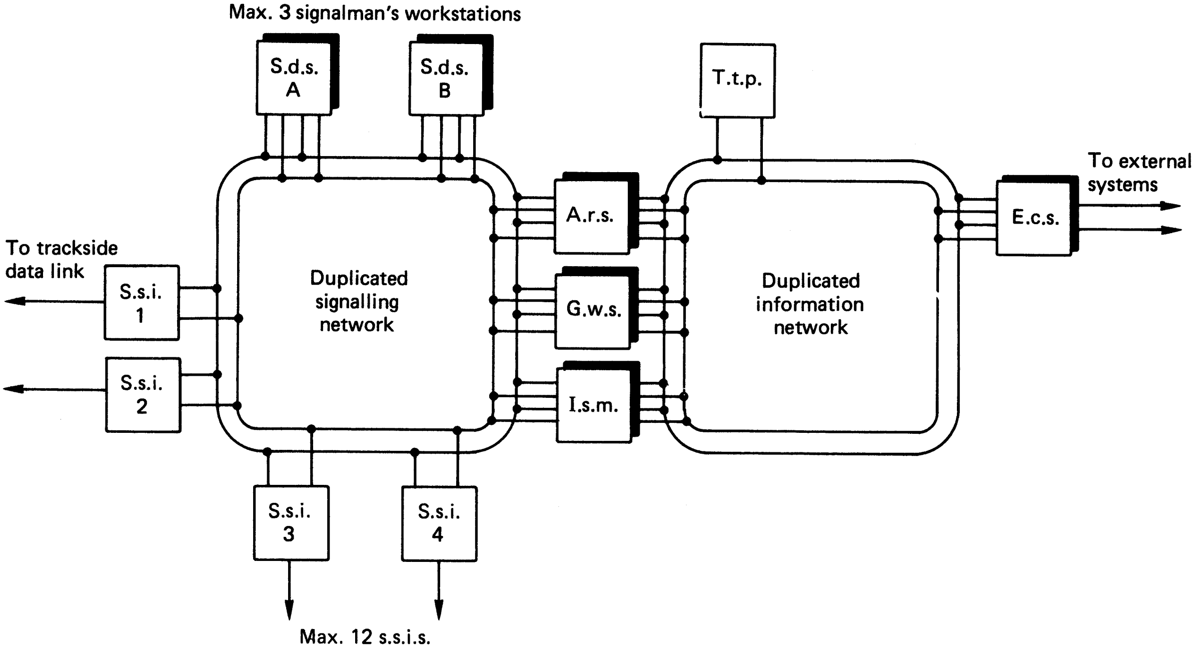


Figure 45.24 I.e.c.c. network

45.4.12 Interlocking equipment

It is necessary to ensure, before a signal clears, that the route has its points in the correct position and that there is no possibility of a train being signalled in an opposing direction. Equally, once a train has been signalled, there must be no possibility of a point in that signalled route being moved or another movement being signalled against the first-signalled move.

A signalled route generally extends from one signal to the next. In it there will probably be a number of track circuits, and a point in one of the middle track circuits. The control of that point is fundamental to the safe passage of trains: (1) the point will be locked in position by occupation of the track circuit that includes it; (2) when the signal has been operated and cleared, it must be locked; (3) if after (2) the signal has to be restored to 'danger' (e.g. in emergency), then the point must be kept locked for a time sufficient for an approaching train to stop—the time normally allowed is 2 min minimum and 4 min maximum, depending on the spacing of the signals to the rear; and (4) when a train has passed a signal which had cleared, on occupation of the first track circuit and subsequently others, the signal would revert to 'danger' and it would be necessary to ensure that the points are still locked for these movements. But there might be movements in opposite or other directions through these track circuits, in which case point locking may not be necessary. Thus, locking by these track circuits is conditional: it is called 'route locking' and is often designed to incorporate the other features listed above; and it is essentially a part of any electrical interlocking system. Another important feature is that of sectional release whereby, as a train clears each track circuit and the sections ahead are still locked, the locking of points in those track circuits that have been occupied and cleared can be released.

45.4.12.1 Solid state interlocking

Whilst a number of older systems remain in use, nowadays safety logic required by railway signalling is performed by solid state interlocking (s.s.i.), which is a purpose-built microprocessor-based system. Safety is provided by redundancy, the s.s.i. employing a two-out-of-three voting arrangement.

The interlocking is usually sited at the control centre and communicates with trackside equipment via a data link and trackside functional modules (*Figure 45.25*). The area of control is limited by the density of trackside equipment but single s.s.i. modules can be connected to form multiple interlockings to control larger layouts. Interface to the signalman's control and display system is via a standard RS422 link driver from panel processor modules which are duplicated to provide fault tolerance.

The interlocking program is data driven and comprises two parts. There is a fixed program written in assembler which has a modular structure to enable clear validation, and geographic data which configure the interlocking for its specific location. Each of the panel processors and the triplicated multiprocessor interlocking modules are fitted with an interchangeable memory module which contains the fixed program and geographic data in erasable programmable read only memory (EPROM).

The three multiprocessor modules which perform the interlocking logic tasks are identical and run the same program. They form a two-out-of-three majority voting system and each contains a redundancy management device which monitors output states, system states and program memory comparing with the other two units. A faulty module can be disconnected by its own or either of the other two management devices allowing the remaining two modules to continue operating as a two-out-of-two redundant system

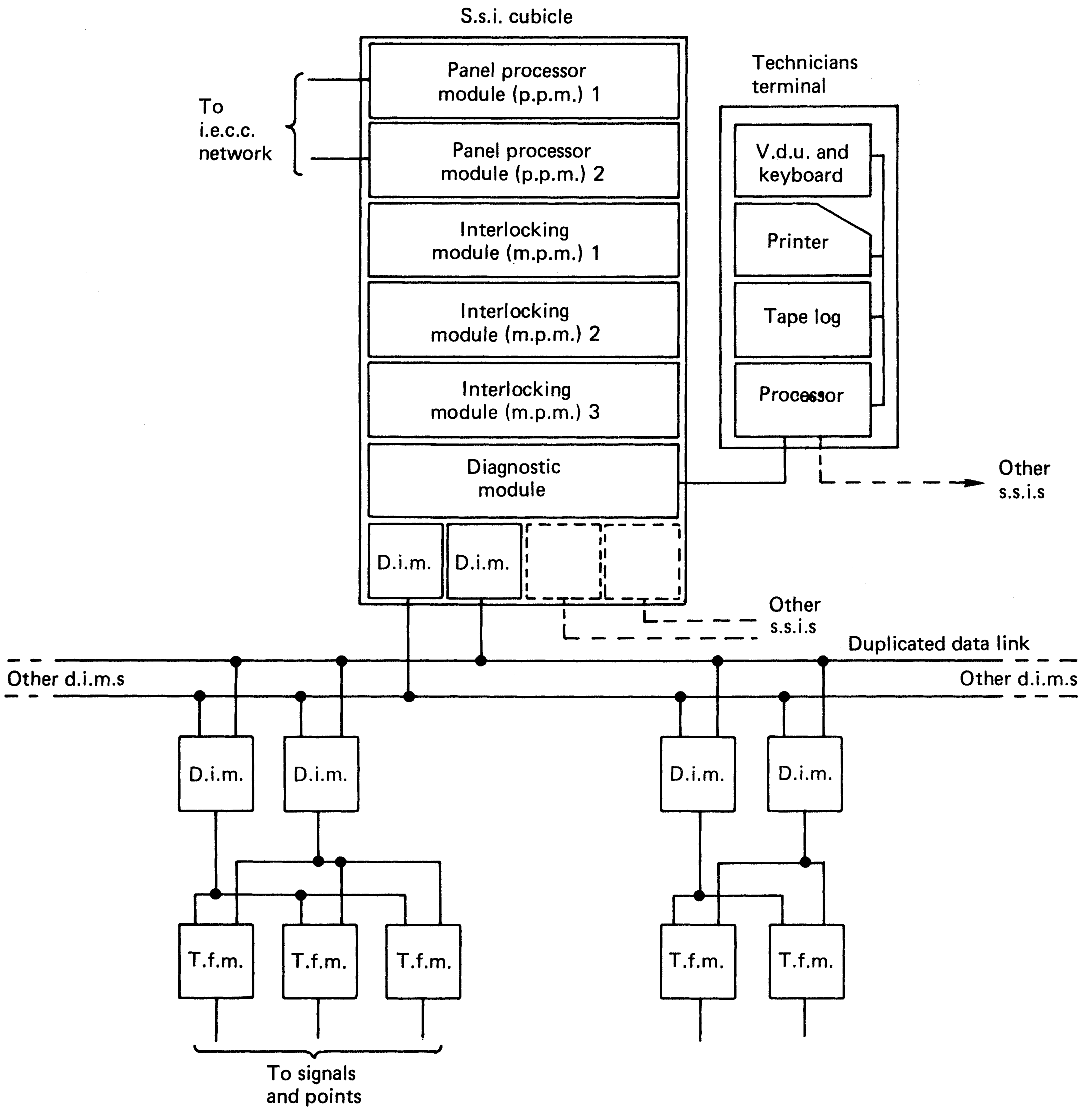

S.s.i. cubicle
To
i.e.c.c.
network
Panel processor module (p.p.m.) 1
Panel processor module (p.p.m.) 2
Interlocking module (m.p.m.) 1
Interlocking module (m.p.m.) 2
Interlocking module (m.p.m.) 3
Diagnostic module
D.i.m.
D.i.m.
Technicians terminal
V.d.u. and keyboard
Printer
Tape log
Processor
Other s.s.i.s
Other s.s.i.s
Duplicated data link
Other d.i.m.s
Other d.i.m.s
D.i.m.
D.i.m.
D.i.m.
D.i.m.
T.f.m.
T.f.m.
T.f.m.
T.f.m.
T.f.m.
To signals and points

Figure 45.25 Scheme of solid-state interlocking. D.i.m., data link module; t.f.m., trackside functional module

without a reduction in safety. If a further fault occurs in either of these remaining modules a complete system shutdown is invoked and signals are retained at red.

Connection to the lineside equipment is made through trackside functional modules (t.f.m.s) connected to a duplicated data cable via data link modules which handle the communications. The integrity of the complete system relies upon the safe transmission of data between the interlocking and t.f.m.s and on each t.f.m. performing in a fail-safe manner. Data are transmitted at a rate of 10 kbaud in each direction in the form of telegrams addressed to and from each t.f.m. Up to 63 t.f.m.s can be addressed by one interlocking, each being polled and reply received in turn. Telegrams take the form of Manchester II code containing direction, addresses, identity, status, signalling controls and parity.

The t.f.m.s contain two processors working in a two-out-of-two voting arrangement for safety, availability being less stringent than for the central interlocking.

The modules act as a serial/parallel multiplexer taking control commands from the interlocking and turning these into outputs to directly drive signal lamps and electro-hydraulic clamp locks or point machines via contactors. Other equipment is interfaced via relays driven from outputs of the t.f.m.s. Likewise, inputs from lineside equipment such as track circuits, point detection and lamp proving are converted to serial data and transmitted back to the s.s.i. (*Figure 45.26*).

Finally, a diagnostic processor is connected to a technician's terminal to provide comprehensive fault reporting and maintenance support.

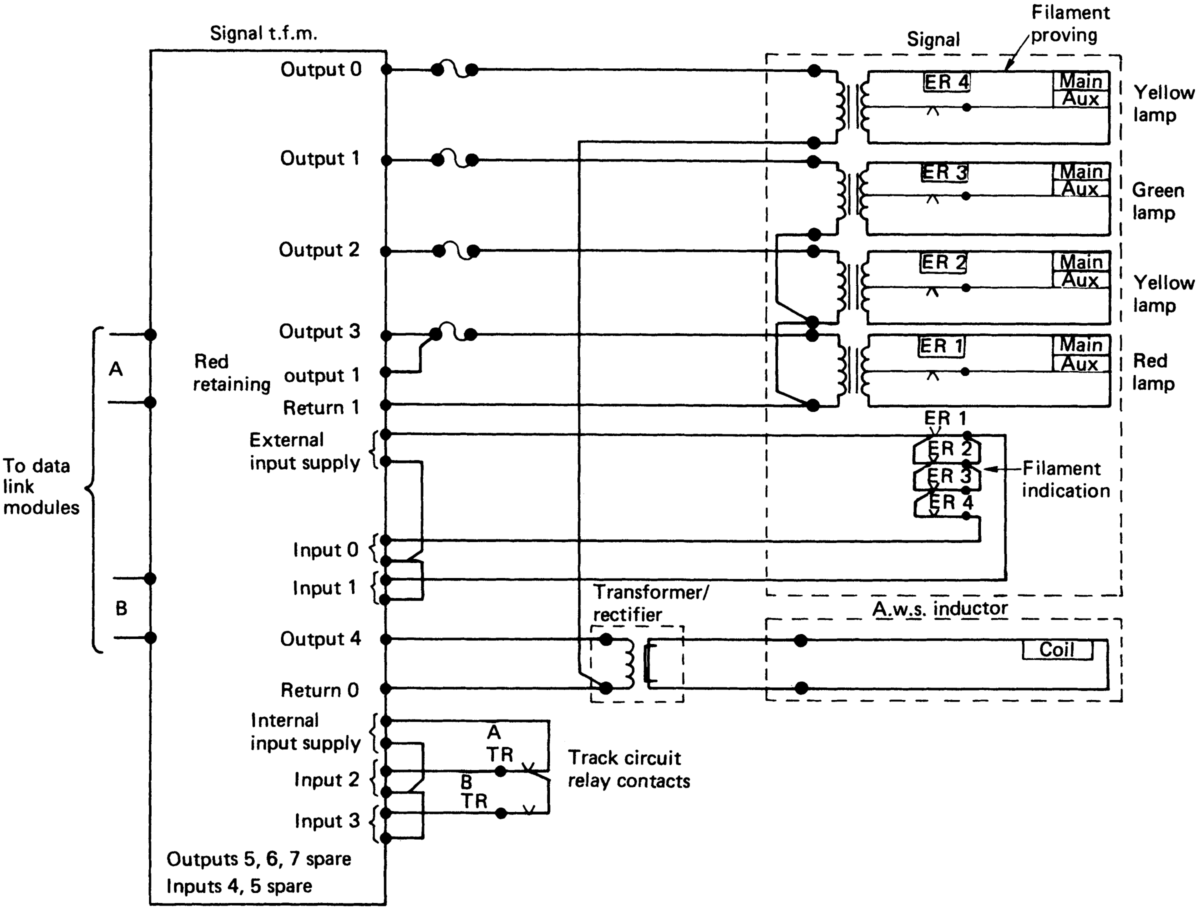

Signal t.f.m.
Output 0
Output 1
Output 2
Output 3
Red retaining output 1
Return 1
External input supply
Input 0
Input 1
Output 4
Return 0
Internal input supply
Input 2
Input 3
Outputs 5, 6, 7 spare
Inputs 4, 5 spare
To data link modules
A
B
Signal
Filament proving
ER 4
ER 3
ER 2
ER 1
Main
Aux
Yellow lamp
Green lamp
Yellow lamp
Red lamp
Filament indication
Transformer/ rectifier
A.w.s. inductor
Coil
A TR
B TR
Track circuit relay contacts

Figure 45.26 Typical trackside functional module connections

45.4.13 Automatic warning system

The standard British Rail form of in-cab signal, fitted to around 70% of the network, is intermittent in operation and gives only two indications, 'caution' and 'clear'.

Two different audible sounds can be given to the driver. If the signal ahead is showing green, an electric bell rings for approximately 1 s; but if the signal displays any other aspect a horn sounds continuously until the driver makes an acknowledgement. If the driver fails to acknowledge, a full brake application is automatically initiated. When the driver makes an acknowledgement, a visual indicator operates to remind him of the fact that he has received and acknowledged this warning, and the indication remains displayed until it is cancelled on approaching the next fitted signal.

45.4.13.1 Track equipment

The link between the track and locomotive or vehicle is magnetic, the track being fitted with two inductors, the first one to be passed over being a permanent magnet and the second an electromagnet. The magnet centres of these two inductors are 787 mm apart. The magnets have a vertical axis, the permanent magnet having its south pole uppermost and the electromagnet, when energised, having its north pole uppermost. The electromagnet is only energised when the signal ahead is clear (green). The inductors are usually fitted about 183 m on the approach side of the signal with their upper surface at rail level. They are sufficiently powerful to operate the equipment on the locomotive or vehicle through a 165 mm air gap at speeds up to 240 km/h. The permanent inductor has a magnet of Alcomax II, fitted with mild-steel spreader plates, top and bottom, to give a projected flux curve that is of sufficient intensity to operate the vehicle equipment over a horizontal distance of 250 mm measured at a height of 165 mm above the top of the inductor casing. The electro-inductor coil has two windings, and the core is similarly fitted with spreader plates to give a flux curve like that of the permanent inductor. The two coils can be connected in series or parallel for 24 or 12 V d.c., the nominal coil consumption being 9 W.

Permanent magnets are also fitted on the approach to permanent and temporary speed restrictions to give a warning and to provoke a response from drivers.

45.4.13.2 Rolling-stock equipment

Rolling-stock equipment has various forms for diesel and electric locomotives hauling vacuum-braked trains, diesel multiple-unit trains with vacuum brakes, and diesel or electric multiple-unit trains with air brakes.

On the underside of the locomotive or vehicle is a receiver responding to the magnetic flux from the track inductors. It consists of a permanent-magnet armature, centrally pivoted and lying between soft-iron poles. The armature carries a flexible beryllium–copper contact strip to act as a single-pole changeover switch when the armature moves under the influence of the south and north poles of the inductors. Its own magnetism holds it in one or other of the two positions against the vibration of the vehicle. A coil wound round one of the poles is energised by the driver when acknowledging a warning indication, the magnetic field causing the armature to return to the normal position from which it had been moved by the flux from the permanent inductor. If the signal has been clear, the receiver armature after being thrown by the south pole of the permanent inductor will be automatically restored by the north pole of the now energised electro-inductor, no acknowledgement by the driver being required.

Power for the automatic warning system (a.w.s.) apparatus is provided from the vehicle battery through a static voltage converter, which serves to isolate the a.w.s. equipment electrically from the driving control and lighting circuits and to provide the correct voltages (12 V for the main circuits and 40 V for the acknowledging or re-setting circuits). Normal power consumption is about 4 W with a peak of 12 W for a few seconds on receiving a signal from the track.

The 'caution' indication is either a vacuum or compressed-air operated horn controlled by an electropneumatic (e.p.) valve normally energised. A second e.p. valve, similarly energised and mounted in the same unit, is used to control the a.w.s. brake valve. On passing over a permanent-way inductor with the electro-inductor de-energised, the breaking of the normal receiver contact de-energises the e.p. relay, which, in turn, energises both e.p. valves. The horn immediately sounds and the brake valve begins to release in the air/vacuum system. Within about 3 s the valve releases fully and a brake application is made to bring the train to a stand.

The driver can acknowledge and re-set the equipment by pressing a re-set plunger within the 3 s period and maintain control of the train. The visual indicator changes to a black and yellow display to act as a reminder that a 'caution' signal has been passed.

45.4.13.3 Mechanical safeguards

On double-ended locomotives and multiple-unit trains, means have to be provided to isolate the a.w.s. equipment in all drivers' cabs except that from which the train is being driven. For vacuum braked stock, a combined electrical and vacuum switch is provided with a loose handle which is carried by the driver. When the driver enters a cab, he inserts the handle into the switch and turns it to close the necessary electrical contacts, which puts supply to the a.w.s. apparatus and opens the a.w.s. brake valve to the train pipe. The handle cannot be removed without first switching off to isolate. To ensure that the driver inserts the loose handle and switches in the a.w.s., additional contacts are provided through which the supply to the main driving controls is taken. The train cannot be moved unless these contacts are made. This switch also carries a second fixed handle, which is sealed in the running position. In case of a failure of the a.w.s. apparatus which would apply a permanent brake, this seal can be broken and the handle moved to isolate the a.w.s. equipment and yet allow the train to be moved.

On air-braked trains the same principles apply, but the independent a.w.s. loose handle is dispensed with and the switching effected by operation of the driver's brake control key.

45.4.13.4 Bi-directional lines

Where a line is used for signalled moves in both directions steps have to be taken to suppress the permanent magnets which are passed over in the wrong direction. This is achieved by switching on a suppressor unit comprising an inductance forming part of the permanent magnet. When switched on a magnetic field is produced that is equal and opposite to that of the permanent magnet. Having said this many lines now fitted with a simplified form of bi-directional signalling do not have this feature built into the a.w.s. track equipment. In these cases the driver must cancel and ignore the warnings received that do not apply to his direction of travel.

45.4.14 Automatic train protection

Although the a.w.s. has made an important contribution to increasing the safety record of British Rail, it cannot ensure that drivers respond correctly to a caution signal or warning of a speed restriction. A programme of fitting automatic train protection (a.t.p.) has therefore been embarked upon. Like a.w.s. the a.t.p. system is intermittent, in that information is passed to the train at fixed locations. However, the amount of information passed between track and train is sufficient to allow complete speed supervision at all times.

The trackside equipment comprises beacons, usually fitted adjacent to signals, into which are encoded data relevant to the route ahead. These data comprise such items as: present position, distance to 'danger' signal, overlap at signal, gradient, maximum line speed, distance to next speed restriction, value of next speed restriction and its length. Variable information such as signal aspect and route information is determined by selection from the signal controls. The data are encoded into a telegram and presented to the track beacon which comprises either a cable loop laid between the rails or a transponder mounted on a sleeper. On passing over a beacon the train-borne receiver passes the telegram to the on-board processor.

45.4.14.1 Speed supervision

The on-board processor is responsible for taking the trackside messages and other train-carried inputs such as speed and direction and generating supervision curves to ensure that the train does not exceed speed limits or pass a signal at 'danger'. A certain amount of fixed train-carried data is required to enable the processor to make the right calculations. This includes: length of train, maximum permitted speed, and braking capability. These are entered by the driver at the start of the journey. When all signals are at 'clear' and no speed restrictions exist in the immediate route ahead the processor will supervise the train speed to ensure that the driver does not attempt to exceed the maximum line speed or maximum permitted speed of the train, whichever is the lower. When approaching a speed restriction or signal at 'danger' the processor calculates a series of braking curves down to the target speed at the signal or restriction. There are four curves (*Figure 45.27*).

The basic curve is derived from the distance required to slow the train D_b from its present speed V to the target

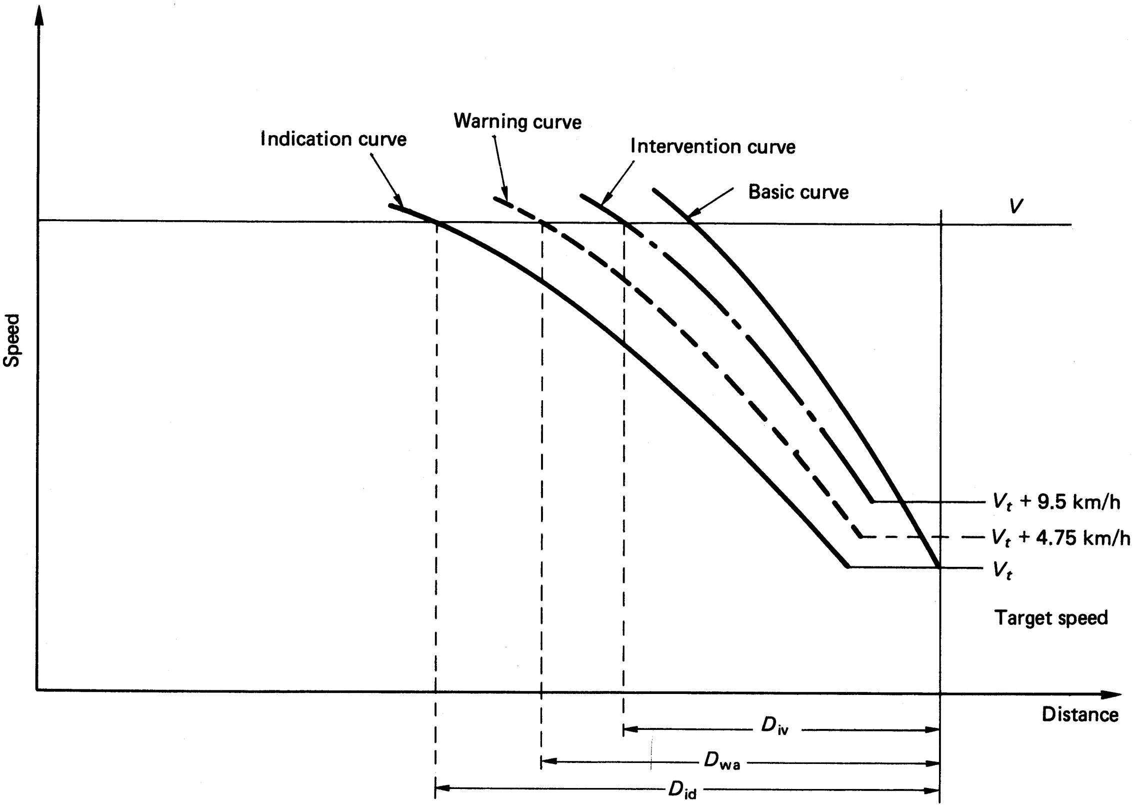


Figure 45.27 Automatic-train-protection generated curves

speed V_t where B is the deceleration on level track brought about by service braking and I is the deceleration resulting from the gradient:

$$D_b = (V^2 - V_t^2)/2(B+I) \qquad (45.4)\Leftarrow$$

In order that the on-board processor can intervene in the event that a train is not being braked sufficiently early, account must be taken of the time required for full service brake force to build up. A second (intervention) curve is therefore derived from the basic curve displaced by the distance travelled at constant speed during the time it would take for the brakes to be fully applied. This distance D_i is derived from the formula

$$D_i = D_b + VT_b \qquad (45.5)\Leftarrow$$

where T_b is the time taken to build up full service brake force. In the normal course of events the system should not be called upon to intervene as the driver will control braking well before any overspeed is detected. To advise the driver should he come close to invoking intervention a third (warning) curve is generated. The warning is based upon giving a constant warning time T_w and, therefore, the displacement of this curve from the intervention curve is the sum of T_w and the time taken for the brakes to fully apply T_b. The warning curve is thus derived from the formula

$$D_w = D_b + V(2T_b + T_w) \qquad (45.6)\Leftarrow$$

The driver requires his indication of target speed to be updated such that he is able to react before receiving a warning from the system and so a fourth (indication) curve is used to update the cab display. Again a constant time T_{id} is allowed for the driver to react to the change of indication and so the distance from target at which the curve begins D_{id} is defined as

$$D_{id} = D_b + V(2T_b + T_w + T_{id}) \qquad (45.7)\Leftarrow$$

The warning and intervention curves result in being 4.75 and 9.5 km/h above the indication curve, respectively.

In order to allow trains to draw up to a signal at red or to pass a previously red signal which has improved to proceed, trains are not supervised to an absolute stop. The release speed is determined by the length of the overlap D_o such that should a train pass a red signal it can still be brought to rest in the overlap. If the release speed is known as V_{max} and the time taken for the emergency brakes to be applied T_{be} then

$$V_{max} = \sqrt{[2(B+I)D_o]} - 2(B+I)T_{be} \qquad (45.8)\Leftarrow$$

45.4.14.2 Driver's interface

The driver's interface comprises green and yellow lights around the dial of a conventional analogue speedometer. The green lights indicate permitted speed and the yellow lights release speed. A three-character alpha-numeric display gives details of status and actions of the system. Four push buttons are supplied to cater for switching on, shunting, passing signals at danger, and regaining control after intervention. A separate keypad and display are used for entering train data. Various audible tones alert the driver to changes of state.

45.4.15 Signal power supply

As industrial power supply reliability cannot be guaranteed, all signal systems have stand-by arrangements. For isolated signals and light loads, the stand-by is in the form of trickle-charged batteries. Where a.c. is essential, inverters have been used. The normal provision for major power signalling systems is a stand-by diesel-engine/generator set. It is accepted that the set may take up to 7 s to be ready for load. Transfer back to normal supply is done manually. An acceptable main power variation is ±10% in voltage and ±2 Hz in frequency. The stand-by equipment is designed to detect wider variations and take over the load in such a case.

A normal power distribution system is based on 650 V a.c. In densely signalled areas ring mains are used, and in all other areas spur mains are used. Power supplies are sited at suitable points according to loads but, on average, power supplies are required, with associated stand-by equipment, every 16 km.

The basic signalling equipment operates at voltages of 110 V a.c. for signals and track-circuit equipment, 50 V, 24 V and 12 V d.c. for relays, and 120 V d.c. for point machines. Transformers are provided at suitable intervals to give 110 V from the 650 V main, and the other voltages used are obtained from the 110 V a.c. Batteries are normally used on telecommunications equipment and at automatic level crossings, the latter to ensure continued operation on safety grounds. In certain track layouts it is possible to operate 12 point machines at once and this would require a large transformer rectifier. As it is only an intermittent load a small transformer rectifier and heavy-duty battery are used to overcome this heavy peaking.

All signalling circuits other than track circuits are insulated from earth.

45.4.16 Protecting signalling against traction currents

A.c. traction systems are unbalanced, and as much as 50% of the return current may reach the substations through the earth, and signalling circuits are subject to induction. In order that induced voltages will not exceed the International Consultative Committee for Telephone & Telegraph (CCITT) limits of 60 V normal and 430 V under fault conditions, circuits are restricted to a length of 2 km, and in the presence of a.c. traction use a.c.-immune d.c. relays.

Connection to colour light signals is restricted to 200 m in electrified areas. Point machines are d.c. operated with a.c.-immune motors and contactors, the latter being located near the machines. For track circuits, a.c. feed is used on d.c. electrified lines and d.c. on a.c. lines. A.c. track circuits operating at frequencies of $83\frac{1}{3}$ Hz or at frequencies which are not a multiple of 50 Hz are acceptable for both forms of traction electrification.

45.4.17 Level crossings

The railway has an obligation to protect road users at level crossings on public highways. Traditionally this was done with gates, but currently there are six main ways of achieving modernised protection. These comprise automatic half barrier (a.h.b.), automatic open crossing (locally monitored) (a.o.c.l.), automatic barrier crossing (locally monitored) (a.b.c.l.), manually controlled barrier crossing (m.c.b.), trainman-operated barrier (t.o.b.) and miniature warning lights (m.w.l.).

All these crossing types, with the exception of m.w.l., are fitted with road-traffic signals. These comprise one yellow and twin red flashing lights located in each of the four corners of the crossing facing oncoming traffic. Additional units are provided for side roads as necessary. Other fixed road signs and audible warnings are provided according to the particular type of crossing and the number of lines crossed.

The road signals are lit on initiation of the crossing and show a yellow light for 3 s followed by alternate red flashing lights at a rate of 80 flashes per minute. The red signals continue to flash throughout the operation of the crossing until the train is clear and, if there are barriers, until these have begun to rise.

45.4.17.1 Automatic half barrier

These are controlled automatically by the approach of trains through track circuits and treadles. Barriers are provided on each nearside road approach only with red and white booms which carry two red lights.

A minimum of 27 s warning is given before the fastest train arrives. At initiation the yellow aspects show for 3 s followed by the flashing red for 7 s before the barriers lower. With the lowering taking 5 s the barriers are down for a minimum of 12 s. No indication of the crossing is given to the train driver but operation and power status together with emergency telephones are monitored from a supervising control centre.

The maximum line speed where a.h.b.s are installed must not exceed 160 km/h and on multiple lines the barriers must remain up for at least 5 s before a second train initiates operation. If this period cannot be achieved the barriers remain down until the second train clears the crossing.

45.4.17.2 Automatic open crossing (locally monitored)

These are controlled in a similar manner to a.h.b.s but no barriers are provided. A minimum of 27 s of warning time is given and the lights must remain extinguished for at least 10 s before a second train initiates operation on an adjoining line. In this instance a separate road signal on each side of the crossing flashes red with the words ‘another train coming’. The equipment and power supply are monitored by the train driver through a rail signal. This normally exhibits a flashing red indication which changes to flashing white once the red road signals are correctly lit. Should the red lamps in one road signal fail, or a failure of the main power supply occur, the rail signal will not show the proceed aspect. The driver is then authorised to pass over the crossing at caution once he is sure it is safe to do so.

Each a.o.c.l. is restricted to a speed dependent upon local sighting conditions but always within the maximum of 88 km/h. Before reaching the crossing and sighting the rail signal the driver will pass an ‘advanced warning board’ at braking distance followed by a crossing speed board advising of the speed restriction.

45.4.17.3 Automatic barrier crossing (locally monitored)

This system is similar in operation to an a.o.c.l. with the addition of road barriers but without emergency telephones. It is similarly subject to a speed restriction based on local conditions. The train driver receives the same trackside approach boards and monitors the operation of the crossing via the red and white flashing rail signal. During a failure situation the barriers remain raised. A plunger is provided for the driver to lower the barriers in this situation before entering the crossing.

45.4.17.4 Manually controlled barrier crossing

This type of installation is provided with full barrier protection which closes off the whole road. The barriers are normally fitted with skirts and the usual yellow and flashing red signals are shown to road users. Whilst in the lowered position the barriers are locked down.

There are several variations on the style of control but each essentially requires an operator to ensure that the barriers are lowered and that the crossing is clear before a train is allowed to approach. Controlled signals are, therefore, provided on each rail approach interlocked with the level crossing.

There are no restrictions on train speed, traffic levels or the number of lines passing over the crossing. Typically operation will be performed by a signalman or crossing keeper adjacent or near to the crossing or remotely by the aid of closed-circuit television. Automatic raising is often fitted and in many instances automatic lowering, whereby an approaching train initiates operation. Before the signals can be cleared for a rail movement, however, the operator must always confirm that the crossing is clear.

45.4.17.5 Trainman operated barrier

Used on rural lines, the train driver is responsible for operating the crossing from a switch or push button which he can reach from the driving cab. A flashing red and white rail signal indicates the status of the road signals and barriers before the driver ensures that the crossing is clear and then proceeds. A special signal is located beyond the crossing to indicate to the driver that the barriers have raised automatically behind the train. If they fail to rise the train must stop and the crew raise them manually.

45.4.17.6 Miniature warning lights

This type of installation is used for footpath crossings or minor level crossings where user operated gates or barriers are fitted. Red and green lights are provided for users on each side of the crossing, operated by the approach of trains. Users may only cross when the green light is showing. No indications are given to the train driver.

45.4.18 Train to signal box radio

To facilitate driver-only operation of suburban passenger trains a system of 'train to signal box radio' is implemented on appropriate routes. This allows for secure direct voice communication between the signalman and driver and for the exchange of fixed data messages.

Each signal box or control centre is equipped with a management processor and one or more signalman's processors each with interfaces comprising a visual display unit (v.d.u.), keyboard and audio equipment. Radio coverage is provided by a number of fixed radio stations along the length of the railway under control. These are connected back to the signalman's processor by four-wire circuits. Coverage through tunnels is maintained by using leaky co-axial cables or directional antennae. The signalman has control of the system at all times, his v.d.u. advising of all incoming calls and alarms and echoing his inputs from the keyboard. The fixed stations transmit on two frequencies, a nationwide signalling (or control) channel used to establish calls and a discrete traffic channel assigned to the area. The system is allocated 19 traffic channels in the ultra-high frequency (u.h.f.) band. To prevent cross-talk between areas continuous tone controlled signalling system (c.t.c.s.s.) is employed. Each base station is allotted a time slot in which to transmit. Synchronisation of time is handled by the management processor from reference to the Rugby radio clock. The train-carried equipment comprises a transmitter, receiver and control equipment. The cab interface includes a keypad and alpha-numeric display.

The system aids safety and operating by giving direct communication during all times without the need for drivers to leave their train and contact the control centre via signal post telephones. When stopped at a signal a driver can for instance send a 'waiting at signal' data message by operation of a function key. The signalman can respond to this with the instruction 'wait' by pressing a function key. This message is then displayed on the driver's display in his cab.

The management processor is kept informed of train whereabouts by the signalling system and is thus able to identify each caller by the train running identity and pinpoint the location of the train.

Bibliography

INSTITUTION OF RAILWAY SIGNAL ENGINEERS, Technical booklets, IRSE, Badlake Close, Badlake Hill, Dawlish, Devon

LEACH, M. E. (ed.), *Railway Control Systems*, A & C Black, London (1991)

NOCK, O. S. (ed.), *Railway Signalling*, A & C Black, London (1980)

46

Ships

R W G Bucknall

N A Haines
Royal Naval Engineering College

Contents

46.1 Introduction 46/3

46.2 Regulations 46/3

46.3 Conditions of service 46/3

46.4 D.c. installations 46/3

46.5 A.c. installations 46/4

46.6 Earthing 46/4

46.7 Machines and transformers 46/4
  46.7.1 A.c. generators 46/4
  46.7.2 Voltage build up 46/5
  46.7.3 Reverse-power protection 46/6
  46.7.4 Single phasing 46/6

46.8 Switchgear 46/6
  46.8.1 D.c. switchgear 46/6
  46.8.2 A.c. switchgear 46/6

46.9 Cables 46/7

46.10 Emergency power 46/7

46.11 Steering gear 46/7

46.12 Refrigerated cargo spaces 46/8

46.13 Lighting 46/8
  46.13.1 General 46/8
  46.13.2 Navigation lights 46/8

46.14 Heating 46/9

46.15 Watertight doors 46/9

46.16 Ventilating fans 46/9

46.17 Radio interference and electromagnetic compatibility 46/9

46.18 Deck auxiliaries 46/9
  46.18.1 Variable speed 46/9
  46.18.2 Deck auxiliary services 46/9

46.19 Remote and automatic control systems 46/10
  46.19.1 Operational modes of machinery spaces 46/10
  46.19.2 Alarms and safeguards 46/11
  46.19.3 Reliability 46/12

46.20 Tankers 46/12

46.21 Steam plant 46/13

46.22 Generators 46/13

46.23 Diesel engines 46/13

46.24 Electric propulsion 46/13
  46.24.1 Methods of propulsion 46/13
  46.24.2 Traditional electrical systems 46/13
  46.24.3 Modern electrical systems 46/14
  46.24.4 Voltage levels and harmonics 46/15
  46.24.5 Electric propulsion employing superconductivity 46/17
  46.24.6 Electromagnetic slip couplings 46/17
  46.24.7 Electromagnetic gearing 46/17

46.1 Introduction

Prior to 1950, electrical installations in ships (other than tankers) were predominantly d.c. This predominance has been reversed because of the higher power requirements of modern ships, for which a.c. systems have lower capital and maintenance costs. The main problems of this change-over have concerned the requirement, based on tradition, for variable-speed deck auxiliaries (winches, capstans and windlasses), pumps and fans. For the latter, makers have accepted single-speed or change-pole two- or three-speed induction motor drives, with additional throttle control in appropriate cases. For deck auxiliaries it was similarly found that for certain trades the change-pole induction machine was adequate, and that Ward–Leonard or other sophisticated systems could meet more demanding duty.

The development of North Sea exploration since 1970 led to the introduction of Mobile Offshore Installations generating large electrical power; their systems are invariably a.c., with d.c. conversion for operation of drilling machinery.

The concept of centralised control stations has been fully developed, with control gear, alarms and instrumentation grouped in enclosed, air conditioned and soundproofed rooms. With advances in the automatic control of steam raising plant, including the re-emergence of coal burning, engine rooms unmanned at night and with reduced manning by day have become the norm.

46.2 Regulations

With very few exceptions, every seagoing ship must comply with national, international and classification society rules and regulations. The leading classification societies (to which the administration of international requirements is sometimes delegated by the government concerned) include Lloyds Register of Shipping, Bureau Veritas, American Bureau of Shipping, Germanischer Lloyd, Nippon Kaiji Kyokai, Norske Veritas and Registro Italiano Navale. Lloyds Register of Shipping's Rules and Regulations for the Classification of Ships, which are reviewed regularly, lays down the necessary guidelines to ensure that all registered vessels are safely operated within the laws of the relevant nations. In addition to the well known larger societies, other smaller ones have emerged and unified requirements are promulgated by the International Association of the Classification of Ships.

Ships are to comply with the 1974 United Nations Resolution for the Convention for the Safety of Life at Sea (SOLAS), the 1986 consolidation and its 1988, 1990, and 1991 protocols. They must also comply with 1991 Marine Pollution (MARPOL) Consolidation which incorporates the International Maritime Organisation's (IMO) 1973 Convention on Marine Pollution and its subsequent protocols and with the acts of the various maritime nations. For British ships this includes the Merchant Shipping Act 1988. Passenger ships (carrying more than 12 passengers) must have a valid Passenger Safety Certificate and cargo vessels a valid safety Construction Certificate. Validity for communications including terrestrial radio and telex is encompassed by the Global Maritime Distress and Safety Systems which is covered in the 1988 amendments to SOLAS. The IMO is responsible for reviewing Conventions and Codes relating to the safety and operation of vessels at sea and any revisions must be accepted and acted on by the leading maritime nations.

Electrical construction and performance standards are set by each of the classification societies, SOLAS, government regulations, International Electrotechnical Committee (IEC) and, in the UK, by the British Standards Institution (BSI) and Institute of Electrical Engineers (IEE). The electrical regulations for offshore oil and gas rigs are incorporated in the IMO Code for the Construction and Equipment of Mobile Offshore Drilling Units 1991. The main IEE publication is the regulations for the Electrical and Electronic Equipment of Ships with Recommended Practice for their Implementation 1990. BSI publications include BS 2949:1960 which covers the construction and performance of rotating machinery and BS 3399:1961 which covers the regulations for transformers. The main IEC regulations are IEC 92 PT101 which cover the general requirements for electrical installation in ships, IEC 92 PT502 for electrical installations in tankers, and IEC 92 PT503 which is applicable to electrical installations with voltages in the range 1 kV to (and including) 11 kV.

46.3 Conditions of service

IEC Publication 92–101 specifies for ships on unrestricted service an ambient air temperature of 45 °C for all equipment other than rotating machines in machinery spaces, in galleys and on weather decks. For rotating machines in machinery spaces it is 50 °C. In all other spaces and for vessels on restricted service (i.e. coasters, tugs and harbour craft operating solely in temperate climate) 40 °C is recognised. For electronic devices, semiconductor diodes, etc., much higher ambient temperature conditions may have to be withstood.

Other onerous conditions are vibration, and inclination up to 15° transversely and with rolling up to $22\frac{1}{2}^\circ$. Voltage variation may be +6 to −10%, with simultaneous frequency variation of ±2.5%. With a.c. generation a momentary voltage dip of up to 15% at the generator is permissible when large motors or groups of motors are switched direct-on-line.

From the point of view of electrical equipment, very severe conditions prevail while a ship is under construction. Welding and painting will be in progress in the vicinity, accompanied by dirt and exposure to the weather.

Skilled maintenance and repairs can be carried out only in ports where suitable facilities exist. This applies particularly to machine windings. Installations must therefore be of a high standard of reliability and suitable for operation over prolonged periods with a minimum of attention. Unlike industrial conditions, in which there are shut-down or reduced-load periods, apparatus on essential ship's services may operate continuously for several days.

46.4 D.c. installations

Standard practice is to use parallel-operated level-compounded generators with equaliser connections and reverse-current protection. For small installations 110 V may be used, but 220 V is the general norm. Installations of up to 3000 kW with individual machine ratings up to 1000 kW were formerly common, but the present preference for a.c. means that d.c. systems do not now exceed 1000 kW total with smaller generating units.

The hull return system of distribution is not permitted for any tanker vessel or for any ships of greater than 1600 t gross tonnage, unless an exception is granted by the appropriate classification society. Exceptions to this rule are:

(1) impressed current cathodic protection, and
(2) insulation monitoring devices.

46.5 A.c. installations

Tankers and passenger ships of recent construction are almost all equipped with a.c. systems—about 40% with a frequency of 50 Hz; the remainder of 60 Hz. As most marine generators and motors are special to this service, the choice of frequency does not have to be related to particular national supply systems. The frequency of 60 Hz gives the advantage of higher operating speed and lower weight. Motors built for 440 V, 60 Hz can operate from 380 V, 50 Hz shore supplies and (if some additional heat can be tolerated) on 415 V, 50 Hz. However, contactors and voltage operated relays may not always be amenable to such conditions; and 50 Hz motors may not operate satisfactorily on 60 Hz, particularly with centrifugal fan and pump loads.

In some tankers and passenger ships, high-voltage generation has been adopted, particularly in those vessels that are electrically propelled. IEC Publication 92, *Electrical Installation in Ships*, Part 503, and BS 3659 (replaced by BS 5311:1991:Parts 1–7) apply to vessels with high-voltage generation.

46.6 Earthing

Regulations permit isolated or earthed neutral systems, except for tankers, in which earthed systems are forbidden. The choice (where there is one) is almost always for isolation, the risk of overvoltage being accepted to avoid the loss of a vital service, e.g. steering, should one earth fault occur. A single earth fault on an insulated system can be detected, and does not result in an outage unless a second fault occurs; and in any case overvoltages are rare compared with the incidence of earth faults. Every insulated distribution system, whether primary or secondary, is required to be provided with means to continuously indicate the state of insulation from earth and is to be arranged in such a way as to give warning of abnormally low levels of insulation. However, care must be taken when designing the installation to protect electronic circuits, particularly those containing semiconductor diodes, from overvoltage 'spikes', direct or induced.

46.7 Machines and transformers

The construction, installation and performance of rotating machinery on ships is dealt with in BS 2949:1960, BS 4999:1992, and IEC 92 PT202. Transformers for high voltage discharge lighting should comply with BS 3535:1990. Other transformers are covered in BS 3399:1961, BS 3535:1990 and by IEC Publication 92–303. It is recommended, except for those used for motor starting, that they should be doubly wound, i.e. with separate windings and that dry types should be used in preference to 'wet' types.

Because relatively large motors and groups of motors are started direct-on-line, the consequent voltage dip is important because of its effect on the system as a whole and, in particular, on lights, contactors and voltage operated relays. Stipulations are:

(1) A limit of 15% voltage dip at the generator terminals and a recovery to within 3% of rated voltage within 1.5 s, when the generator is subjected to a suddenly applied symmetrical load of 60% rated current at a power factor between zero and 0.4 lagging; the recovery may be increased to 4% within 5 s for emergency generators.

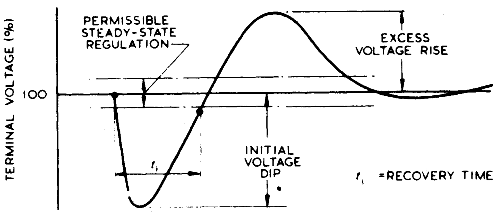


Figure 46.1 A typical voltage response

(2) A limit of 20% excess voltage following initial recovery.
(3) Under steady-state conditions the system, which may include an automatic voltage regulator, must be maintained within 2½% rated bus-bar voltage (3½% for emergency sets). The conditions are shown in *Figure 46.1*.

These requirements are for normal installations. Special conditions apply if the impact load exceeds 60%, or if the deck machinery consists of groups of multi-speed cargo winches liable to be switched simultaneously, or in any other special conditions.

It is common practice in modern ships to use self-excited compounded a.c. generators, or brushless machines with a.c. exciters and shaft mounted rectifiers (*Figure 46.2*). In order not to nullify short-circuit protective gear, such generators must maintain adequate voltage under short-circuit fault conditions. BS 2949:1960 specifies a current of at least three times rated value for 2 s unless provision is made for a shorter duration without impairing safety.

46.7.1 A.c. generators

Self-excited self-regulating a.c. generators are now available in all sizes for marine service, with excitation obtained from a three-phase exciter through semiconductor rectifiers. The principles are the same whether a slip-ring or a brushless form is applied. Voltage and current transformers connected to the generator output feed field excitation through a three-phase rectifier to give a compounding effect. A typical arrangement (*Figure 46.3*) combines compounding with closed-loop control: the effect on the excitation due to the compounding is supplemented by a fine control, the response of which is determined by the divergence of the generator voltage from a pre-set reference. A transient dip of generator voltage during the first period following a

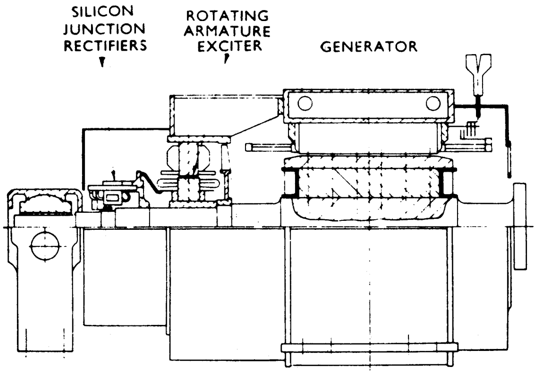


Figure 46.2 Cross-section of a brushless a.c. generator

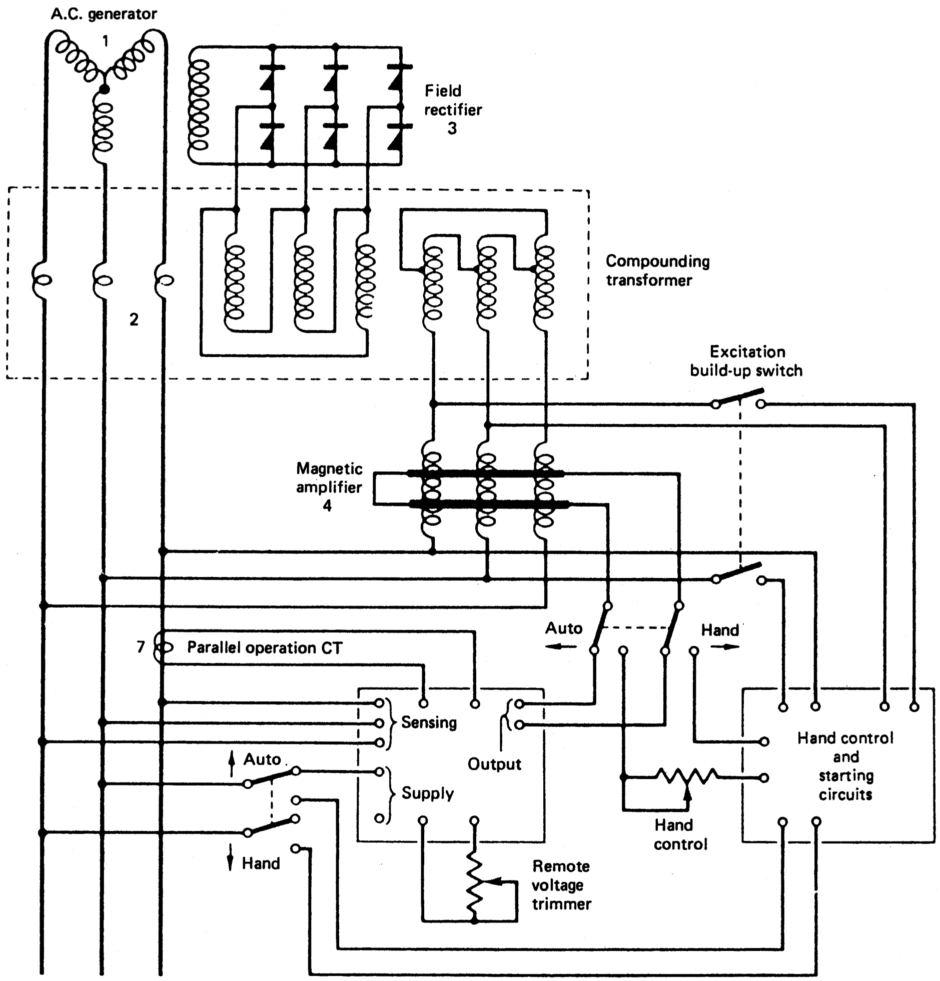


Figure 46.3 Schematic diagram of a static-excited a.c. generator

disturbance is inevitable, but no further change due to armature reaction occurs, because the excitation is rapidly corrected by field forcing through the action of the series windings of transformer 2.

Under short-circuit conditions saturation of the static excitation components limits the excitation to about 1.5 times full-load value, and the sustained short-circuit current is about three times rated value. Variations in field resistance are swamped by the series impedance of the magnetic amplifier. The circuit self-compensates for speed changes, as the amplifier current rises as the frequency falls. The automatic control circuit incorporates a Zener diode which controls a silicon transistor driving a thyristor. The three-phase rectifier is connected to the generator main field through slip-rings.

An example of a brushless system is given in *Figure 46.4*. Excitation is provided partly by saturable current transformers and partly by a three-phase linear inductor, all of small rating and easily accommodated on the switchboard. There is sufficient magnetic remanence in the a.c. exciter to ensure starting. By means of the automatic voltage regulator (a.v.r.) a steady voltage regulation of ±1% is obtained with high-speed response, typically within 0.5 s.

Governors on prime-movers are required to automatically maintain rated speed within a momentary maximum variation of 10% and a maximum steady state variation, i.e droop not exceeding 5%. These figures are to be met during tests whereby a fully loaded generator is subjected to a step 50% load reduction for a short period before full load is restored.

46.7.2 Voltage build up

In *Figure 46.3* the no-load terminal voltage is applied to magnetic amplifier 4; its current lags this voltage by nearly a quarter-period, passes through the compounding current transformer 2 and the field rectifer, and supplies the no-load excitation. When the generator is on load, the load current passes through the series coils of current transformer 2; the secondary current is then the phasor sum of the inductor current and a current proportional to the load current. By correct proportioning the static excitation is appropriate for all normal loads, even at low lagging power factors.

With brushless generators two rectifiers in parallel are provided in each phase of the rotating element to provide

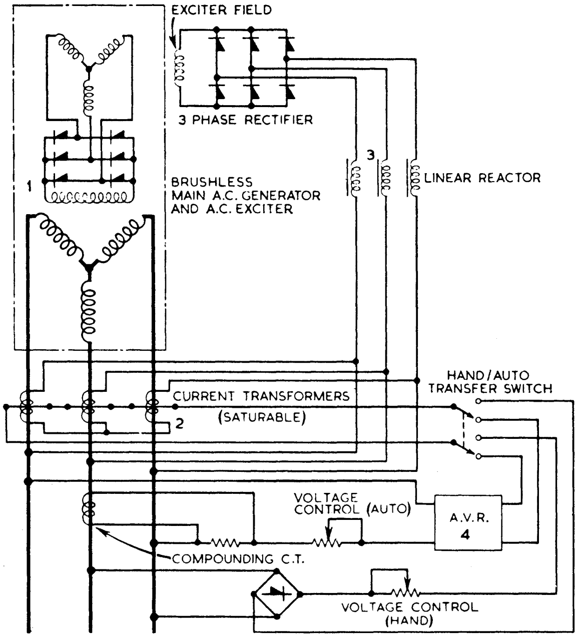


Figure 46.4 Schematic diagram of a brushless a.c. generator. C.T., current transformer; A.V.R., automatic voltage regulator

back-up should one fail. The commonest diode failure is a short circuit, so that each diode must be fused. A failed doide produces an unbalance and a ripple in the exciter field current. This can be used to detect failure.

46.7.3 Reverse-power protection

Loss of power from the prime-mover may be accidental or intentional. If the generator was left connected in parallel with other generators, it would act as a motor and continue rotating, with possible damage to the prime-mover. With d.c. generators this is taken care of with a reverse-current tripping relay, but with a.c. a reverse-power relay is necessary. To prevent inadvertent operation of reverse-power relays due to power surges, particularly when synchronising, a time-delay feature is incorporated. With diesel engines a fairly coarse setting of the order of 10–15% of full power is suitable, but steam turbines absorb very little power when motored and a fine setting of 2½–3% is necessary.

The alternative to reverse-power relay protection is to provide electrical interlocks or contacts which will respond to predetermined occurrences, such as failure of lubrication, closing of fuel or steam admission valve, operation of overspeed governor or excessive back pressure.

46.7.4 Single phasing

A common cause of motor burn-out is single-phasing, which can arise from broken or faulty connections or, more commonly, the blowing of one of the three-phase fuses. For this reason the IEE Regulations for the Electrical and Electronic Equipment of Ships with Recommended Practice for Their Implementation 1990, stipulate that matching cartridge fuses be used. In addition, overload relays with alarms and indication of motor failure to start must be fitted. This is particularly important if automatic starts are employed or if the machinery spaces are operated unmanned.

Undervoltage releases do not provide protection. Should the open circuit occur on the motor side of the circuit-breaker, the coil will continue to be fed from the supply, and if the open circuit is on the supply side, the coil will be fed by voltage induced in the motor.

A three-phase motor will not restart with an open-phase connection: out of sight and sound it may remain stalled, and if the overload protection is set too high or with too long a time delay, the motor may burn out. Pilot lights across one phase will give a false indication if the fault is on another phase. The remedy is to set overload protective devices closely, with suitable time lags, or to adopt some form of single-phase protection.

46.8 Switchgear

46.8.1 D.c. switchgear

Open switchboards have all essential switchgear exposed on the front. Some owners prefer dead-front construction. For the open type all parts, front and back, are readily accessible for maintenance, an important consideration when it is remembered that these operations have to be performed on a live board.

Regulations state that for each d.c. generator installed which does not run in parallel, a double-pole circuit-breaker or a fuse for each pole and a double-pole switch is required. For generators that run in parallel in a two-wire-insulated system a double-pole circuit-breaker is required for each machine. With compound generators the equaliser switch is required to be interlocked with the circuit-breaker of each machine so that it is made before the circuit-breaker can be closed and not opened until the circuit-breaker is opened.

Regulations require preferential tripping in large installations so that non-essential loads can be automatically switched off when the generators become overloaded. This can be done in either one stage in a simple installation or in two or three stages in larger systems. A typical arrangement is shown in *Figure 46.5*.

46.8.2 A.c. switchgear

Because of the greater risk of shock in a.c. systems, the open construction is not permitted and switchboards must be ‘dead-front’. In the usual construction the switchboard is divided into cubicles. Circuit-breakers can be withdrawn and isolated from the bus-bars for maintenance and adjustment. Access doors are interlocked to prevent access to live parts. A similar construction is used for control gear.

Preferential tripping as described for d.c. systems is also required. Instrumentation must include a synchroscope (see *Figure 46.6*). Automatic synchronising is now becoming necessary in installations comprising prime movers which start automatically, and SOLAS and its protocols require automatic start-up and circuit-breaker closure within 45 s to ensure continuity of supply with unmanned installations.

Circuit-breakers are required to comply with the requirements of BS 4752 (replaced by BSEN 60947:1990). Air-breakers and miniature circuit-breakers are common at 440 V; high fault levels have led to the introduction of current-limiting breakers. At 3.3 kV and above, vacuum

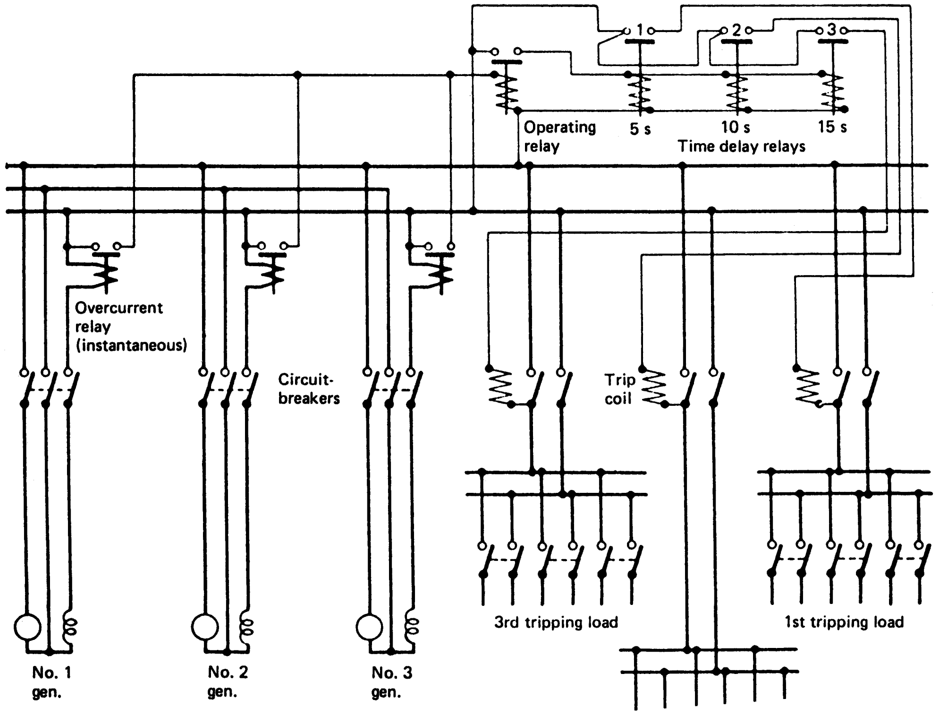

1
2
3
Operating relay
5 s
10 s
15 s
Time delay relays
Overcurrent relay (instantaneous)
Circuit-breakers
Trip coil
3rd tripping load
1st tripping load
No. 1 gen.
No. 2 gen.
No. 3 gen.

Figure 46.5 Diagram of typical preferential tripping circuits

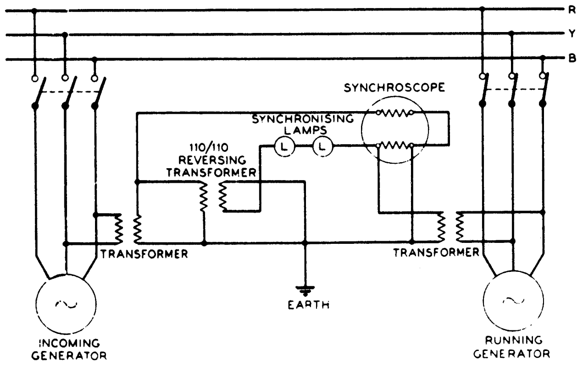

R
Y
B
SYNCHROSCOPE
SYNCHRONISING LAMPS
110/110 REVERSING TRANSFORMER
L
L
TRANSFORMER
TRANSFORMER
EARTH
INCOMING GENERATOR
RUNNING GENERATOR

Figure 46.6 Schematic diagram of a synchroscope and synchronising lamps for a 'lamps bright' system

contactors are often used; active consideration is given to SF_6.

46.9 Cables

EPRCSP, PVC and XLPE, all with metallic sheath and PVC overall, are most common, being flame retardant as required by SOLAS. MICC and silicone rubber withstand high temperatures and have appropriate applications.

Cables which are required to be flame proof, for example those in tankers and gas carriers, are required to retain insulating properties after severe fire damage. The regulations governing these types of cable are covered by IEC 331 and by BS 6387:1991.

46.10 Emergency power

SOLAS and its protocols have extended for all ships the services to be supplied; for cargo ships the period for which this power must be available is stated.

Passenger ships are to have emergency power from a generator or battery adequate for 36 h duration. With a generator there must be a transitional battery to supply specified services for 30 min to come into operation automatically.

Cargo ships have similar requirements for 18 h duration, but do not need a transitional battery if the emergency source can start and be connected automatically.

Mobile offshore installations are covered in the MODU Code.

Both nickel–cadmium and lead–acid batteries are used for emergency or stand-by services to vital circuits such as computer communication, fire detection, etc., without 'excessive volt drop', a term interpreted by the British authority as a drop, at the end of the specified emergency period, not exceeding 12½% of nominal system voltage. The voltage should be within the limits −10% to −12½% from the fully charged condition to that at the completion of the prescribed duty.

Battery installation, ventilation and maintenance correspond to the best practice ashore.

46.11 Steering gear

Electric steering can be either all-electric or electrohydraulic. Its function is to control the position of the rudder through an angle of 35° each side of the central position. Regulations require that the time taken to put the rudder from 35° on

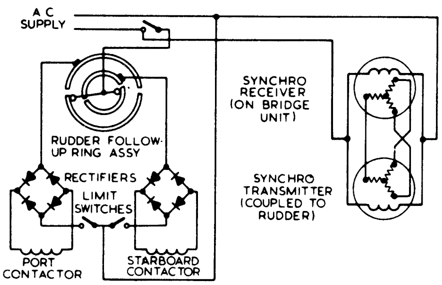

A C SUPPLY
SYNCHRO RECEIVER (ON BRIDGE UNIT)
RUDDER FOLLOW-UP RING ASSY
RECTIFIERS
LIMIT SWITCHES
SYNCHRO TRANSMITTER (COUPLED TO RUDDER)
PORT CONTACTOR
STARBOARD CONTACTOR

Figure 46.7 Schematic diagram of a rudder follow-up system and the synchro transmitter and receiver (Sperry system)

one side to 30° on the other is not to exceed 28 s at maximum service speed. The steering gear must be provided with running indicators, one on the bridge at the main steering position and one in the engine room. A usual arrangement is that one quarter-turn of the steering wheel corresponds to one degree of rudder movement. Automatic steering on a set course controlled by a gyro-compass can be superimposed on a power-operated system.

One such system is shown in *Figure 46.7*. Mounted in the bridge control unit is a follow-up ring assembly consisting of two pairs of silver rings mounted on insulating formers. Carbon brushes on the outer rings are connected to rectifiers which, in turn, are connected to contactor coils. Follow-up rollers make contact with the inner surfaces of the rings, one half-ring of each inner ring being connected to the complete outer ring. The carbon brushes are fixed, but the rings and the bracket carrying the inner contact rollers are free to rotate. The inner roller assembly is rotated by a synchro receiver; the contact rings by the pilot wheel when under hand control or by the gyro-compass transmitter when under automatic control. In this system the transmitter is geared to the rudder so that it adopts a corresponding angular position. When the rudder has reached the angle set by the pilot wheel (or by the compass), the rollers will have caught up with the gap in the ring and the contactors will open.

Electrohydraulic systems usually depend on a continuously running motor coupled to a variable-delivery pump supplying oil pressure to one or other of a pair of hydraulic cylinders. Operation from the bridge is by telemotor control, in which the rudder motion continues until its position coincides with that of the bridge setting.

All-electric systems depend either on a direct-coupled reversing motor or on Ward–Leonard control. In the latter the generator voltage may be controlled by a voltage divider or by field windings of opposite polarity. In push-button systems the rudder movement continues so long as the push-button is held closed.

Steering is a vital service. SOLAS now requires duplication of power supplies and control circuits, the latter to be supplied specifically from the associated power circuits. Alarms and transfer switching are to be sited on the navigating bridge. Large ships require steering power to be available, automatically within 45 s, from the emergency source on loss of main power.

46.12 Refrigerated cargo spaces

The system now commonly used consists of batteries of brine cooled pipes over which air is circulated by fans and distributed uniformly to all parts of the hold. The fans vary in size and number according to the holds, and a ship may require up to about 40 fans of 1–8 kW rating. Compressor motors up to about 200 kW at constant speed are commonly required. The fans are located in the holds and their lubrication presents a problem, as the ambient air temperature can vary between tropical conditions at the loading port and approximately −20°C when fully refrigerated.

Temperature control of the holds within narrow limits is essential for some cargoes (e.g. bananas), so that provision must be made for accurate and sensitive sensing and control at numerous points. Electrical thermometers have been specially developed to read to 0.1°C over a range of approximately −20°C to +15°C. A tolerance of ±0.1°C at the freezing point of water is attainable under working conditions.

Container ships with typical installed power of 6 MW, 450 V, 60 Hz and considerable instrumentation are now common. Up to 3000 containers of rating 3.5 kW may be supplied through flexible connections.

46.13 Lighting

46.13.1 General

For general lighting both tungsten and fluorescent lamps are used. For deck lighting halogen floodlights or high-pressure mercury lamps may be installed. Where colour rendering is not important, sodium lamps, which have high efficiency and long life, are suitable for high-level general lighting. The lighting load for large public rooms can be appreciable, and fluorescent lighting can reduce power and heat, which, in turn, assists air conditioning.

Under the Merchant Shipping Act, the Department of Trade requires artificial lighting in crew accommodation to be well diffused, avoiding glare and deep shadows. Bunk lights must be provided, and 25 W or 40 W tungsten or 15 W fluorescent lamps are considered satisfactory. In cold stores separate light fittings of robust construction, with a switch outside the compartment and a pilot light, are required.

For general lighting the level of lighting measured at a height of 0.85 m above floor level and midway between adjacent lamps, and between any lamp and a boundary of the space, is prescribed in precise terms. Good lighting in galleys is important, and brings faster working, fewer mistakes and accidents, and better hygiene.

Statutory regulations (when applicable) require emergency lighting supplied from the emergency source of power to be embodied in the lighting system. All boat stations, lifeboat launching gear and all public and crew areas, alleyways, service spaces, stairways and exits must be provided with emergency lighting.

Voltage variation affects the light flux output and life of both filament and fluorescent lamps.

46.13.2 Navigation lights

The requirements are prescribed in the International Regulations for Preventing Collisions at Sea 1972, as amended. All ships are to be provided with 'steaming lights', masthead, side, stern and anchor. Each navigation light should be provided with primary and alternative lanterns. Classification rules require that navigation lights shall be connected to a distribution board reserved solely for this service, and connected directly or through transformers to

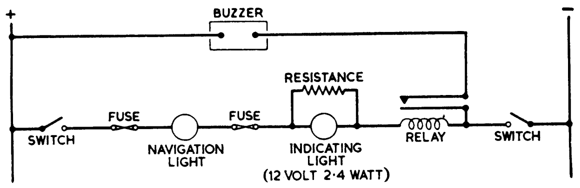


Figure 46.8 Schematic diagram of a typical navigation light indicator

the main or emergency switchboard. Each light must have aural and/or visual indication of failure. If only an aural device is fitted, it must be battery operated. If a visual signal is connected in series with the navigation light, there must be means to prevent extinction of the navigation light through failure of the signal lamp. Typical arrangements are shown in *Figure 46.8*. The volt drop across series connected indicators must not exceed 3% of the system voltage. The use of double-filament navigation lights is not permitted.

46.14 Heating

SOLAS prescribes that electric heaters, where used, must be fixed in position and must not have an element so exposed that clothing, curtains, etc., can be scorched or set on fire.

Except in ships employed solely in the tropics, crew accommodation has to be provided with a heating system of prescribed minimum performance. With certain exceptions a temperature of 19°C must be maintained in a ship regularly employed otherwise than as a ‘home-trade ship’, and 15°C in the case of any other ship, when the outside air temperature is −1\°C.

46.15 Watertight doors

Watertight doors, under certain conditions prescribed by regulations, have to be power operated and to conform to specific requirements, including testing at maker’s works in the presence of a surveyor.

They may be either electrically or electrohydraulically operated and controlled, either in groups or individually from a central control position, usually the navigating bridge. In addition, hand operated gear must be provided which can be worked at each door or from a position above the bulkhead deck. At each control position an indicator must show whether the door is open or closed. An audible warning device situated near the door and operated by a time switch giving about 10 s notice functions before it starts to close.

The central control station has overriding features which decide whether the doors are to close; they can be opened locally but re-close automatically. They can also be closed locally independently of the central control.

An essential feature of the powering system is that it should exert an initial high effort to withdraw the door from the grip of door wedges which come into effect in the fully closed position. Thereafter this force diminishes and the rate of opening increases.

46.16 Ventilating fans

When trunking carries ventilating air to accommodation and cabins, any noise arising from fans or fan motors must be avoided. For this reason sleeve bearings are generally preferred. A variable speed or a choice of speed is usually necessary for control purposes.

Regulations require means to be provided for stopping fans from remote positions in the event of fire.

46.17 Radio interference and electromagnetic compatibility

Radio communication is vital and must as far as practicable be interference-free. ‘Noise’ may originate in the ship’s electrical installation, in the rigging, from other radio frequency apparatus (e.g. public address systems) and from electromedical appliances. The radio installation can also cause interference with other electronic circuits; hence the concept of electromagnetic compatibility.

Suppression of unwanted noise is an economic problem to be shared by the supplier of the radio equipment, the shipbuilder, the electrical contractor and the manufacturers. Usually interference cannot be entirely eliminated, so permissible levels are prescribed. Interference can be alleviated by careful planning of radio installations and aerial systems, by applying certain techniques to the rigging, and by fitting suppression devices to electrical equipment and wiring. Interference arising from the electrical installation can be of two kinds—i.e. radiated and picked up by the aerial or conducted by cables entering the radio cabin. BS 1597:1985 deals with remedies generally considered necessary for the electrical installation and prescribes the permissible level of interfering emissions in the various wavebands. Precautions to be taken in the construction of the rigging and in the installation of cables are included.

This subject is detailed in an Appendix to the IEE Regulations for the Electrical and Electronic Equipment of Ships 1990.

46.18 Deck auxiliaries

Deck auxiliaries include cargo winches, cranes, capstans, warping winches, windlasses and hatch cover winches. For some classes of ship a variable speed for deck auxiliaries is still preferred, but, in general, two- or three-speed change-pole induction motors are suitable. For cranes a high-speed facility is needed for rapid return of the empty hook.

46.18.1 Variable speed

Where variable speed is essential in a.c. systems, slip-ring motors or Ward–Leonard control may be provided, the latter with a motor generator for each winch or for a group. In recent years there has been a trend to employ a.c. variable-speed drives. With d.c. systems, a variety of methods has been used: e.g. series resistance, Ward–Leonard, boost and buck, and lowering by dynamic braking or regeneration. Dynamic braking is generally achieved by means of diverters in shunt with the armature giving a potentiometer connection. In nearly all cases load discriminators may be necessary to control speed in relation to load, i.e. to permit of higher speeds with empty or lightly loaded hooks.

46.18.2 Deck auxiliary services

Some of the additional factors concerned with particular applications are given below.

46.18.2.1 Cargo winches

Emergency centrifugal brakes are fitted, where necessary to prevent excessive speeds when heavy loads are being lowered. Provision must also be made to prevent the load from running back in the event of a power failure. Light-hook speeds are 3–4½ times normal full-load speed.

Serious generator loading problems may be introduced where several winch motors with direct-on-line starting are in use. Large currents at low power factor cause generator and cable heating. *Table 46.1* gives empirical data for the effective current (in per-unit of the full-load current per winch) of a group of n winches ($n > 2$). Load-assessment curves are given in the IEE Regulations for the Electrical and Electronic Equipment of Ships.

46.18.2.2 Warping

Warping is frequently done with an additional barrel on care winches or windlasses. A flange prevents the hawser from running over the rim, and at the inner end of the frame a projection prevents it from becoming jammed between frame and warp end. If a foot-brake is fitted, its effect should be limited to the full-load torque of the winch, or torque-limiting relays should be fitted.

46.18.2.3 Capstans

Capstan barrels are normally mounted on a vertical shaft and the motor mounted below deck to leave the deck free.

46.18.2.4 Windlasses

Anchor windlasses are vital to the safety of the ship and have no stand-by. They are subject to classification and governmental requirements. The cable lifter is shaped to fit the links of the cable, and will normally accommodate four or five links around its circumference, although only two links are actually engaged at one time. The lifter can be declutched so that it runs freely for lowering, the speed being controlled electrically or by band brake. A slipping clutch is fitted to prevent excessive stress which could otherwise occur when heaving or when entering the anchor into the hawsepipe. A crawl speed is necessary to enable the anchor to be housed safely and to allow the motor to stall when it is fully home. It is also necessary while the anchor is still holding.

The overall efficiency of a windlass is about 60% and as much as 30% can be lost in friction unless, in accordance with modern practice, rollers are fitted. Windlasses can perform warping duties and extra control refinements are sometimes incorporated for these. They provide for a pull of about one-third of that of the cable lifters but at a greater speed. This speed is also required for recovering lines cast off from the quayside.

46.18.2.5 Mooring winches

Mooring winches are similar to warping winches except that one end of the line is fixed to the barrel.

The St Lawrence Seaway regulations require short-period stalling while docks are navigated, also constant tensioning against rising and falling tides or lock waters and during rapid loading or unloading.

46.19 Remote and automatic control systems

Since 1960 automation of ship's machinery spaces has increased considerably to the point where it is now generally recognised as an indispensable part of modern propulsion plant. In 1981 about 55% of the ships classed with Lloyd's Register of Shipping were designed to be operated with an unattended machinery space (UMS), and approaching 80% of the ships, including those with a UMS notation, had machinery spaces operated from a centralised control station.

Automation has also become indispensable in such areas as cargo handling, disposal by incineration of obnoxious waste, and pollution prevention. With high discharge rates, critical incinerator temperatures and specified limits of contamination, control systems are required to anticipate demand and give rapid response if catastrophes are to be avoided on board or at the terminal.

The proper functioning of control systems demands careful planning, since their viability may be governed by the nature of the ship's trading, the type of machinery to be controlled and the shipowner's manning and maintenance policies. A full specification is required of the extent of control facilities to be provided to ensure compatibility of controls with machinery, the marine environment and the operating personnel.

46.19.1 Operational modes of machinery spaces

Modern applications of control engineering systems permit two basic operational modes of the machinery space: (1) continuously attended, but with a high degree of remote and automatic controls to allow operation from a centralised control station; (2) periodically unattended machinery spaces so that engineers need not be tied to traditional watch-keeping routines and, for example, may leave the machinery space unattended during the night.

On vessels where centralised control is adopted, it is usual to incorporate all controls, alarms and instrumentation in an enclosed control room which is sound and vibration proofed and air conditioned. With the withdrawal of the engineer from the machinery space, it is essential that the controls and instrumentation provided be such that supervision of the machinery plant from the control room is as effective as it would be under direct supervision. The arrangements must give provision for corrective actions to

Table 46.1 Effective group current (in p.u. of the full-load current of n similar winch motors in a group, for $n > 2$) of winches

Part of system	*D.c. motors (series resistance)*	*A.c. cage motors*	*A.c. slip-ring motors*	*A.c. Ward–Leonard*
Cables and switchgear	$0.33n$	$3.3 + 0.3n$	$1.6 + 0.2n$	$1.2 + 0.15n$
Generators	$2.2 + 0.2n$	$5.0 + 0.4n$	$2.0 + 0.25n$	$2.2 + 0.2n$

be taken at the control station in the event of faults such as stopping of machinery, starting of stand-by machinery, adjustment of operating parameters, etc. These actions may be effected by either remote or automatic control.

With the advent of microprocessor-based control systems the concept of centralised control has been extended to incorporate techniques known as 'totally distributed control'. This enables all machinery within specified areas to be controlled and monitored by one integrated system. To implement this, individual microprocessor-based interface units are placed at various locations throughout the machinery space and are interconnected to the central control station by a data link.

Figure 46.9 shows diagrammatically a totally distributed control system. The outstations, interface units, have two functions: (1) to receive data from the various sensors, and (2) to output information to the control actuators. The system is arranged so that information is transmitted to the central station only as and when requested, or when a fault condition develops. If the data link between the central station and the outstation is broken, the outstation is capable of continuing operation.

In order to transmit information from the central station to the outstations, and vice versa, a multiplex system has to be used. This arrangement eliminates the need for interconnecting individual signals from each outstation to the central station. Multiplexing is a technique whereby each individual signal is given a specific address which can be transferred from one station to another along the cable in a very short time. The address is recognised only by the unit it is intended for, i.e. it will search many 'go-no go' gates until it finds the one 'go'. The information will then be used by the unit for action. *Figure 46.10* shows diagrammatically how multiplexing functions.

At the control station the complete operation of the system is organised; it consists of the central computers along with visual display units, keyboards, printers, analogue recorders, etc. In early distributed-control systems conventional analogue controllers were used at the outstations with the facility for the set points to be changed from the central station. However, as microprocessor systems have become more reliable, analogue controllers have been replaced by software generated digital control algorithms. Thus, the desired requirement for each control loop is retained in software rather than hardware. These systems are called direct digital control (DDC) systems. All the facilities that were available with the analogue controllers are built into these DDC systems.

46.19.2 Alarms and safeguards

For periodical unattended operation of a machinery space, all controls and safeguards necessary for centralised control are required, but safety actions must be automatic. In addition, it is necessary to extend the machinery space alarm system to the bridge and accommodation areas so that engineering personnel are made aware when a fault occurs. Control of propulsion must also be extended to the bridge to enable the navigating officers to carry out manoeuvres if the need arises. It is important, when an engineer responds to an alarm and enters the machinery space alone, that other personnel are aware of his well-being. It is usual to configure the alarm system so that the navigating officer is also made aware of a machinery fault, when it is being attended to and when it is corrected. *Figure 46.11* shows the functioning of an alarm system to meet these requirements.

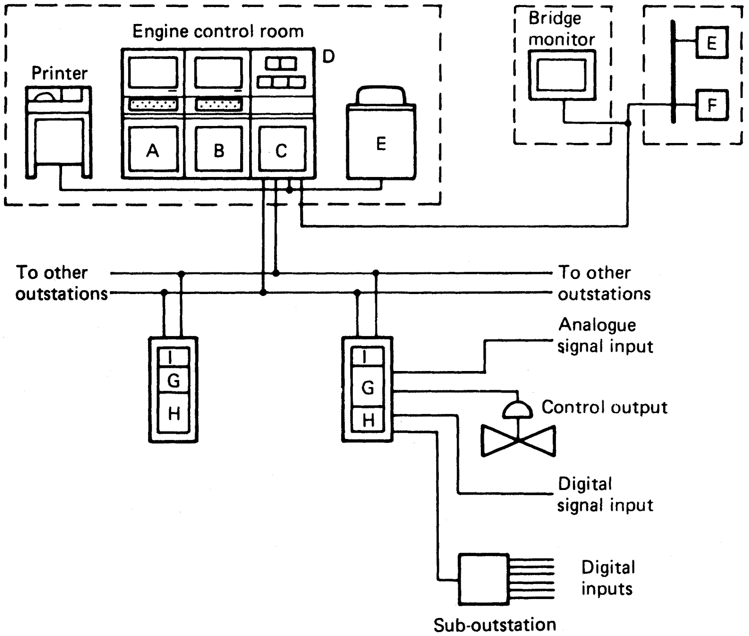


Figure 46.9 A totally distributed control system. A, Operator station alarm; B, operator station control; C, multiplexer and control unit; D, trend recorders; E, magnetic storage unit and general-purpose computer; F, extension alarm system; G, process interface analogue/control and alarms; H, process interface digital/alarms; I, multiplexer unit

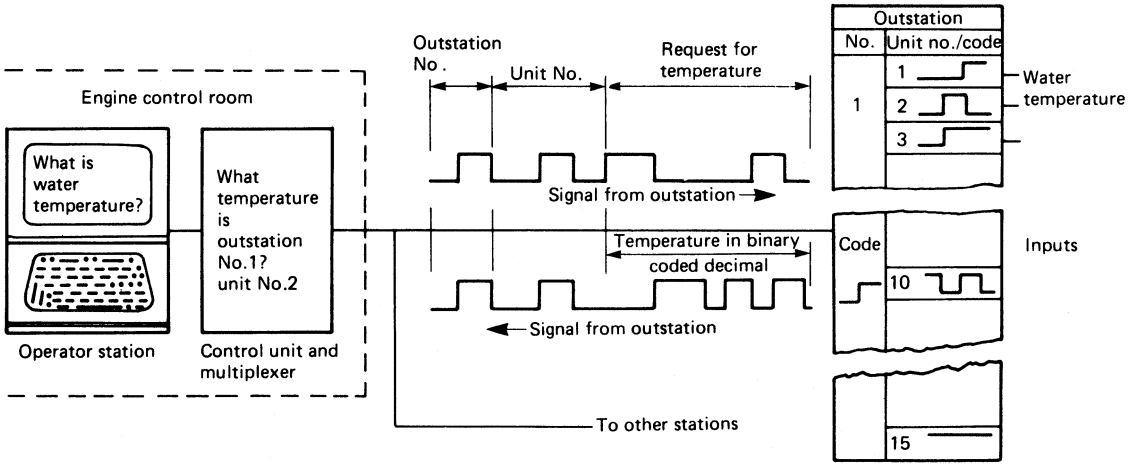

Engine control room
What is water temperature?
Operator station
What temperature is outstation No.1? unit No.2
Control unit and multiplexer
Outstation No.
Unit No.
Request for temperature
Signal from outstation
Temperature in binary coded decimal
Signal from outstation
To other stations
Outstation
No.
Unit no./code
1
1
2
3
Water temperature
Code
10
15
Inputs

Figure 46.10 A totally distributed control system—form of signal on the data link

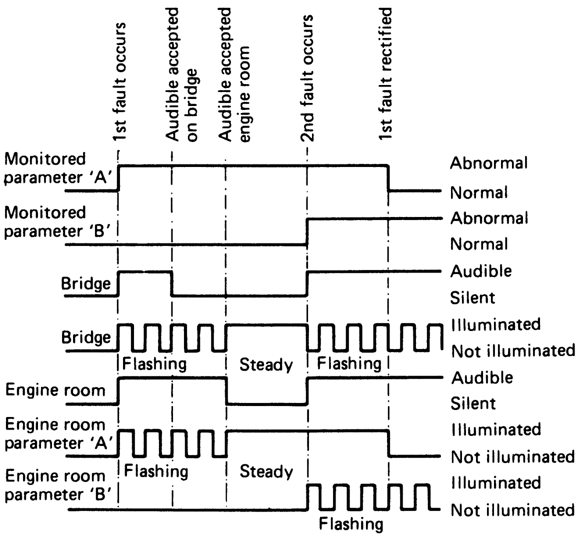

1st fault occurs
Audible accepted on bridge
Audible accepted engine room
2nd fault occurs
1st fault rectified
Monitored parameter 'A'
Monitored parameter 'B'
Bridge
Bridge
Flashing
Steady
Flashing
Engine room
Engine room parameter 'A'
Flashing
Steady
Engine room parameter 'B'
Flashing
Abnormal
Normal
Abnormal
Normal
Audible
Silent
Illuminated
Not illuminated
Audible
Silent
Illuminated
Not illuminated
Illuminated
Not illuminated

Figure 46.11 Functional diagram for an alarm system

46.19.3 Reliability

No matter how comprehensive the control and protection systems provided, they are of little value if the equipment used is not reliable. Experience has shown that many items of control equipment do not operate satisfactorily at sea. Control equipment, designed for use ashore, is all too often unsuitable or unreliable when used on board ship.

In order to improve reliability, the majority of classification societies have adopted and developed criteria to which equipment should be tested. This is known as 'type approval', a procedure whereby a prototype or a production unit is tested under conditions of environmental and mechanical stress that simulate severe shipboard operating conditions.

Basic minimum requirements specified by Lloyd's Register of Shipping include the following.

(1) *Visual inspection*: to ensure that workmanship is good and materials used are adequate for the duty of the equipment.
(2) *Performance tests*: to ensure that the manufacturer's specified limits of accuracy, repeatability, etc., are fulfilled.
(3) *Fluctuations in power supply*: voltage variations, steady, ±10% with simultaneous frequency variation of ±5%. Transient, ±20% voltage with ±10% frequency. For hydraulic and pneumatic systems supply pressure variations of ±20%.
(4) *Vibration tests*: testing in frequency range 1–13.2 Hz at ±1.0 mm displacement and 13.2–100 Hz at ±0.7 g. Endurance tests carried out at each major resonant frequency for 2 h.
(5) *Humidity*: 90–100% for 12 h at 55 ±2°C, then reduced to 20 ±5°C over a period of 1–3 h and remaining at the lower temperature for not less than 6 h. This test repeated over two full cycles.
(6) *Dry heat*: at manufacturer's stated maximum operating temperature if greater than 55°C for 16 h.
(7) *Inclination test*: at least $22\frac{1}{2}$° each side of the vertical in one plane, repeated in a second plane at right angles to the first.
(8) In addition, for electrical equipment, high-voltage and insulation resistance tests are required.
(9) *General*: further requirements may include low-temperature and salt mist tests. As applicable, intrinsic safety certification may be called for.

46.20 Tankers

It is of vital importance to take account of bending stresses in the hull resulting from unequal buoyancy. These may arise from ballasting or from different grades of oil, or may occur during loading or discharging. A large number of valves control the filling and unloading of cargo tanks, and these must be operated in logical sequences. Trim and list must also be taken into account. There is obviously a fertile field for computer controlled centralised operation. Because of explosion risks, hydraulic power is favoured for valve operation. For instrumentation and control purposes, pneumatic and intrinsically safe electronic circuits are suitable, subject to classification approval.

46.21 Steam plant

Correct relationship between fuel supply, feed-water supply, temperatures, forced draught, engine load and sea-water temperatures depends on a large number of interdependent controls. Optimum efficiency is rarely, if ever, achieved without some form of automatic control. The essential factors are: (a) maintenance of steam pressure under varying loads by fuel control; and (b) optimum control of combustion air flow and fuel/air ratio.

46.22 Generators

Two or more generators are always provided, but for economy it is undesirable to run more sets than necessary for the prevailing load. Preferential tripping of non-essential loads has already been dealt with. This is only a temporary expedient and, to prevent overloading, systems are available for automatic starting, stopping and synchronising of sets according to demand. In confined waters it is necessary for safety reasons to have a margin in reserve, and it is therefore usual practice to have two sets on the board. An overriding provision should be included in automatic schemes.

Advances in technology have led in recent years to the introduction of micro-based systems for efficient power management. Escalating fuel costs and the need for reliable unmanned operation have stimulated this development.

46.23 Diesel engines

When bridge control of main propulsion engines is installed, it is necessary to make provision for a repeat start if the first sequence is not completed. After a set number of false starts (usually three) an alarm operates. If acceleration to the firing speed is not reached in about 3 s, a further period of about 4 s is allowed and the complete cycle is then repeated.

For auxiliary generator plant, standby sets can be prewarmed by continuous circulation of hot water from the main engine cooling system. It is usual to bring in another set before full load is reached. However, a short time delay is advisable to prevent starting on a spurious load demand.

46.24 Electric propulsion

The popularity of propelling ships using electric motors has varied considerably over the past 80 years, but electrical propulsion has always offered several advantages over traditional mechanical systems. These include, installation flexibility, high torque across the whole speed range, low vibration and noise, smooth variation of speed and reduced fuel consumption and maintenance (not always so in d.c. systems). It should be noted, however, that both the initial cost and size of an electrical system are generally greater than those of mechanical systems.

46.24.1 Methods of propulsion

The mechanical power produced by the electric motor may be transmitted to propel a vessel in a number of different ways:

(1) by direct coupling, or via a reduction gear box, to a conventional fixed pitch propeller (FPP) or to a controllable pitch propeller (CPP);
(2) by hydrojet thrusters;
(3) by an azimuthing or fixed ducted propeller assembly; or
(4) by a cycloidal propeller system.

The electric motor may be the sole means of propulsion, or it may be used in conjunction with a mechanical drive, as in some types of naval vessel. In these types of ship electric motors are used at low speed, often when the vessel is required to minimise radiated noise for sonar activity, and at high speed gas turbines are used to provide the boost power as required.

Steering of a vessel may also be achieved by the propulsion unit, as in the case of the azimuthing propeller assembly, but such systems must conform to specifications as laid down by Lloyds Register of Shipping's Rules and Regulations for the Classification of Ships.

46.24.2 Traditional electrical systems

46.24.2.1 D.c. generators, d.c. motors, fixed pitch propellers

In series-connected d.c. systems the generators and motors are all connected in series in one circuit. In the parallel-connected system all generators and motors are arranged in parallel. In both cases the speed and direction of rotation of the motors is changed by varying the generator and motor excitations, as shown in *Figure 46.12.*

Both parallel and series-type systems have been used widely in the past, particularly in ice breakers, but the build up of carbon dust from the commutator brush gear has always been a problem with d.c. systems. As the size of vessels has progressively increased, the popularity of this method of propulsion has declined, and there are only a few vessels still in service employing full d.c. propulsion systems.

46.24.2.2 A.c. generators, a.c. motors, variable-frequency supply

An a.c. generator, often driven by a steam turbine, supplies a synchronous motor. Speed control of the motor is achieved by varying the speed of the prime mover and hence the generator and motor frequency, as shown in *Figure 46.13*. Below 25% of normal speed, i.e. when manoeuvring is being carried out, it is usual to operate the propulsion motor as an induction machine, i.e. with the motor field removed. Precautions need to be taken during this manoeuvring period because high slip-ring voltages can be induced.

During the past two decades there has been a steady decline in the number of turbo-electric systems in operation, mainly because prime-mover limitations make it difficult to achieve high torques at low propeller speeds. The most notable ship still employing such a main propulsion system is the passenger ship *Canberra*, but low-power turbo-electric systems have seen some resurgence for bow thruster applications.

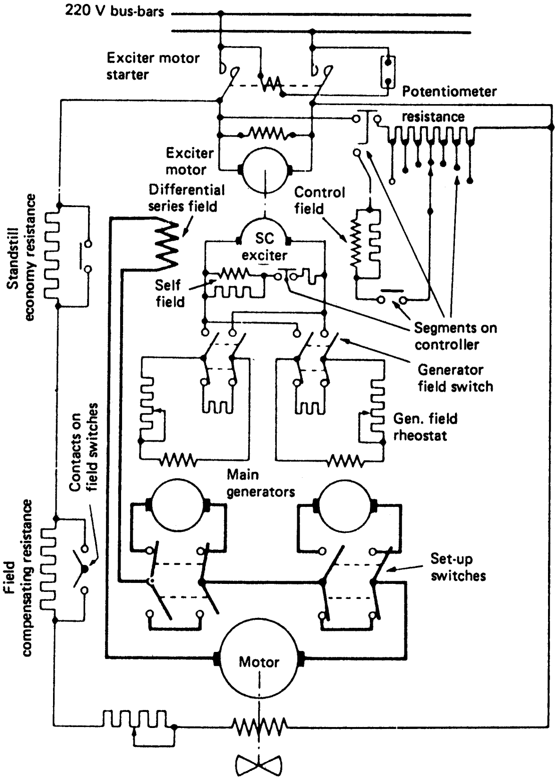

220 V bus-bars
Exciter motor starter
Potentiometer
resistance
Exciter motor
Differential series field
Control field
SC exciter
Self field
Segments on controller
Generator field switch
Gen. field rheostat
Main generators
Set-up switches
Motor
Standstill economy resistance
Contacts on field switches
Field compensating resistance

Figure 46.12 Diesel-electric d.c. system with modified Ward–Leonard control

46.24.3 Modern electrical systems

46.24.3.1 A.c. generators, d.c. motors, fixed-frequency generation and fixed pitch propeller (FPP)

The d.c. propulsion motor is normally supplied by a fully controlled bridge rectifier fed from the a.c. generator supply system, as shown in *Figure 46.14*. The speed of the motor is controlled by changing the firing angle of the silicon controlled rectifiers (thyristors), which in turn change the voltage supplied to the armature of the d.c. motor. Motor reversal is easily achieved by reversing the polarity of the field of the motor.

There is, however, an empirical limit to the size of d.c. motor that can be built, which depends on the product of speed and power:

$$1.5 \times 10^6 = \text{Power (kW)} \times \text{Speed (rev/min)} \Leftarrow$$

As an example, this limits the power to 10 MW at 40 rev/min. In vessels requiring propulsion powers greater than those obtainable with a single d.c. motor, the choice is really between multiple d.c. motors on the same shaft or a full a.c. system.

46.24.3.2 A.c. generators, fixed-speed a.c. motors, fixed-frequency generation

In fixed-speed systems a constant-speed a.c. motor is used in conjunction with a controllable pitch propeller, as shown in *Figure 46.15* for a North Sea support vessel. The motor may drive the propeller directly, such as in the system employed in the New Zealand ferry *Arahura*, or via a gearbox, an example of this being the *BP Iolair*. Both synchronous and induction motors are used in these types

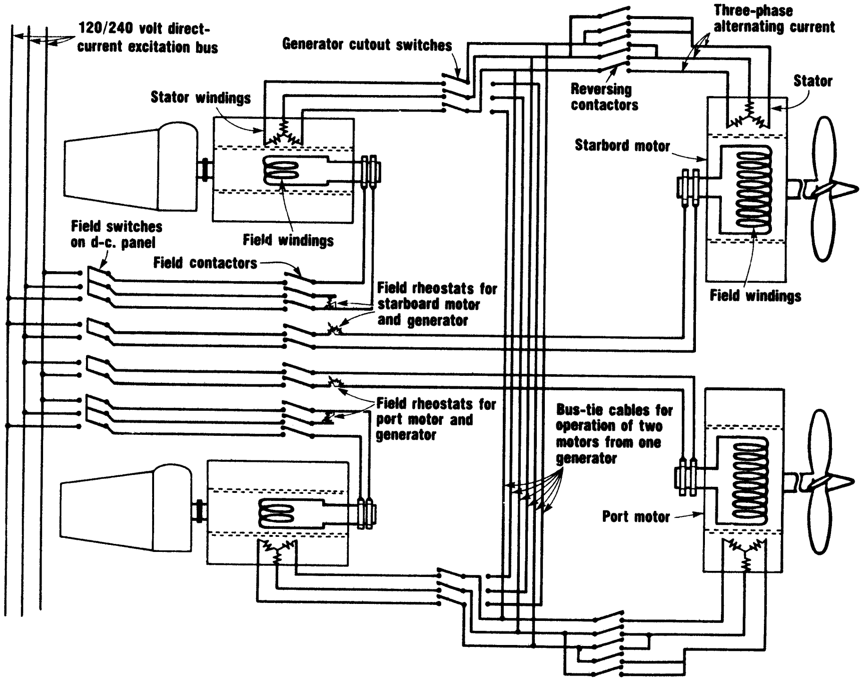

120/240 volt direct-current excitation bus
Generator cutout switches
Three-phase alternating current
Stator windings
Stator
Reversing contactors
Starbord motor
Field switches on d-c. panel
Field windings
Field contactors
Field rheostats for starboard motor and generator
Field windings
Field rheostats for port motor and generator
Bus-tie cables for operation of two motors from one generator
Port motor

Figure 46.13 Electrical propulsion system with variable frequency generation

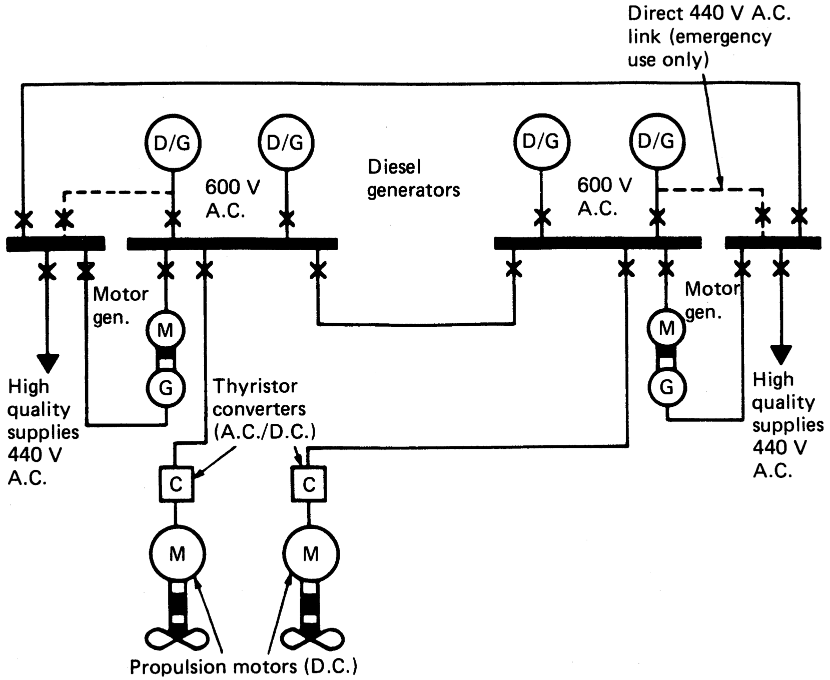

Direct 440 V A.C. link (emergency use only)
D/G
D/G
600 V A.C.
Diesel generators
D/G
D/G
600 V A.C.
Motor gen.
M
G
Motor gen.
M
G
High quality supplies 440 V A.C.
High quality supplies 440 V A.C.
Thyristor converters (A.C./D.C.)
C
C
M
M
Propulsion motors (D.C.)

Figure 46.14 A.c. generators/d.c. motor configuration employing motor generator sets to ensure that harmonics generated by the fully controlled converter are not transmitted on to the 440 V a.c. bus-bars

of electric propulsion system, the motors being fed directly from the 50 or 60 Hz generator system or via transformers. Speed control of these vessels is achieved solely by the use of the controllable pitch propellor (CPP), so this type of system is employed in vessels that are operated extensively at high speed or in vessels that require dynamic positioning.

46.24.3.3 A.c. generators, variable-speed motors, fixed-frequency generation

In this type of system the speed of the a.c. motor is controlled by a frequency converter, and is generally used in conjunction with a fixed-pitch propeller, as shown for an ice-breaker in *Figure 46.16*. Shaft-speed variation is achieved by varying the supply frequency to the propulsion motor, since the motor speed is proportional to frequency/number of poles:

$$\text{Rev/min} = \frac{120 \times \text{Frequency (Hz)}}{\text{No. of poles}}$$

A synchronous motor is normally used in preference to an induction motor in many applications. This is because synchronous motors can be manufactured with larger air gaps to withstand arduous mechanical conditions. They can also be operated at unity power factor across the whole speed range, and generally have higher overall efficiencies. In addition, the synchronous motor has the highest power/weight ratio of any electrical machine. Induction motors, because of their reduced construction costs, tend to be used in low-power propulsion systems only.

There are two main types of static converter used to feed the synchronous machines with the variable frequencies they require, the cycloconverter and the autosynchronous inverter. Both are reversible and, if necessary, both systems can be designed to provide motor braking provided dynamic resistors are included.

Induction motors are generally fed using autocommutated or forced-commutated inverters, although cycloconverters have recently found favour in systems requiring power above 8 MW.

This type of electric drive has become the accepted system in many different types of vessel, including cruise liners and ice-breakers. The *Queen Elizabeth 2* employs a modified version of this system, and includes CPPs for manoeuvring at lower speeds, and an autosynchronous drive that is used for speed adjustment at higher speeds.

46.24.4 Voltage levels and harmonics

The choice of system voltage for an electric propulsion system is determined predominantly by the size of the propulsion plant and the short-circuit fault currents. Where the capacity of individual generating sets exceeds 2.5 MW, or the fault level under normal operation exceeds 50 MV-A, a high-voltage system is usually selected as specified in Part 6, Chapter 2.1 of the Rules and Regulations of the Classification of Ships. Standard marine high voltages now in use are:

50 Hz	*60 Hz*
3300 V	4160 V
6600 V	6000 V
	10 000 V

IEC 92:Part 304 also applies to the installation and performance of static converters using semiconductor

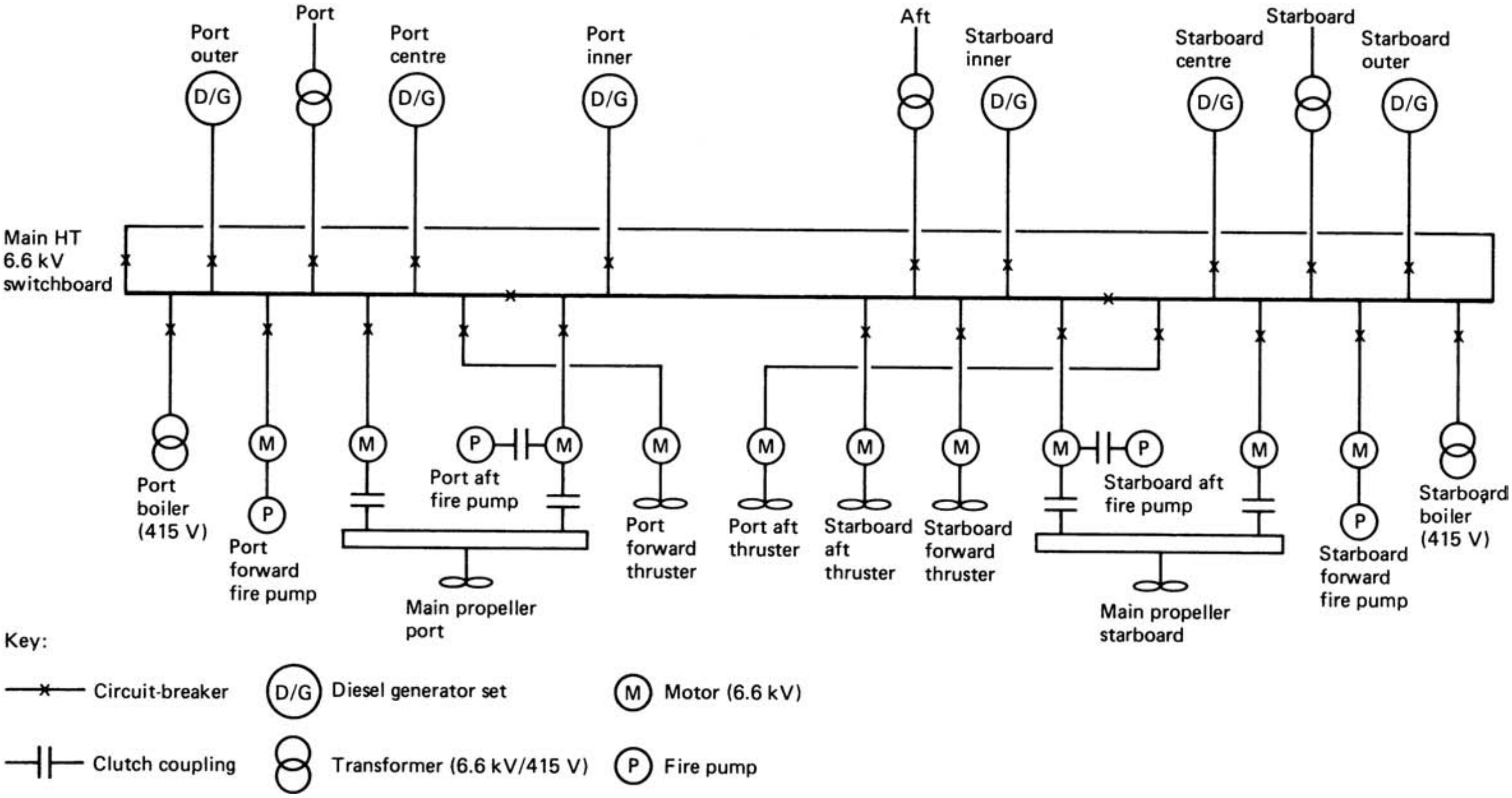

Port outer
Port
Port centre
Port inner
Aft
Starboard inner
Starboard centre
Starboard
Starboard outer
D/G
Main HT 6.6 kV switchboard
M
P
Port boiler (415 V)
Port forward fire pump
Port aft fire pump
Main propeller port
Port forward thruster
Port aft thruster
Starboard aft thruster
Starboard forward thruster
Starboard aft fire pump
Main propeller starboard
Starboard forward fire pump
Starboard boiler (415 V)
Key:
Circuit-breaker
Diesel generator set
Motor (6.6 kV)
Clutch coupling
Transformer (6.6 kV/415 V)
Fire pump

Figure 46.15 A.c. generators/a.c. motor configuration employed on a typical North Sea oil-field support vessel. Speed variation is achieved solely by the use of controllable-pitch propellers

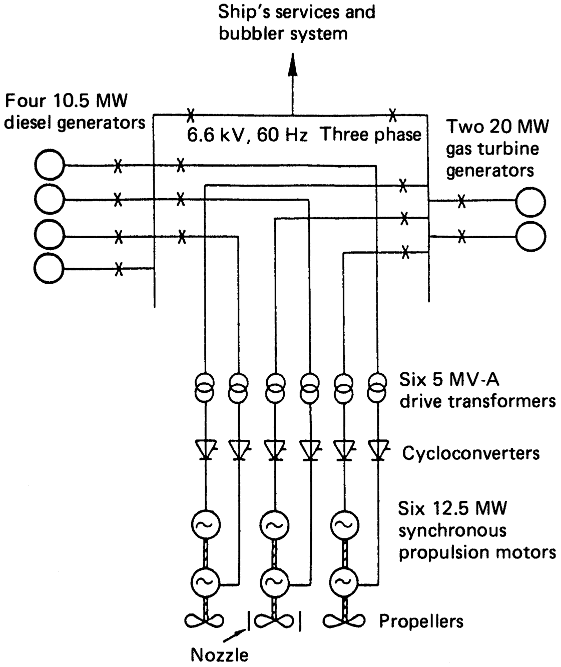

Ship's services and bubbler system
Four 10.5 MW diesel generators
6.6 kV, 60 Hz Three phase
Two 20 MW gas turbine generators
Six 5 MV-A drive transformers
Cycloconverters
Six 12.5 MW synchronous propulsion motors
Propellers
Nozzle

Figure 46.16 A.c. generators/a.c. motors employing cycloconverter variable-speed drives commonly found on ice-breakers

devices. In ships using converters, harmonic problems, such as electrical resonance, may occur in the electrical system. In some instances the problems can be quite severe, and the high power levels of the converters usually makes in-line filtering impractical. The recommendations for eliminating harmonic problems using motor generator sets, or low-power filters local to the equipment to be protected, is governed by the IEE Regulations for the Electrical and Electronic Equipment of Ships with Recommended Practice for their Implementation 1990.

46.24.5 Electric propulsion employing superconductivity

The development of niobium–titanium d.c. homopolar motors in the UK and USA during the 1960s demonstrated that high shaft torques could be produced with almost no motor power loss if the motor could be cooled using liquid helium. An electromagnetic thruster system has been developed in Japan. In this system superconducting coils are used in a linear motor configuration. An electric current is passed through sea-water and a propulsion thrust is generated without the use of rotating propellers. The disadvantage of superconducting systems is that large cooling plants are required to maintain the motors at the very low temperatures needed (−268.6°C). Recent research has been directed towards obtaining super-conductivity at ambient temperatures using ceramics rather than alloys but, although significant advances have been made, it is unlikely that super-conductivity will become a viable commercial option in the near future.

46.24.6 Electromagnetic slip couplings

In diesel-engined ships in which the propeller is driven through mechanical reduction gear, up to four engines per shaft can be coupled to the gear through slip couplings. It is necessary to protect the gears from torsional vibration transmitted by the engines, and the slip coupling serves both this purpose and that of a disconnecting clutch, enabling the numbers of engines in service to be altered without stopping the engines already operating. A typical four-engine arrangement is shown in *Figure 46.17.*

Manoeuvring can be carried out by having some engines running ahead and the others astern and selecting the direction required by switching.

A coupling can exert a starting torque from rest, with the engine at full speed, equal to full-load torque. The efficiency is high, the only losses being windage, excitation and the I^2R loss due to slip. Slip varies with speed and rating and is generally 1–2%.

46.24.7 Electromagnetic gearing

Elements of a speed reduction and reverse gear which can also act as an auxiliary a.c. generator, clutch and flexible coupling are shown in *Figure 46.18.* An inner rotor carries the input shaft, field windings and slip-rings. The outer rotor carries the electromagnetic coupling armature winding on the inner side which is connected, through a switching device, to the synchronous motor field windings on the outside, and two sets of slip-rings.

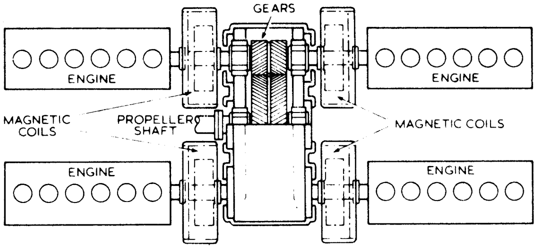

GEARS
ENGINE
ENGINE
MAGNETIC COILS
PROPELLER SHAFT
MAGNETIC COILS
ENGINE
ENGINE

Figure 46.17 Typical four-engine coupling arrangement

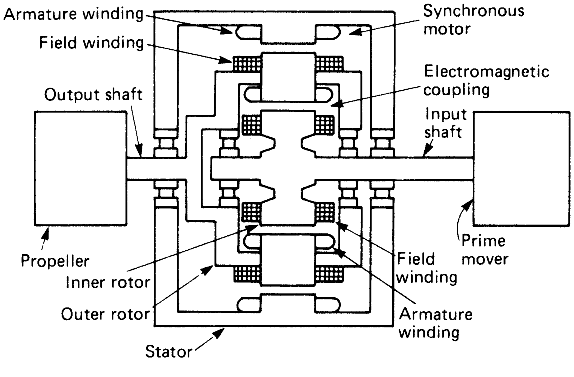

Armature winding
Field winding
Output shaft
Propeller
Inner rotor
Outer rotor
Stator
Synchronous motor
Electromagnetic coupling
Input shaft
Prime mover
Field winding
Armature winding

Figure 46.18 Versatile electromagnetic gear

47

Aircraft

M J Cronin FRAeS, FIEE, AIAA, IEEE (US)
Advanced Aeronautical Systems Company, USA

Contents

47.1 Introduction 47/3

47.2 Engine technology 47/3

47.3 Wing technology 47/4

47.4 Integrated active controls 47/6

47.5 Flight-control systems 47/6

47.6 Systems technology 47/7

47.7 Hydraulic systems 47/8

47.8 Air-frame mounted accessory drives 47/9

47.9 Electrohydraulic flight controls 47/11

47.10 Electromechanical flight controls 47/12

47.11 Aircraft electric power 47/12

47.12 Summary of power systems 47/13

47.13 Environmental control system 47/14

47.14 Digital power/digital load management 47/17

47.1 Introduction

Modern aircraft have come a long way since the Wright brothers built their first aircraft in their bicycle shop in Dayton, Ohio, and made the first powered flight of an airplane in 1903, at Kitty Hawk, North Carolina. This first crude biplane, however, embodied the two basic elements of all modern airplanes; namely a propulsion system and lifting devices (wings). Since that time, the technologies of propulsion and wings have made significant strides in the science of aviation.

Progress in technology, as in medicine, is stimulated significantly during wartime and World Wars I and II played a major role in advancing the basic design concepts of the modern (military and civil) airplane, as well as bringing about major improvements in the power plants. Prior to World War II, aircraft were powered initially by internal combustion engines in the 1600–2800 hp range, but during that war the first jet-powered aircraft emerged.

47.2 Engine technology

From the simple rudimentary two-stroke gasoline piston engines developing a mere 60 hp, engines moved from the small 150 hp piston engines that powered the World War I biplanes (such as the Bristol Bulldog, Sopwith Pup, Avro 504, the French SPAD and the German Fokker aircraft) to the 1600 hp/1800 hp 'radial' and 'in line' piston engines that were used in the four-engined bombers in World War II. The radial engines were designed for propeller/ram-air cooling: the air stream from the former being used to provide blast-air cooling of the 'finned' piston housings (during static conditions on the ground) and ram-air cooling, during taxi and flight conditions. The 'in-line' engines, in contrast, were liquid cooled, as exemplified by the Rolls-Royce Merlin engine that powered the early Spitfire at some 340 mile/h in its record-breaking test run at a Schneider Cup Race in 1927. The advantages of the 'in-line' engine in this case were its low frontal area, which permitted power-plant streamlining and, therefore, reduced power-plant drag.

From these early World War II engines, the designers explored other piston-engine technology such as Bristol's 'sleeve-valve' engine, named the Centaurus, that had a take-off power rating of approximately 2800 hp. This engine was unique in that it eliminated the complex 'poppet-valves', with their push rods, cams, etc. In the Centaurus engine, the equivalent of the conventional inlet/exhaust valves was provided by an inner 'sleeve' cyclinder, within the piston housing, that had machined ports (openings) that successively exposed the inlet and exhaust manifolds to the combustion chambers; as the sleeve moved up and down with an oscillatory motion. Napier also developed a double '12-pot' engine arranged with the rear 12-pot engine 'staggered' (for cooling purposes) behind the first: this 24-cylinder engine was rated at 2850 hp. Another novel engine development by Napier was the Nomad engine which uniquely combined a turbine engine with a diesel engine, to yield attractive fuel and weight savings. The following is Napier's weight estimate of typical four-engined aircraft, carrying the same payload over a range of 2860 miles.

4 × turbojets (10 000 lbf each) 153 000 lb total
4 × turboprops (3780 shp each) 140 000 lb total
4 × compounds (3780 shp each) 123 500 lb total

(where lbf is pounds-force and shp is shaft horsepower).

Higher horsepower engines were developed in the 1940s and were used during World War II. The most notable US fighter at that time was the Curtiss 1-40 Warhawk which used an Allison 12-cylinder liquid-cooled in-line engine that developed 1200 hp. This period was also of note in that engine (propulsor) technology moved on from the internal-combustion-engine design to the turbojet and turboprop designs. Frank Whittle's gas turbine engine, built and tested in 1930, was the precursor of these jet engines which were to set the trend for these advanced propulsors in the military and commercial aircraft. The first early military aircraft application of the jet engine in World War II was in Germany's Messerschmitt Me262A which had a maximum speed of 540 mile/h. This jet-engine-powered aircraft was followed by the UK's Gloster-Meteor aeroplane with two Rolls-Royce Derwent engines, developing 2000 lbf each: these two engines provided the Meteor with a top speed of 520 mile/h.

Early turbine engines were rated in the 2000–5000 lbf range, but this was soon to increase to 10 000 and 20 000 lbf. The Pratt & Whitney PW7000 is an example of a modern 20 000 lbf engine.

During the 1950s, engine thrusts increased to 48 000 lbf as in the case of the Rolls-Royce RB.211, the GE-CF6 and United Technologies P&W JT-9D engines. These engines powered the Lockheed L-1011, Douglas DC-10, Boeing B-747 and many other aircraft. These engines foreshadowed the emergence of even larger engines, in the 80 000–100 000 lbf range, which will power the Boeing B-777X and other new aircraft.

During these years, the engines were defined as 'low-pass' and 'high-pass' engines. In the latter type an increasing amount of the ram-air was bypassed around the 'engine core' and then mixed with the core-flow air in the 'nozzle': such engines are known as 'turbofans'. To pump the bypass air, a large 'fan' was mounted in the front of the engine and this was powered by the engine's low-pressure turbine, which furnished the necessary shaft power for the fan.

The thrust power developed by these engines was derived by accelerating the air-mass flow through the engine to effect a significant change in air energy momentum (lbf ≡ MV). Fuel economy being a quintessential objective of all modern aeroplanes, engine technology was directed toward increasing engine compressor ratios, decreasing tolerances and improving the science of high-temperature metallurgy. These high-temperature materials were developed primarily for the hot turbine sections of the engine to permit higher turbine inlet temperatures. These advanced metallurgical developments led to the use of high temperature aluminium alloys, titanium alloys and Inconel. Ceramic-coated turbine blades were also used in the first and second turbine stages to permit turbine inlet temperatures in excess of 2500 °F. The following is the power relationship of typical gas turbine engines.

$$\mathrm{HP_t} = \Delta{}^\circ F \times \mathrm{pps}/2.95 \qquad (47.1)$$

where

$$\Delta{}^\circ F = \eta \times T{}^\circ R \times \left[1 - \frac{1}{(\mathrm{PR})}\right]^{[(k-1)/k]}$$

η is the turbine efficiency, k is the ratio of specific heats, $\mathrm{HP_t}$ is the turbine horse power, pps is the engine core flow in pounds per second, PR is the (turbine) pressure ratio, T is the temperature, R is the gas constant, Δ°F is the temperature drop across turbine, and T°R is the temperature in degrees Rankine (T °F + 460).

As a concomitant technology objective, turboshaft engines were developed in which nearly all the core energy of the engine was converted to shaft power, rather than thrust power. From the earlier 1600–2800 shp ratings of

World War II, large turboshaft/turboprop engines were designed, worldwide, in ratings up to 5500 ehp (equivalent horsepower) and higher. In these engines, the residual thrust power in the engine nozzle was added to the shaft power to provide a total 'equivalent' horsepower. These turboshaft engines found wide application in aircraft, marine and utility installations. The turboshaft engines were particularly suitable in aircraft where fuel economy was a key consideration. Typical of these aircraft were the antisubmarine warfare (a.s.w.) and the aircraft early warning (a.e.w.) aircraft, which were required to remain 'on station' for anywhere from 10 to 20 h. Another example of a 'widely used' turboshaft-powered aeroplane is the Lockheed C-130 (Hercules) which has a take-off gross weight of approximately 155 000 lb and is powered by four Avco/Lycoming T-56 engines rated at approximately 4500 eshp each: this aircraft has a payload capacity of approximately 44 000 lb.

Marked improvements in fuel efficiency (low specific fuel consumption) are obtained with turboprop engines compared to turbojet aircraft. In fact the engines are 'propfans' which use large highly curved, highly twisted blades that deliver superior performance, *vis à vis* the conventional turboprops.

Helicopters are also prime users of turboshaft engines as in the case of short take-off and landing (s.t.o.l.) and vertical take-off and landing (v.t.o.l.) aircraft. The US V-22 Osprey is an entirely novel application of a turboshaft v.t.o.l. airplane in which the turboprops are rotated vertically for take-off and translated back to the horizontal position during flight. The significant design advantage of this is that where the speed of a conventional helicopter is limited to a percentage of the propeller-tip speed to some 140–180 mile/h, the V-22 can operate as a conventional turboprop aeroplane, in straight and level flight, and achieve speeds of 300 mile/h and above.

Turboprop engines are exemplified by their fuel efficiency and by their fast response to changes in power settings, which derives from the fact that the propellers typically operate at constant speed and the thrust power is changed simply by a change of the propeller pitch. The manoeuverability and the ability of turboprop aircraft to execute a short take-off run and a rapid pull-up (after take-off) are, therefore, the main attributes of these aircraft. These turboprop aircraft are also very fuel efficient and this commends them to a.s.w. and a.e.w. roles.

In addition to the unique applications of the turboshaft/turboprop engines, a.s.w. aircraft such as the Lockheed S-3 (GE-TF-34 tf), Lockheed P-3 (4 × Lycoming T-56 tp) and a.e.w. aircraft such as the Grumman E-2C (2 × Lycoming T-56 tp), Boeing E-3A (4 × P&W TF33 tf) used turboprops and turbofans. These a.s.w./a.e.w. aircraft are the backbone and workhorse of the US patrol and surveillance activities.

For reconnaissance activities, the US used the McDonnell-Douglas RF-4C (2 × GE-79 tj), Northrup's RF-5E (2 × GE-85 tj), but *the most notable reconnaissance aircraft were the Lockheed SR-71 (2 × P&W J-58 tj) and the Lockheed U-2 aircraft (1 × P&W-75 tj)*.

To improve the fuel-efficiency of the *turbojet* engine, NASA sponsored and funded the development of the energy-efficient engine (e.e.e.). This engine is identified by a very high bypass ratio and a very small core section. For these fuel-efficient engines, the number of compressor stages was increased and the compressor ratio changed from the typical value of 16:1/20:1 and later 60:1. A caveat, implicit in the e.e.e., is that its low amount of core-flow air makes it *very sensitive to compressor bleed-air extraction*. As a result, its prospective use for powering passenger services such as the environmental control system (e.c.s.) and engine/wing de-icing is highly penalising to the engine's aerothermodynamic efficiency.

47.3 Wing technology

In the preceding section an overview was given of the significant progress in aircraft propulsion technology, as derived from the design and development of the internal combustion engine, the turbojet, turbofan, turboprop and the energy efficient engines. It is axiomatic, however, that unless the engine technologies were matched by comparable air frame/wing technologies, they would abrograte and offset the benefits of the new engines.

The two key elements highlighted for *aircraft efficiency are the 'engine' and 'wing' technologies*. Fortunately, these technologies have seen major advances in the last 30 or more years, resulting in outstanding military aircraft such as the World War II Spitfire, the Lockheed F-104G, the Lockheed P-38, the Messerschmitt Me262A, the USSR MIG-24, the French Mirage and many other high performance military aircraft built in the USA, Europe and other countries worldwide. Similar aerodynamic improvements were implemented in commercial aircraft such as the Boeing B-757/B-767, B-727, the Lockheed L-1011, Douglas DC-10, the BAC A-320/330, as well as many European and Russian/Chinese commercial aircraft.

In the following discussion, wing technology is reviewed from the standpoint of the different wing plan forms and their respective advantages *vis à vis* high lift/drag efficiency, etc. Simply, an aeroplane's operating efficiency η can be defined as

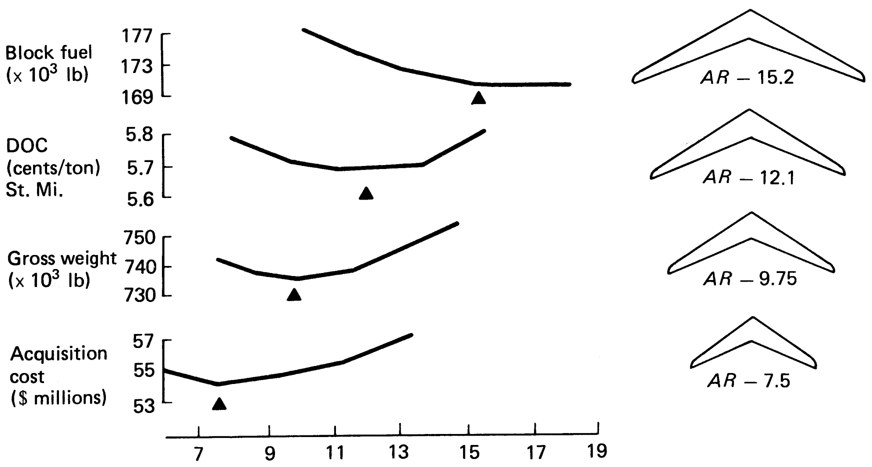
$$\eta = M \times L/D \tag{47.2}$$

where M is the mach number, L is the lift and D is the drag.

Figure 47.1 shows a number of 'optional' wing plan forms, which the aircraft designer evaluates from the standpoint of which parameter best fits his particular airplane. Where 'fuel efficiency', for example, is key (because it impacts on the direct operating costs), a wing with a *high* aspect ratio is desirable. (Note: the aspect ratio may be defined approximately as the relationship of the wing span to the average wing mean chord.)

When fuel efficiency is the important design option, an aspect ratio of 10:1 to 15:1 would be efficacious, recognising that these are heavier wings. However, if 'acquisition costs' are important, then (at the other end of the scale) an aspect ratio of 7.5 would be appropriate. Clearly, there is no single 'optimum' aspect ratio that meets all criteria, so the designer must make an acceptable compromise. Aircraft weight and wing weight are important to the overall performance of an aeroplane so the designer's objective is to achieve a maximum L/D ratio for a minimum wing weight. An integral part of this design activity centres on an evaluation of wing materials that yield the maximum strength-to-weight ratios. Other than this, composites (non-metallic materials) such as Kevlar (aramid and graphite), light-weight materials, such as new aluminium alloys, titanium, lithium alloys, etc., are candidates. Most composites have the advantage that double curvatures and unique aerofoil configurations can be fabricated, which are difficult to achieve with metallic structures.

The wing 'aerofoil' cross-section is another key aspect of wing performance and the wing camber must be such as to optimise the lift and suction forces across the wing. The important objective is to obtain a laminar airflow over as much of the wing as possible and to delay the break-up of airflow, which results in turbulence and drag. Addressing such wing designs, aerodynamicists have developed customised airfoil sections (such as used in 'supercritical' wings), whose distinctive aerofoil section keeps the airflow laminar, over much of the wing's forward cross-section. Complementary to this, the designers have evaluated and



DOC = direct operating costs
AR = aspect ratio

Figure 47.1 Design aspects of wing aspect ratio transport

developed 'natural' and 'forced' techniques of laminar flow control. In the former case, natural laminar flow is accomplished by an efficient wing-camber design, such as the 'supercritical wing'. With forced laminar flow control, a complex wing manufacturing design is involved in which slots or small openings are fabricated in the wings to permit the wing's boundary air to be sucked close to the wings, by means of 'suction air' derived from engine compressor bleed air.

Wing design is a key to fuel efficiency. *Figure 47.2* shows a comparison of an early wing design and an 'advanced' wing design; the latter is derived from computer modelling, wind-tunnel testing, etc. The curves show that the advanced wing design yields a 11–12% improvement in fuel consumption (at the 25% and 40% positions of the wing's mean aerodynamic chord (m.a.c.)). Furthermore, as the centre of gravity (c.g.) moves aft of these neutral points, another 1.5–3.0% improvement can be achieved. However, when the c.g. travels to 35% m.a.c., it reaches a point of relaxed static stability (r.s.s.) and, further back, to a point of 'negative' static stability, where the aircraft becomes longitudinally unstable. In this condition, the pilot cannot fly the aeroplane manually and must rely on a sophisticated (and highly reliable) stability augmentation system (s.a.s.). This type of system dictates the adoption of an electronic flight control system, known as a *digital flight control system* (d.f.c.s.). Apart from providing stable flight management (of a basically unstable aircraft), the primary function of the d.f.c.s. is to reduce the trim drag on the tail of the aeroplane. This trim drag is due to a down force on the tail, when the c.g. of the aeroplane is forward of the wing's m.a.c. and it results in a fuel penalty.

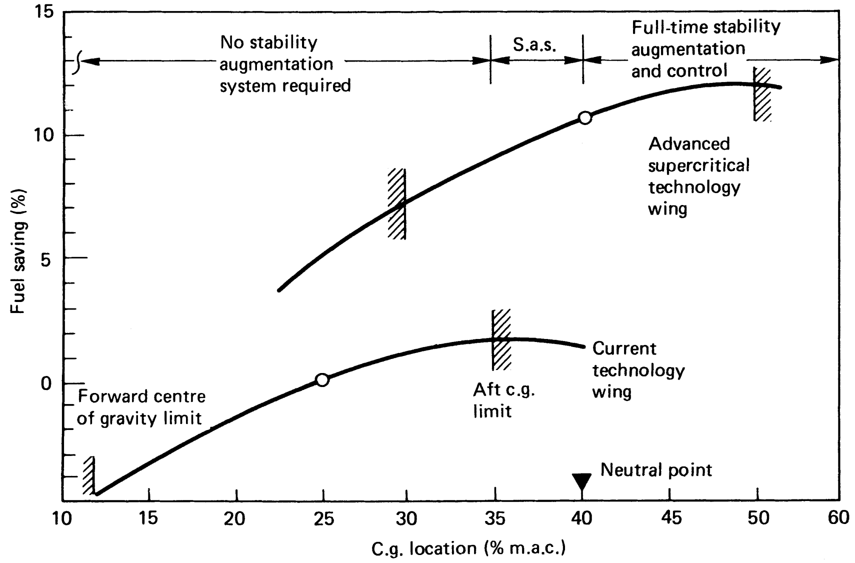


Figure 47.2 Lift/drag versus centre of gravity (c.g.) position. m.a.c., Mean aerodynamic chord

47.4 Integrated active controls

A fall-out 'benefit' from the adoption of a sophisticated d.f.c.s., is a reduction in wing weight which can be achieved by the use of integrated active controls (i.a.c.s). Active control technology (a.c.t.) involves a duplicative use of the ailerons, such that the ailerons can be operated symmetrically in addition to their normal (differential) mode of operation. The purpose of a.c.t. is to modify the spanwise wing-lift distribution from the wing root out to the wing tip. Normally, the spanwise wing-lift distribution is elliptical as shown in *Figure 47.3*, but when a.c.t. is used the wing-lift distribution is reshaped as indicated. This results in a reduction in the wing-root bending moment, to permit either a lower weight wing design or a wing design with a higher aspect ratio. Lockheed opted for the latter in the L-1011 aeroplane and added a 3.2 ft extension on each wing to achieve a higher aspect ratio wing and a more fuel-efficient aeroplane.

In summary, it is clear that aircraft designers have kept up with engine designers and have taken advantage of minimum-weight materials (and composites) to achieve efficient aerofoil designs. These physical changes, when combined with the use of a.c.t., have enabled the aeroplanes to operate in regions of r.s.s. and so achieve improved fuel efficiency (see *Figure 47.2*). When a.c.t. is applied to the three axes of the aeroplane it affords fuel-saving benefits. The Boeing B-757/769, Douglas MD-11 and the Airbus Industrie A-310/A-330 series (and other aircraft) exemplify the application of sophisticated flight-control systems.

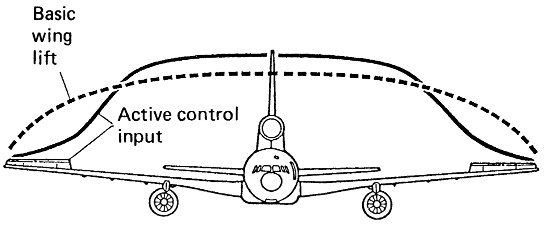


Figure 47.3 Spanwise wing-lift distribution

47.5 Flight-control systems

Historically, the flight-control surfaces in early aircraft were operated directly by mechanical control cables (or torque tubes) running between the control stick and the surfaces. Such mechanical controls were in fact still used on modern aircraft such as the Lockheed L-1011. These systems though reliable are mechanically complex, as they involve many bell cranks, walking beams and differential levers, etc. The complexity of the design, the close manufacturing tolerances and the necessary adjustments (to minimise backlash problems, etc.) characterise this as a high maintenance support flight-control system. Given the prospective obsolescence of such mechanical systems, the aerospace industry moved to the adoption of electric/electronic flight-control systems based on the use of fly-by-wire (f.b.w.) and power-by-wire (p.b.w.) systems.

In a f.b.w. system, triple or quadruple redundant computers transmit a serial digital bit stream to the power electronics of the remote f.c.s. actuators, which operate the control surfaces. A block diagram of a typical f.b.w./p.b.w. system is shown in *Figure 47.4*: the actuators can be duplex hydraulic jacks, electric or electrohydraulic. A view of a typical rotary electric actuator shown in *Figure 47.5* is used for spoilers.

Early f.b.w. systems used analogue data transmission and they go back to the early applications in the Mercury Space Vehicle (1960) and the XV-4B experimental aeroplane (in the mid-1960s): the US Air Force F-8 (1974) was the first aeroplane to adopt a *digital* flight control system (d.f.c.s.) (see *Figure 47.6*).

As discussed in Section 47.3, modern d.f.c.s.s brought to commercial and military aeroplanes benefits that were not possible with mechanical control systems. In addition to the ability to fly unstable aircraft, i.a.c.s provided better 'ride quality', eliminated (or damped) incipient wing-flutter conditions and allowed the aeroplane to fly through strong gust conditions without structural damage (*Figure 47.7*). Computerised flight control (with different flight options) also became possible and autoland systems were implemented that could meet the FAA's (Federal Aviation Association) category IIIb/IIIc class conditions for landings in low-visibility conditions.

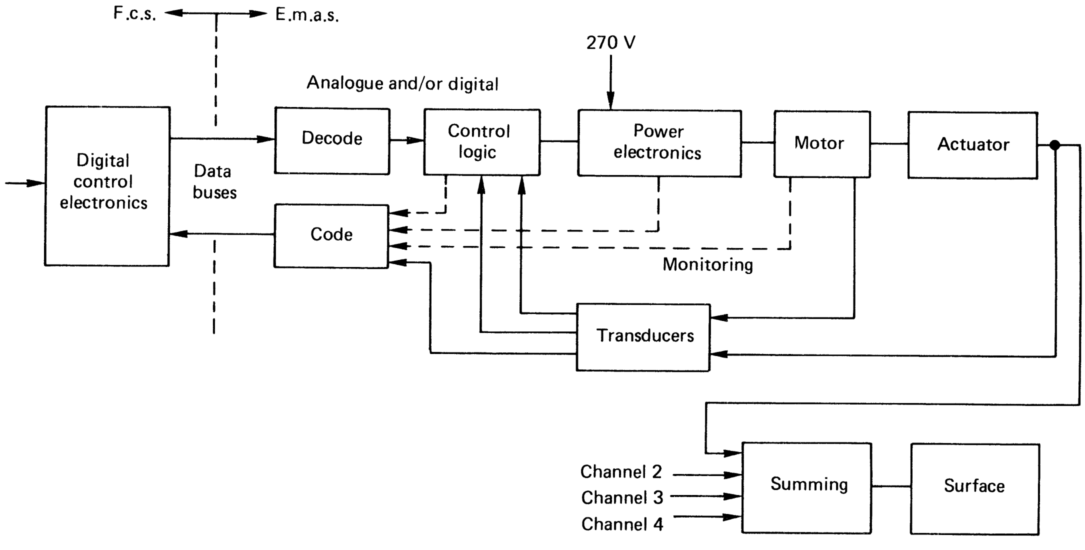


Figure 47.4 Digital fly-by-wire/power-by-wire system

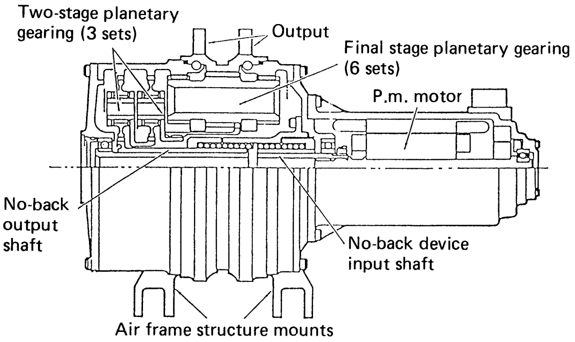

Two-stage planetary gearing (3 sets)
Output
Final stage planetary gearing (6 sets)
P.m. motor
No-back output shaft
No-back device input shaft
Air frame structure mounts

Figure 47.5 Typical rotary electromechanical actuator. (Courtesy of Sundstrand Aviation)

In the military aeroplane, manoeuvre load control (m.l.c.) is a feature that can be added to the d.f.c.s. and this enables the pilot to pull high g forces, during tight combat manoeuvres, without exceeding the vehicle's structural load limits. The d.f.c.s. also provides superior flight-management features in high-performance aircraft, which use interactive engine/f.c.s. controls. In these aircraft, the engine's two-dimensional thrust-vectoring system complements the flight-surface controls. The engine's sophisticated management system is typically effected via a computerised full authority digital electronic control system. These technologies are key to the complex control and flight management of aircraft such as the Harrier, AV-8A and other sophisticated aircraft.

47.6 Systems technology

Whilst an aeroplane's performance and operation are strongly tied to the engine, wing and f.c.s. technologies, the aeroplane is also strongly dependent on systems technology. As an aeroplane's 'propulsor system' is its 'primary' power source, the systems are defined as the *secondary power system* and they consist primarily of:

(1) engine bleed air,
(2) pneumatic,
(3) hydraulic, and
(4) electric.

As associative (dependent) parts of these 'power-source' systems, there are 'utility' or 'subsystems' such as the environmental control system, landing gear system, nose gear steering system, thrust reversing/wheel braking system, hydraulic, pneumatic and electric systems. These subsystems all extract power from the propulsor system (in pneumatic, hydraulic or electric form).

In a turbojet aeroplane, pneumatic power is extracted from the engine's compressor via annular 'taps' at the mid and last stages of the compressor. This high-pressure high-temperature air is then used typically for 'hot-wing' deicing, where hot pressurised air is injected into a double-wall leading edge via 'piccolo' tubes and manifolds with 'stub outs'. Engine deicing and engine starting are also normally powered by engine bleed air.

Pressurisation and cooling of the cabin/cockpit air-conditioning system are normally accomplished with bleed air by ducting the air from the power plants up to the flight station, where 'air-cycle packs' are located. Each of these packs comprises a bleed-air-powered turbine which drives the compressor (mounted on the same shaft) and at the same time provides expansion cooling as the air passes through the turbine. The air-cycle pack is lightweight and compact, because the turbo machinery runs on air-foil bearings at speeds of 80 000 to 100 000 rev/min.

In addition to providing bleed-air power, the engine normally has a 'waist-section' gearbox (located at the 6 0/C position) which provides multiple output (flange) pads to drive fuel/lube pumps, hydraulic pumps and electric generators. The engine gearbox also typically carries a pneumatic

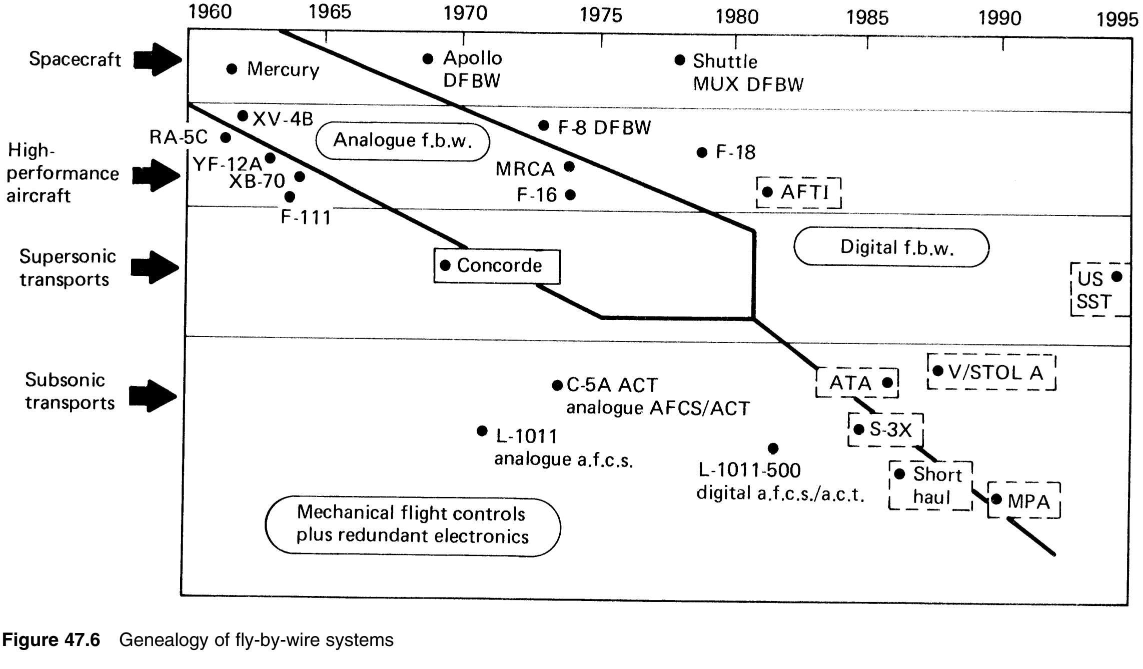


Figure 47.6 Genealogy of fly-by-wire systems

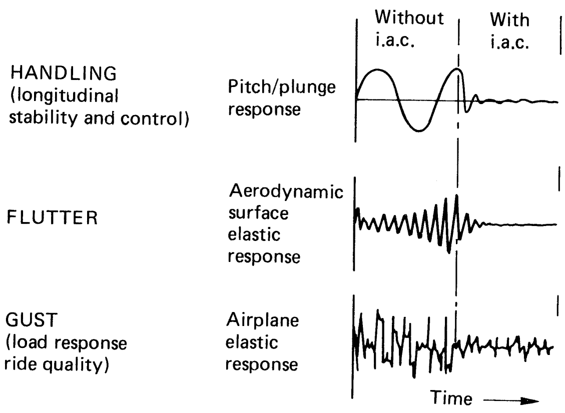


Figure 47.7 Benefits of integrated active controls

starter. *Figure 47.8* shows a typical transition from a conventional accessory gearbox to an all-electric configuration using an integrated engine generator/starter.

A typical accessory gearbox configuration is that mounted on the Rolls-Royce RB.211 engine. This is of a novel modular design, being driven by a power shaft from the high-pressure spool. The accessory gearbox and its accessories are typically cooled by engine-driven oil-lube pumps. The Rolls-Royce RB.211 is classified as a 'high bypass' turbofan engine which is able to furnish low and high pressure bleed-air; however, at high altitudes (35 000 to 42 000 ft) the engine core flow is much decreased, compared to low-altitude core flows. In a large commercial aeroplane carrying 300–400 passengers, there is a large e.c.s. bleed-air demand that, unfortunately, remains constant up to the high altitude and, thereby, creates an unfavourable fuel penalty. In contradistinction to this, if the e.c.s. were powered electrically the fuel penalty would only be some 25% of the bleed-air penalty. The difference between the thrust loss as applicable to the RB.211 in a commercial transport aircraft such as a Lockheed L-1011 or a McDonnell-Douglas DC-10 is shown graphically in *Figure 47.9*.

47.7 Hydraulic systems

Almost without exception, hydraulic systems (in past and current aircraft) used direct or indirect, engine-driven pumps that were typically of the constant-pressure/variable-displacement type. These pumps operated with delivery pressures of 3000, 5000 or $8000\,\text{lb/in}^2$ and at typical flow rates of 5–65 gallons per minute (gpm). The pumps normally have good efficiency (80–85%) at high flow rates, but at low flow rates, the efficiency falls off sharply. Therefore, under conditions when the flaps, spoilers, slats and landing gear have all operated (and the aircraft is in cruise flight) the fluid demand is very low, but the pumps still deliver some 92% of the take-off flow rate; consequently, there is a major mismatch between supply and demand! Another problem with conventional hydraulic systems is that the pumps exhibit constant heat dissipation, regardless of the flow rates (*Figure 47.10*). As a result, a comprehensive oil-to-fuel thermal management system is required.

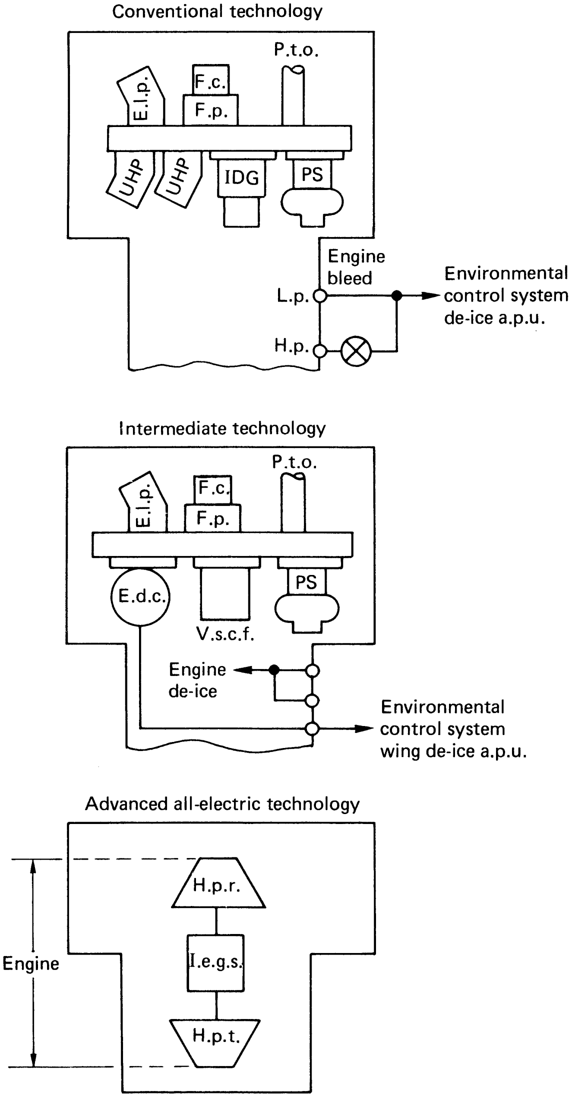


Figure 47.8 Secondary power system transition. E.l.p. engine lube pump; f.c., fuel control; f.p., fuel pump; u.h.p., utility hydrogen pump; i.d.g., integrated drive generator; p.s., pneumatic str.; e.d.c., engine driven component; h.p.r., hydraulic-pump rotor; h.p.t., hydraulic-pump turbine; i.e.g./s., integrated engine generator/str.; p.t.o., power take-off; v.s.c.f., variable speed constant frequency; a.p.u., auxiliary power unit

Hydraulic systems are, in the main, complex labour-intensive systems that comprise accumulators, reservoirs, pressure regulators, filters, noise attenuators and other fittings; but it is the 'distributed' hydraulic system that introduces the main problems of hydraulic systems because of leakage and contamination problems associated with the use of long fluidic lines.

In the majority of aircraft, the sources of hydraulic fluid power are typically pumps driven off the engine's accessory gearbox or off air-frame mounted accessory drives (see below). The latter are normally associated with high performance, small aircraft and their characteristics are discussed later in this text. In a centralised hydraulic system, one or

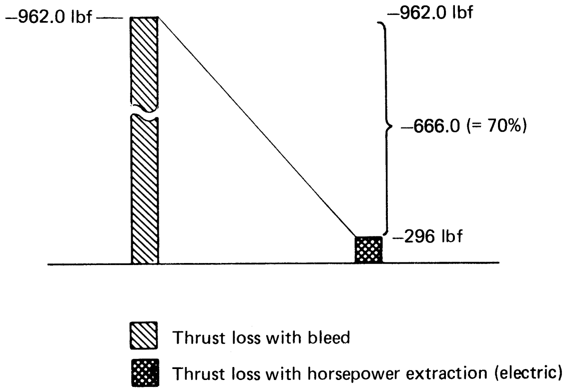


Figure 47.9 Thrust loss/bleed air versus shaft power extraction

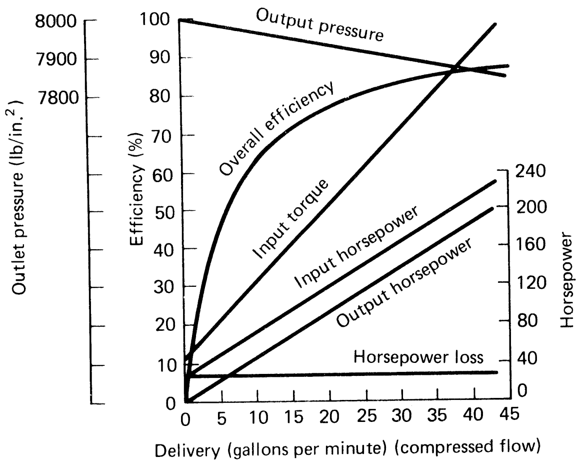


Figure 47.10 Typical performance curve for a hydraulic pump

more 'hydraulic load centres' are set up in the aircraft and the engine pumps feed into these load centres from which the hydraulic power is then distributed to the hydraulic loads (flaps, slats, landing gears, etc.). A multiple load-centre configuration, as used in the Lockheed P-7A ASW aeroplane, is shown in *Figure 47.11*.

In both these systems, many, many high-pressure (3000, 5000 or 8000 lb/in^2) lines thread their way through the aircraft creating a difficult physical installation. Also, because of vibration, the line joints are exposed to incipient leakage problems and, in military aircraft, to prospective fire hazards when missiles penetrate the aeroplane. The facts, nonetheless, are that hydraulic systems have a significant historical record of reliability over 40 years, and it is difficult to compete with the simplicity of a duplex (double piston) hydraulic jack. Furthermore, it is only recently that electrics could be used to take over the hydraulic functions such as landing gears, and the more sophisticated loads associated with the aircraft's primary/secondary flight-control surfaces. A new trend, however, has now been established under the umbrella of the US Airforce/WRDC MEL programme, which could lead to the use of electrohydraulic and electric actuators (see later sections). A typical two-channel hydraulic configuration is shown schematically in *Figure 47.12*.

47.8 Air-frame mounted accessory drives

The secondary power system normally comprises direct engine-driven electric generators (or integrated drive generators (i.d.g.s)) and hydraulic pumps. These accessories are typically mounted on a 'waist-section' accessory gearbox (a.g.b.) normally located at the 6 0/C position on the engine. On the Lockheed L-1011 aeroplane all the accessories such as the engine lube/fuel pumps, tachogenerators, generators and hydraulic pumps are driven by this gearbox. However, in advanced performance aircraft, such as the US Air Force YF-22 and YF-23 aeroplanes, two air-frame mounted accessory drives (a.m.a.d.s) are remotely driven (via disconnectable drive shafts) and the secondary power system components are mounted on these gearboxes.

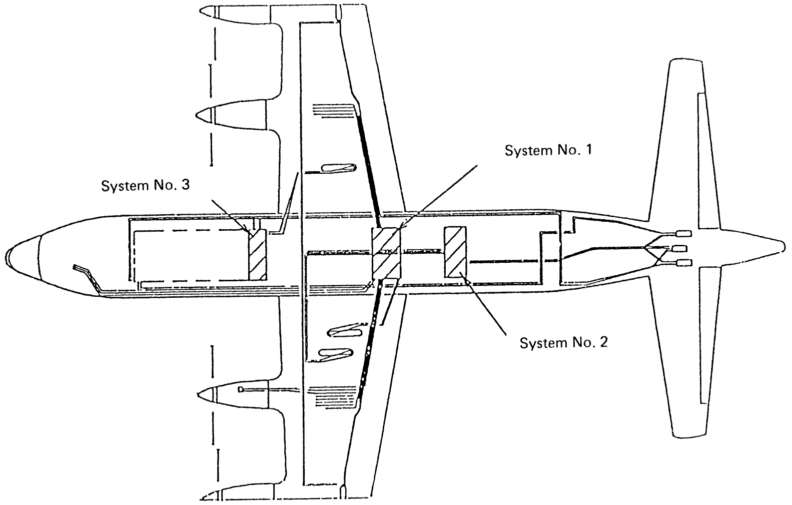


Figure 47.11 Diagram to show the location of the hydraulic bays in the Lockheed P-7A ASW

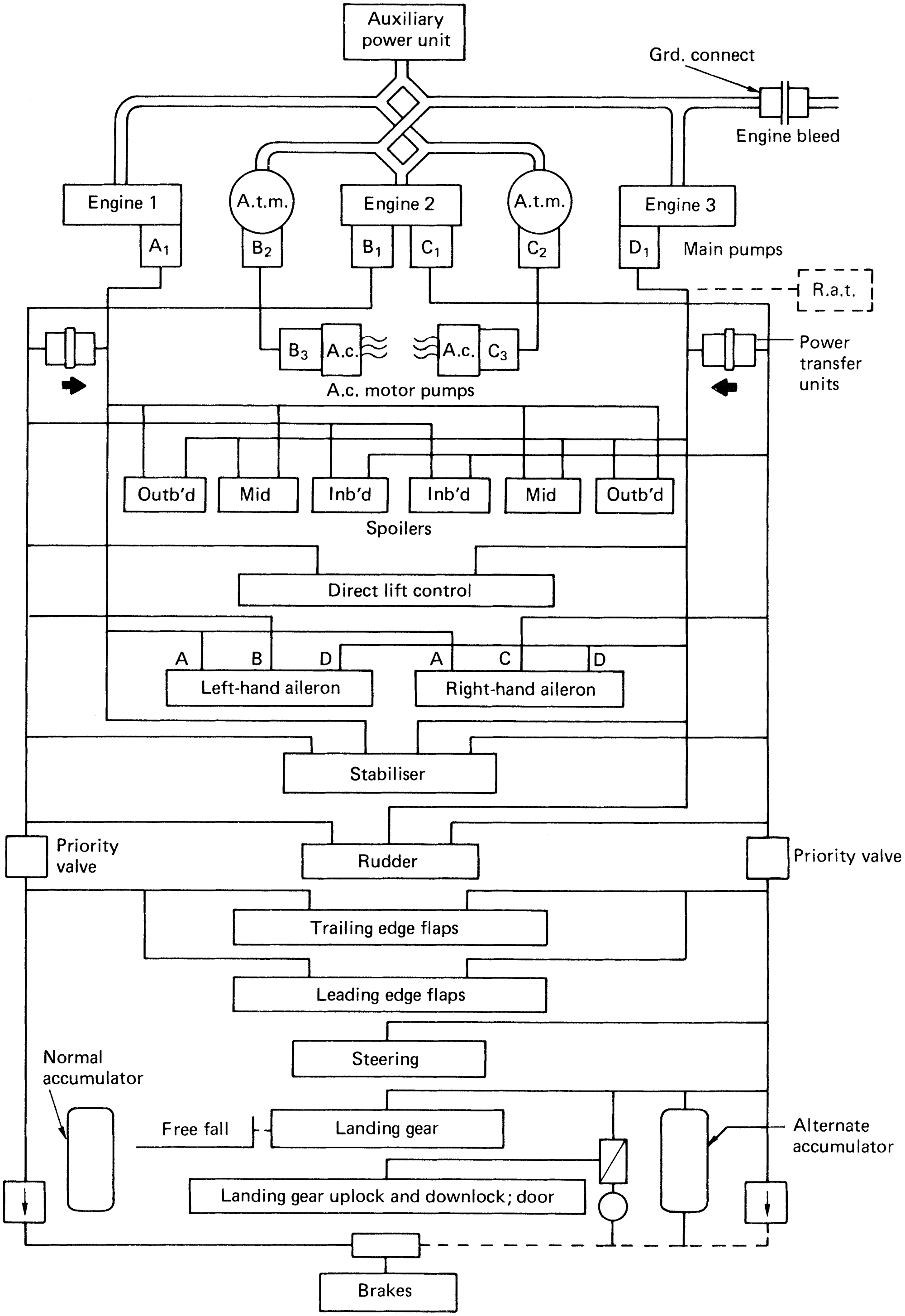

Auxiliary power unit
Grd. connect
Engine bleed
Engine 1
A.t.m.
Engine 2
A.t.m.
Engine 3
A_1
B_2
B_1
C_1
C_2
D_1
Main pumps
R.a.t.
B_3
A.c.
A.c.
C_3
Power transfer units
A.c. motor pumps
Outb'd
Mid
Inb'd
Inb'd
Mid
Outb'd
Spoilers
Direct lift control
A
B
D
A
C
D
Left-hand aileron
Right-hand aileron
Stabiliser
Priority valve
Rudder
Priority valve
Trailing edge flaps
Leading edge flaps
Steering
Normal accumulator
Free fall
Landing gear
Alternate accumulator
Landing gear uplock and downlock; door
Brakes

Figure 47.12 Dual hydraulic system. a.t.m., Air turbine motor; r.a.t., ram air turbine. (Courtesy of Lockheed Aeronautical Systems Co.)

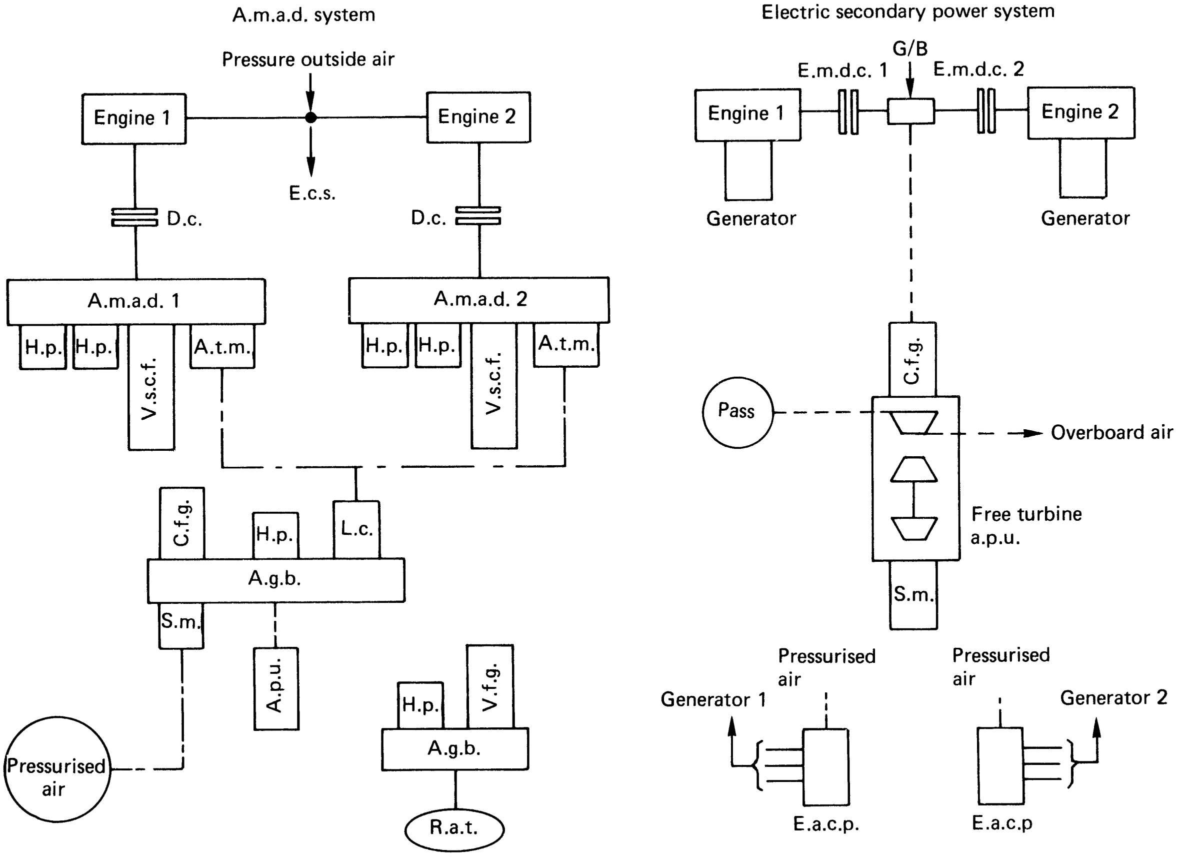


Figure 47.13 Comparison of a.m.a.d. and electric secondary power systems. a.m.a.d., Air-frame mounted accessory drive; d.c., disconnect coupling (engine); e.c.s., environmental control system; h.p., hydraulic pump; v.s.c.f., variable speed constant frequency; a.t.m., air turbine motor; a.g.b., accessory gearbox; e.a.c.p., electric air cycle pack; a.p.u., auxiliary power unit; c.f.g., constant frequency generator; r.a.t., ram air turbine; e.m., electric motor; l.c., load compressor; s.m.,...; v.f.g.,... Engines are turbojet/turboshaft engines

The reason for the a.m.a.d. configuration in the advanced technology fighters (a.t.f.s) is that these aircraft are committed to high sortie rates (and short turn-round times) and these objectives cannot be met with hydraulic pumps mounted directly on the engines. A.m.a.d.s were therefore the practical alternative, since they could be mechanically isolated when the engines were dropped from the aeroplane. This also avoided the prospective leakage and contamination problem that occurs when hydraulic lines are opened up. The a.m.a.d.s tend to be a legacy of the flight controls and, as shown in *Figure 47.13*, they are a complex, heavy, costly addition to the aeroplane. In addition, a.m.a.d.s decrease mission reliability and abrogate the possibility of achieving low maintenance-support objectives. Another problem with a large-capacity system using say four 156 fhp pumps is the *constant* heat dissipation which requires a dedicated oil-to-fuel heat exchanger to absorb the approximate 5800 British thermal units (Btu)/min heat losses: this is equivalent to about six 5 ton air conditioners!

In addition to the hydraulic pumps (and a variable speed, constant frequency (v.s.c.f.) generator), each a.m.a.d. carries an air turbine motor, which is used for ground and emergency flight operation. As shown, the aircraft's auxiliary power unit (a.p.u.) is interfaced with the a.m.a.d.s via pneumatic ducts. The a.p.u. also has an electric starter and a gearbox, on which are mounted a hydraulic pump, electric generator and a compressor. The compressor provides hot pressurised air (at approximately 52 lb/in^2) that can be selectively routed to either a.m.a.d. to permit running and checking of the a.m.a.d.s. The a.p.u. can also be operated in flight in the event of a single or dual engine 'flame-out'. In the YF-22/YF-23, there is almost critical dependence on an in-flight operation of the a.p.u., since the a.t.f.s are able to operate in regimes of aerodynamic instability. The a.p.u. also has an important ground-operation role, providing the a.t.f.s with power autonomy when they are dispersed at remote bare sites around an airfield.

In summary, a.m.a.d.s are a result of the selection of hydraulic flight controls; but with the trend to 'more electric' (m.e.l.) aircraft and the use of electric flight controls, electric (starter) generators could be the only components mounted on the engines. These generators however, could power electro-hydraulic and other advanced flight-control actuators in addition to the aircraft electrical and electronic systems.

47.9 Electrohydraulic flight controls

Electrohydraulic flight control units were used in the Vickers VC-10 (commercial transport) in the 1960s and each consisted of a compact integration of a motor pump,

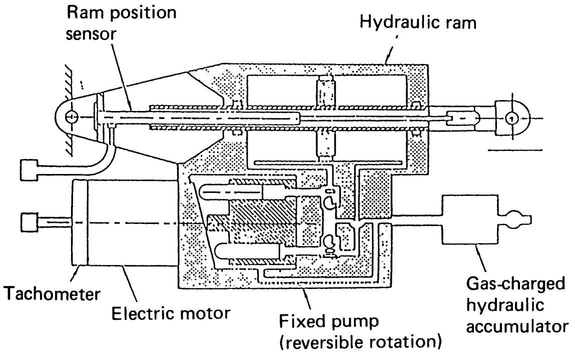

Ram position sensor
Hydraulic ram
Tachometer
Electric motor
Fixed pump (reversible rotation)
Gas-charged hydraulic accumulator

Figure 47.14 Electrohydrostatic actuator

a servo-control, a self-contained fluid, etc. A total of 11 units were used in that aeroplane; four *aileron* units, four *elevator* units, and three *rudder* units. Two units in each group incorporated an integral autostabiliser/autopilot control consisting of an electrical torque motor, operating on electric signalling. These primary flight controls were operated successfully and they complemented the three spoilers in each wing, which provided air-brakes, direct lift control and back up for the ailerons.

Recognising the major technology improvements that have taken place since the 1960s, e.h.a.s are now a low-risk technology. Also, many international companies produce a variety of electrohydraulic actuators. Some employ a constant speed/variable displacement/bidirectional pump or a constant-speed pump operating with a servo-valve control. More recently, however, an 'electrohydrostatic' actuator was developed which operates uniquely as a 'pressure-on-demand' actuator *with low heat losses* (*Figure 47.14*). A US company, H. R. Textron, California, also offers a novel actuator design which uses dual hydraulic jacks to impart a rotary via a high-lead-angle ball screw jack.

47.10 Electromechanical flight controls

Electromechanical actuators (e.m.a.s), like e.h.a.s, are attractive candidates for the new flight-control systems since they offer the same operational flexibility and advantages as e.h.a.s. In addition, losses in the motors and electronics decrease at light loads and e.m.a.s only dissipate high heat during the brief periods of high hinge moments. In an e.m.a. the motor current is proportional to torque:

$$T = K \times I_A \phi \cos\theta \qquad (47.3)$$

where T is the motor torque, ϕ is the stator flux, I_A is the rotor current, and $\cos\theta$ is the motor power factor.

Typically, the preferred implementation of an e.m.a. is a motor-driven 'power hinge' using a relatively high speed motor driving a 'power hinge' (attached to the control surface) via a planetary reduction gear train. However, linear e.m.a.s using high-efficiency ball screw jacks have also been used and, where high mechanical loads were present, the French-designed 'transrol' actuator offered advantages. In the field of novel actuators technology a US designed 'eccentuator' is of note. This utilises a 'bent shaft' which is rotated while its drive end (located in a 'carrier') is rotated in the opposite direction, so as to yield a planar output (*Figure 47.15*).

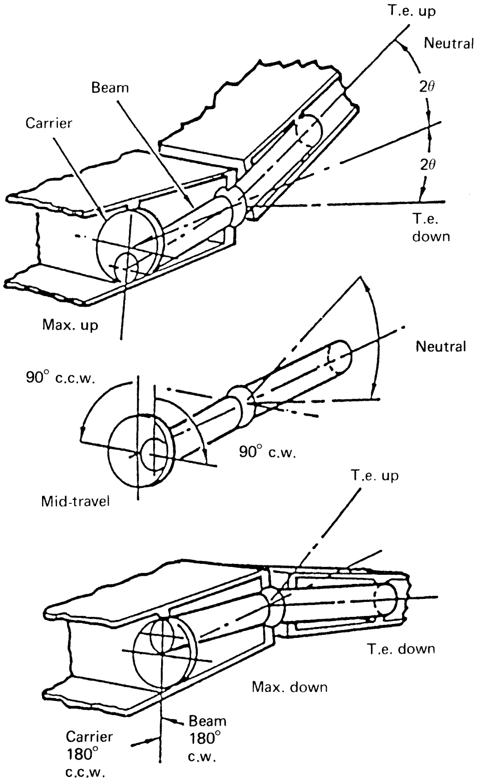


Figure 47.15 An eccentuator. The eccentuator motion compensates for the lateral displacement while doubling the output power. T.e., trailing edge θ, Beam bend angle; c.w., clockwise; c.c.w. counter clockwise.

With digital flight-control systems (f.c.s.) tri- or quadri-redundant computers transmit a digital bit-stream which contains rate, surface-displacement and directional information to the power electronics modules, that control electric actuators. Passive stabilisation networks, loop closure signals, etc., provide the necessary stability and performance characteristics of the actuators. The e.m.a.s provide high static/dynamic stiffness and bandwidth response, comparable with the best hydraulic actuators.

As an example of e.m.a. status, a 'motor controller' design by General Electric for a 12 hp actuator has dimensions of approximately 1 in × 3 in × 6 in. These modules actually integrate low and high voltage chips on the same substrate and they interface directly with a 1553B type digital data bus.

With the adoption of electric flight controls and other electric technologies, there ensues a major simplification of the aeroplane, due to the elimination of the labour-intensive (distributed) hydraulic systems and the elimination of the untenable practice of bleeding the engine compressors for high-pressure high-temperature pneumatic power.

47.11 Aircraft electric power

In World War II, the electric demands of the large four-engine bombers were met by four 28 V d.c. 6 kW generators which supplied the power demands of some highly electrified aircraft using electric turrets, electric de-icing, and nominally large electric/avionic systems. In the immediate post-World War II phase (1946 onwards), the aerospace industry realised that these low-voltage d.c. systems could no longer meet the projected future demands for electric power, so Advisory Boards were set up (in the USA and the UK) to evaluate higher voltage, higher power electric systems. These systems were premised on the use of higher d.c. voltages of 112, 240 and 340 V and higher a.c. voltages of 120 to 400 V; based on frequencies of 250, 400, 800 Hz and above.

However, the trend was towards a.c. power systems because of their ability to transform voltages, the ease of power switching a.c. currents, and the availability of highly reliable, rugged squirrel-cage induction motors. The three-phase, 200 V, 400 Hz system therefore became the popular choice and was selected by the author for the Bristol Brabazon I, Mark I aeroplane; a very large commercial aeroplane, designed for cross-Atlantic operation. The Mark I aeroplane was powered by eight Centaurus 2800 bhp (internal combustion) sleeve-valve engines buried in the wings. The later Mark II aeroplane was slated to use eight 3500 shp Proteus turboprop engines, but the Brabazon Mark II was not brought into prototype flying status. Both the Mark I and Mark II aircraft were configured with four pairs of engines, each pair driving two Rotol contra-rotating propellers, driven from a propeller reduction gearbox.

The electric system in the Brabazon I, Mark I, was probably the first major a.c. power system in the world and it consisted of six 40 kV-A Rotax a.c. generators and two Rotol 40 kV-A auxiliary power units (a.p.u.s).[1] A companion *large*-aircraft programme was the Saunders-Roe Princess Flying Boat which was to be powered by 10 Proteus engines. Uniquely, the electric system was designed around a 250 Hz system, *using a.p.u.s rather than engine-driven generators*. A variety of electric power systems were therefore evaluated in those early post-World War II years, but the trend was strongly established for three-phase 400 Hz a.c. power in the military and commercial aircraft. Constant speed hydromechanical drives and the later integrated drive generators (by Sundstrand) dominated the electric systems in these post-World War II aircraft and these transmissions furnished constant-frequency power over the variable-speed range of the engines. Other constant-speed drives were developed by General Electric and Lycoming. General Electric's drive was based on a 'ball pump' configuration, while the Lycoming drive was designed around a roller/toroidal (mechanical) transmission. These drives, unfortunately, encountered some pervasive development problems and there was only limited production. The problems (characteristic to mechanical and hydromechanical drives), are the high Hertzian stresses, incident upon the use of back-to-back hydraulic motor pump units and the mechanical complexity.

In an attempt to offset the perceived problems of mechanical/hydromechanical drives, the aerospace electrical companies developed electronic approaches to generating constant-frequency power. These were known as variable speed constant frequency (v.s.c.f.) power systems and they were based on 'cycloconverter' and 'd.c. link' conversion techniques. These systems held the prospective advantages of higher efficiency and higher reliability; also they were able to take advantage of the burgeoning technology in power electronics and very large scale integrated (v.l.s.i.) chip technology.

Nonetheless, constant-speed hydromechanical drives and integrated drive generators remained the power systems of choice in modern aircraft, except for the new trend towards 270 V d.c. However, when large-capacity systems (300–500 kV-A) become necessary, a *direct-driven generator, operating with a constant volt/cycle characteristic, could become an efficient and practical alternative*. This would enable the generator to operate over a 2:1 speed ratio, without compromising its weight and size *vis à vis* an optimally sized constant-speed machine. In addition, since it would be desirable to use such a large generator to start the engines, the synchronous generator could be operated as a synchronous motor, deriving its power from a *converter producing programmed voltage-to-frequency (V/F) power*. Furthermore, once installed, the inverter/converter can be used for other purposes in the aeroplane. For example, a motor-driven e.c.s. pack could be driven at speeds of 80 000 to 100 000 rev/min, permitting the use of 'air-foil' bearings, as used with bleed-air turbine/compressor units. For these speeds, a two-pole (low slip) induction motor would require a frequency of 1600 Hz for a speed of 94 000 rev/min, approximately.

Implicit in the adoption of a three-phase a.c. power system, operating with a constant volts per cycle voltage control, is the fact that the generator operates with a 'power proportional speed' characteristic. Thus, for the typical 2:1 speed range of the turbojet engine, the generator output would be reduced to 50% on the ground. There is, however, synergy in this because, in the real world, the generator would never be required to provide full power during ground operation since loads such as windshield heating, galley, wing/engine de-icing, the utility electric loads, and the avionic loads would never all be switched on at one time. Similarly, electrohydraulic pumps, landing gears, etc., would not be operating.

In contrast to the above, *aircraft power generators in almost all aircraft today are effectively 2:1 oversized*. For example, in the case of a generator designed to provide 270 V d.c. at 50% per unit (p.u.) speed, it *'wants' to produce 540 V at 1.0 p.u. speed*, but it is harnessed by the voltage regulator to 270 V. Also, when the 270 V d.c. is developed by rectifying the a.c. generator output by a three-phase full-wave silicon controlled rectifier (s.c.r.) bridge rectifier, only a small part of each (rectified) sine wave is used to produce the 270 V d.c.; as shown in the lower half of *Figure 47.16*. There is the irony also that the bridge-rectifier/filtration components are only efficient on the ground, at 0.5 p.u. engine speed!

In the case of a hydromechanical transmission, the 'speed penalty' shows up as a 'power mismatch'. As shown in *Figure 47.17*, the integrated drive generator can produce full power (at half speed), *but the ground-power demand is very low*. It is also of note that the parabolic (power) curves tend to be asymptotic at the 0.5 and 1.0 p.u. speeds, so the electrical, fluidic and mechanical losses tend to remain high at these point extremes.

47.12 Summary of power systems

Clearly, there is no panacea for the problems typical of the 2:1 speed range of turbojet engines, and this is somewhat unique to the aircraft power systems. In utility electric systems, steam turbines drive generators at constant speed, so they intrinsically produce constant voltage/constant frequency power. In contrast, the aerospace electrical engineer must resort to innovative approaches to meet the speed-range problem. Basically, the choices are:

(1) constant-speed drives,
(2) v.s.c.f. cycloconversion,
(3) v.s.c.f. d.c. link conversion,
(4) 20 kHz resonant power conversion,
(5) constant volt/cycle, or
(6) 270 V d.c.

None of the above is without some disadvantages with regard to costs, reliability, weight, efficiency, maintenance

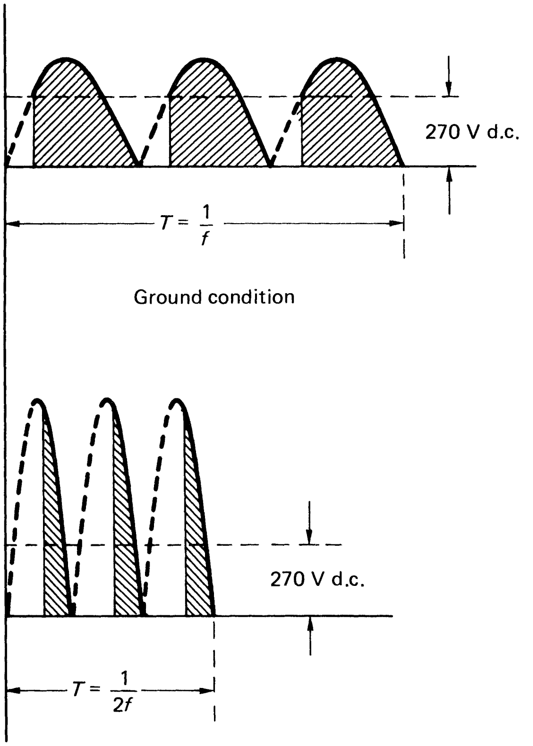


Figure 47.16 Aircraft power generation system: 270 V d.c. power system

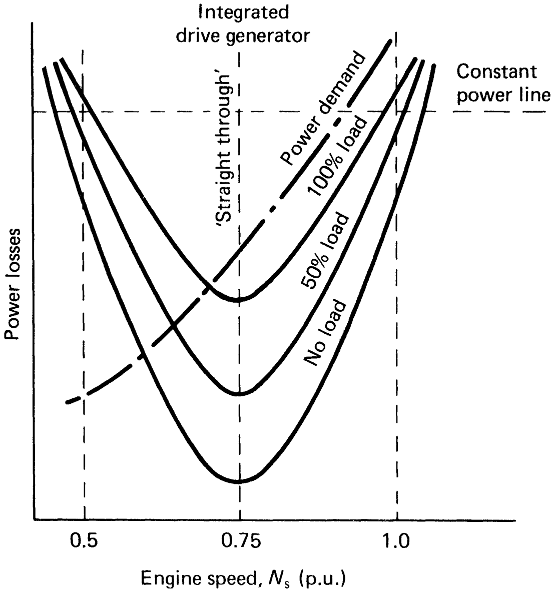


Figure 47.17 Aircraft power generation system: constant-speed drive system

support, life-cycle costs, etc., but it is possible to draw on the expertise, resident in the above power generation systems to develop a simple 'universal' type power system that could furnish multiple, 'dedicated' power supplies for the different loads. One such system, shown in *Figure 47.18* utilises a single high speed (direct driven) 500 kV-A generator, which can operate at a speed of 24 000, 36 000, or 48 000 rev/min. The figure shows a simplified scheme of a generator supplying three, or more, power supplies, such as:

(1) three-phase 200 V, 400 Hz (or three-phase 400 V, 800 Hz),
(2) 270 V d.c.,
(3) 28 V d.c., or
(4) other

Examples of such systems include the following.

(1) A Westinghouse 500 kV-A unit furnishing multiple and different characteristic outputs—120 kV-A of three-phase, 200 V, 400 Hz power; 270 kW, 13.2 kV d.c., 17 kW (pulsed load), 28 V d.c.
(2) An Allied Signal 500 kV-A generator with a weight of only 136 lb, when driven at a speed of 48 000 rev/min. This speed is somewhat high for a direct engine-driven generator, but it is suitable for a turbine drive.

The advantages offered by this hybrid system are that *very high-capacity low-weight generators can directly power large loads* such as landing gears, swing-wings, tilt-tails, de-icing systems, space-heating, lighting, utility loads, and (in commercial aircraft) galley loads. The dedicated loads such as 270 V d.c. and 28 V d.c. (and special high frequency power supplies) can be derived on an 'as-required' basis. However, with advanced power generation systems, *in the 150–500 kV-A per channel capacity, 3% or less conversion may be more than adequate for each of the dedicated power supplies*. This, therefore, significantly reduces the required sizes of the power diodes/filters, etc., compared to the sizes required in a direct engine-driven 270 V d.c. generator. Difficulties in rupturing the prospectively high short-circuit currents and the heat dissipation of solid state power controllers would be a concern in such high voltage d.c. systems.

In contrast, the selection of an a.c. power system simplifies the switchgear problems and permits the use of transformers to provide multiple different voltages for windshields, props/spinners, etc. Also, three-phase a.c. power permits the use of highly reliable, lightweight squirrel-cage induction motors for many loads such as electromechanical actuators, fuel/oil motor-pump units, passenger/cargo/weapon doors, hermetically sealed Freon compressors, motor-driven, bootstrap/air-cycle machines, etc.

Loads such as constant-pressure electrohydraulic pumps may be operated over a 2:1 (or more) speed range and still provide constant torque (*Figure 47.19*(*a*)), as demanded by 3000 or 5000 psi pumps. The efficiency–torque characteristics (*Figure 47.19*(*b*)) also show that *the efficiencies remain high for the four V/F curves* shown.

One of the problems with 270 V d.c. systems is that the 270 V d.c. motors cannot operate directly on 270 V d.c., so the d.c. voltage must be 'inverted' to power the three-phase a.c. stators of the many drive motors; ironically, these inverters would actually furnish constant V/F power!

47.13 Environmental control system

The environmental control system (e.c.s.), unlike many other services, represents an almost constant-horsepower

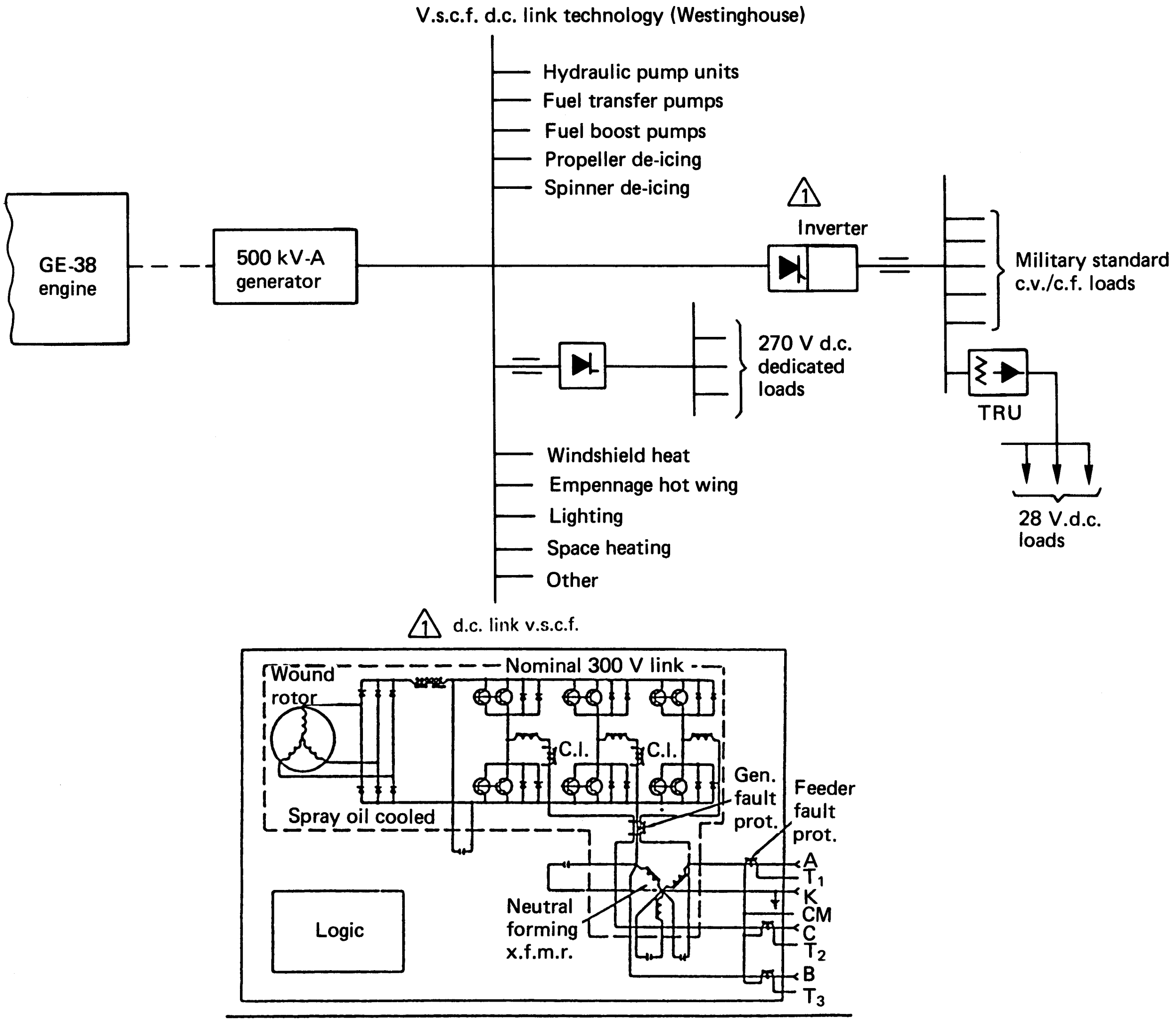

V.s.c.f. d.c. link technology (Westinghouse)
Hydraulic pump units
Fuel transfer pumps
Fuel boost pumps
Propeller de-icing
Spinner de-icing
GE-38 engine
500 kV-A generator
1
Inverter
Military standard c.v./c.f. loads
270 V d.c. dedicated loads
TRU
28 V.d.c. loads
Windshield heat
Empennage hot wing
Lighting
Space heating
Other
1 d.c. link v.s.c.f.
Wound rotor
Nominal 300 V link
C.I.
C.I.
Spray oil cooled
Gen. fault prot.
Feeder fault prot.
A
T1
K
CM
C
T2
B
T3
Logic
Neutral forming x.f.m.r.

Figure 47.18 Advanced electric power system (with d.c. link v.s.c.f. power)

demand over the complete flight envelope, regardless of engine power settings and engine speeds. This, in the past, has been a significant problem for the bleed-air system and it impacts on any mechanical power system, including an electrically powered e.c.s.

For a 500 passenger advanced transport aircraft (a.t.a.), the increased fuselage length, the larger number of windows, and the higher thermal transconductance all result in a significant increase in the capacity of the e.c.s. Without taking advantage of air recirculation (50% or more), the horsepower demand could be excessive. Therefore, it is axiomatic that a.t.a.s utilise a degree of air recirculation: the usual problems of doing this will be ameliorated by the prospectively lower level of vitiated air which can be expected, in the future, when there will be less nicotine contamination of the air. Also, improved filtration methods are now available.

Basically, the e.c.s. performs the following functions:

(1) cabin pressurisation,
(2) cabin cooling,
(3) cabin heating, and
(4) cabin humidity.

Typically, cabin altitude is maintained at 6000 ft, which is equivalent to about 11.8 lb/in^{2}a. At normal cruise altitudes, an intermediate tap on the engine compressor supplies the pressurisation needs, but at idle descent letdown it is necessary to change over to a last-stage bleed. Duct pressures and temperatures near the bleed ports are typically 100–300 lb/in^{2}a and 200–700 °F. These higher pressures and temperatures are necessary, in conventional aircraft, where the hot pressurised air is used for wing anti-icing and for air-turbine motor drives. However, these temperatures and pressures are actually too high for the cabin e.c.s. requirements, so intercoolers and air cycle machinery are necessary to condition the air before it is introduced into the cabin. Outflow valves modulate to maintain the cabin pressure at the approximately 6000 ft altitude.

For all-electric e.c.s. (a.e.e.c.s.) pressurised air is derived from motor-driven compressors. For a typical 500 passenger a.t.a., the induction motor-driven compressors are

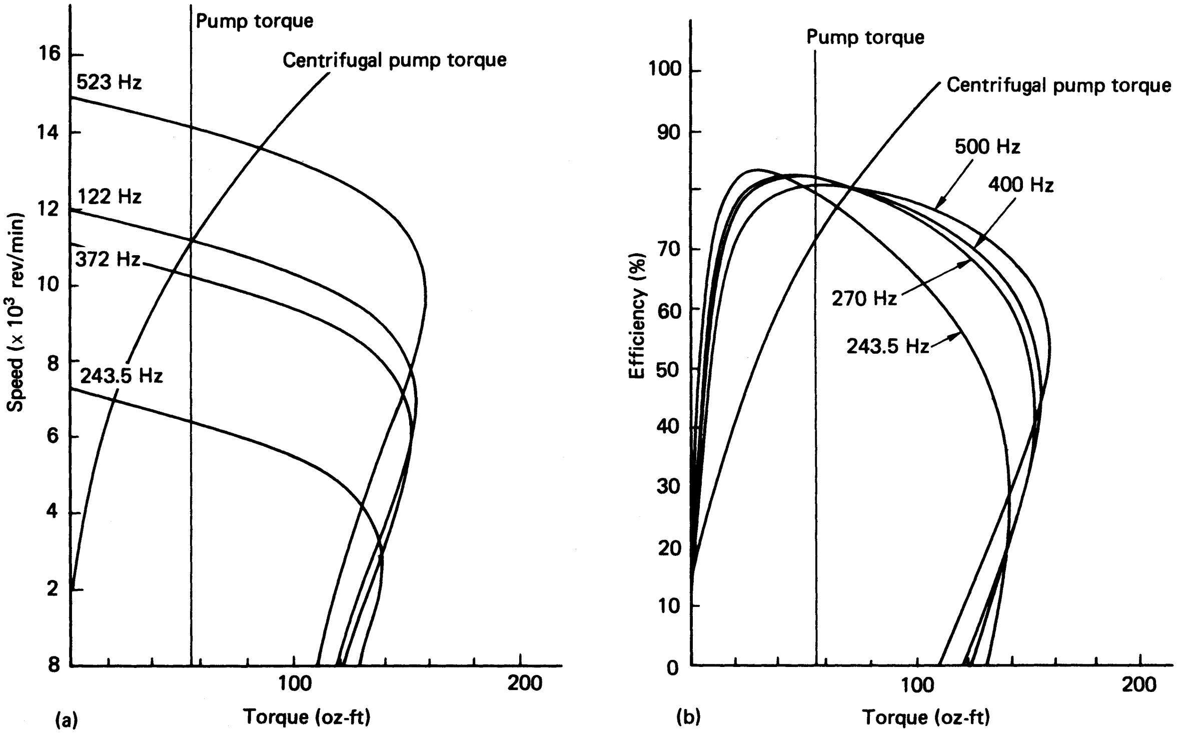


Figure 47.19 Performance of electrohydraulic pumps on constant V/F power: (a) speed–torque curves; (b) efficiency–torque curves

designed to furnish approximately 1.2 ppm per passenger, made up of 300 ppm of fresh air and 300 ppm of recirculated air. The pressure ratio of each compressor is approximately 3.4:1, so the heat of compression can be used to supply the heat requirement during winter operation. The temperature rise across such compressors is defined as:

$$\Delta F = \frac{T_a^o + 460}{\eta_c}[\mathrm{PR}^{\frac{K-1}{K}} - 1] \qquad (47.4)\Leftarrow$$

where T_a is the ambient temperature (in degrees Farenheit), PR is the pressure ratio, K is the ratio of the specific heats, and η_c is the compressor efficiency. Δ°F is temperature rise in degrees Fahrenheit.

The power to drive each compressor is given by

$$\mathrm{HP} = \frac{\dot{\omega}\Delta T}{2.95} \qquad (47.5)\Leftarrow$$

where $\dot{\omega}$ is in pounds per second and ΔT is ΔF in equation (47.4).

Equations (47.4) and (47.5) show that the ΔT value for air across the compressors is approximately 225 °F and the compressor power is approximately 120 hp. To avoid overloading the motor compressors during low-altitude, high engine speed conditions, the compressors incorporate an inlet-guide-vane control and a pole-changing motor, which changes speed from 48 000 to 24 000 rev/min nominal.

The cooling requirements can be furnished by three vapour-cycle cooling units, which employ a reverse-Rankine cycle with a R-114 refrigerant: the moisture content, which affects the capacity of the cooling system, was taken as 130 gr/lb. Condenser cooling is effected by fan-induced air flow on the ground, and ram air in-flight. Inlet-guide-vane control is used to modulate the Freon flow and cooling capacity during light-load conditions.

Cabin cooling is a major load on the ground, with a fully loaded aircraft and outside air temperatures at ≥100 °F. The metabolic load of 476 passengers and 24 crew members, the solar input, and other internal heat dissipation factors add up to a total cooling load in excess of 375 000 Btu/h for the large a.t.a. In a conventional aeroplane such as the L-1011, cooling is derived from three air-cycle units, which are an integral part of three e.c.s. packs. To furnish this power on the ground, an a.p.u. driven compressor is used.

Cabin heating tends to be a less significant requirement than cooling in that there is still a demand for cooling, during high-altitude cruise when there is a maximum passenger complement. The heating demand therefore applies mainly to cold-day conditions with low passenger densities, i.e. low internal losses. Evidently, with bleed air, there is an adequate supply of hot air.

As is typical with industrial or commercial vapour-cycle cooling systems, the liquid R-114 is flashed to a cold two-phase low-pressure liquid gas system, as it flows through the expansion valve; this liquid gas is supplied to the evaporators (for cooling the motors driving the compressors).

While a minimum weight was not emphasised in this e.c.s. design, a 1500 lb weight saving was achieved by using an all-electric e.c.s. instead of a conventional bleed-air powered e.c.s.

Figure 47.20 shows an a.e.e.c.s. that is a power regenerative system which utilises 'air cycle' and 'vapour cycle' cooling. The former consists of an air-cycle pack in which a compressor and a turbine are driven at high speeds (75 000–90 000 rev/min) by a low-slip, squirrel-cage induction motor supplied with high frequency. This whole

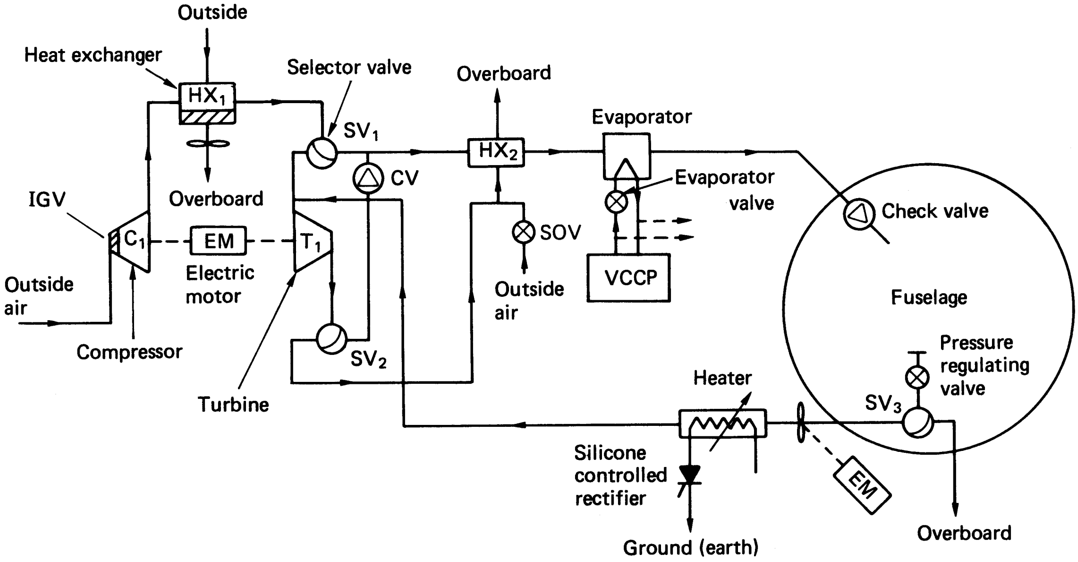


Figure 47.20 Electric e.c.s. with air cycle and vapour cooling. EM, electric motor; IGV, inlet guide vanes; C, compressor; T, turbine; SV, selector valve; VCCP, (electric) vapour cycle cooling pack; HX, heat exchanger; PRV, pressure regulating valve; CV, check valve; SCR, silicon controlled rectifier; EV, expansion valve

assembly runs on air-foil bearings to provide long life and reliability. The compressor employs an inlet guide vane to improve efficiency and to modulate the air mass flow through the single-stage centrifugal compressor. The turbine itself furnishes two functions: it offsets the amount of power supplied to the compressor by the electric motor and provides cold air to the cabin. As mentioned above, the purpose of the e.c.s. is to provide conditioned air to the cabin and passengers.

The simplified scheme given in *Figure 47.20* shows the e.c.s. operating in a typical flight condition (at 35 000 ft) where power 'regeneration' is effective. For example, at 35 000 ft, the atmospheric pressure is 3.5 lb/in^{2}a and the cabin pressure (for 6000 ft equivalent altitude) is approximately 11.6 lb/in^2: this is equivalent to a pressure ratio of 3.3:1. This pressure ratio is then effective across the turbine, as the cabin outflow air is dumped overboard (to modulate the 11.6 lb/in^2 pressure). As the outflow air discharges through the turbine, it develops shaft power to offload the motor-compressor drive. (Note: this a.e.e.c.s. is shown operating as a 100% 'fresh-air' system, but it can utilise a 50% 'recirculation' system; if this were adopted, however, the power regeneration would be reduced by 50%.)

During hot-day ground operating conditions, the total cabin-heat load is very high, so in this mode the compressor and turbine are connected in series by changing the position of the selector valves. With the selector valves in this position some of the heat of compression is removed by the heat exchanger (HX_1) and further heat is removed by the turbine: The actual drop in temperature across the turbine is determined using Equation (47.4). From the turbine, the air passes through a second air-to-air heat exchanger (HX_2) and then through the Freon evaporator; the flow of Freon through the evaporator may be modulated in accord with the desired evaporator exit temperature (typically 40 °F). In this ground operation mode, the pressure rise across the compressor is reduced in the same ratio as the pressure drop across the turbine. The air entering the cabin is, therefore, at the same pressure as the outside (14.7 lb/in^2) and is discharged overboard with minor pressure loss. On cold winter days, electric heat may supplement the metabolic heat load of the passengers by passing fan-recirculated air over an electric heater, whose power-dissipation is varied by a silicon controlled rectifier.

47.14 Digital power/digital load management

Presently used aircraft are still flying with 1960–1975 electrical technology and very little design activity has taken place to bring electric systems into line with the digital avionic and engine management systems. The NASA and the US Air Force/Navy laboratories have long promoted advanced digital power systems that would interface with solid state power controllers (s.s.p.c.s), controlled and monitored by a 'data handling system'. The US Navy Air Development Center (NADC) has also installed two A-7 mock-ups in their laboratories at Warminster and have successfully demonstrated the advantages of modern power management over conventional power management.

Another key objective of the NASA and US military services is to adopt a 'distributed power bus' system, which would simplify power distribution to the loads and constrain the *power wires to the areas in which the loads are located.* This technology also inhibits the 'bus-proliferation' problem, which is a legacy of the conventional radial power systems now used in aircraft. For example, power feeders typically route from the engine-located generators to a remote (forward fuselage) main electric load centre. Radial power feeders then feed back to remote areas where the power feeders originated!

In the advanced power distribution system, insulated power feeders traverse the wings and the fuselage. The concept is that in a four-engine four-generator configuration *there are only four distributed buses.* The feeders are also redundant, i.e. insulated/isolated quadri-redundant feeders route in the fuselage and two sets of feeders in each wing are

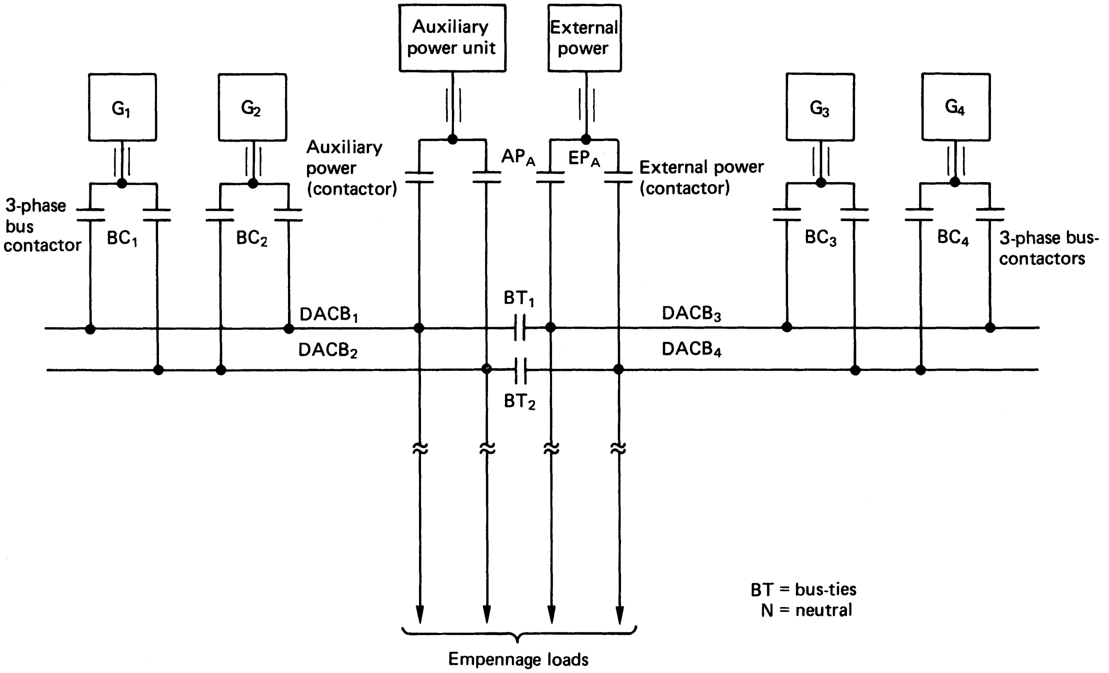


Figure 47.21 Advanced power generation system with distributed buses. APU, auxiliary power unit; G, generator; DACB, distributed a.c. bus; BC, bus contactor; EP, external power contactor; AP, APU power contactor; BT, bus-tie/buss

interconnected by bus ties. The distribution system may be operated as a 'closing-ring' system in which, in an extended-emergency condition, any one generator can feed all bus sections (*Figure 47.21*).

In the implementation of such a distributed bus system, subfeeder taps are made into the buses (as they traverse past key load areas) and they are protected *at the tap points*: this is essential *in order not to compromise the primary feeders*. The key and enabling technology in this advanced power system is a digital power/digital-load management system (d.p./d.l.m.s.).

To preserve the philosophy of isolated quadri-redundant power supplies, *the generator channels are not paralleled.* In addition, the health and status of each generator is monitored by a microprocessor-based generator control unit (*Figure 47.22*) that communicates with the flight station computer display unit over a 1553B, or other dedicated digital data bus (*Figure 47.23*). In addition to the management of the individual power channels, the d.p./d.l.m.s. controls the bus ties and the interfaces with the external/a.p.u. power systems.

Figure 47.24 shows a simplified scheme of the d.p./d.l.m.s. management system designed by the Leach Corporation. The electric load management centre (e.l.m.c.) interfaces with the 1553B bus and a MIL-STD-1750 processor (with a 128k word memory) manages the complete electric system via a dedicated load management distribution bus. This bus interconnects the 'remote power modules' (similar to the one identified in the e.l.m.c.) to control the ON/OFF status of the s.s.p.c.s and the smart remote power contactors (s.r.p.c.s).

The digital load management system makes possible the control of all loads via the distributed bus system and discrete (multi-bit) address code and a logic code. This code also ensures that the s.s.p.c. closes only when all the circuit parameters are correct. In addition, the individual loads are given a 'priority tag' which establishes their level of importance in the overall hierarchical structure of the loads in the aeroplane. Loads may therefore be in, say, four levels; in which loads in group 1 would be first isolated, followed by those in groups 2, 3 and 4, as the level of emergency increases. For example, if smoke occurs in the vehicle, all four groups would be disconnected, leaving only the flight-essential loads connected. From this, it is clear that power reduction is effected *not by dropping dedicated buses, but by a software-controlled dumping of individual loads.*

The s.s.p.c./s.r.p.c. combines the features of a relay and a circuit-breaker but, unlike circuit-breakers, they can detect open-circuit and open-phase faults. By the use of the d.p./d.l.m.s., multiple circuit-breaker panels are eliminated and there is a significant reduction in the amount of load wiring.

Reference

1 CRONIN, M. J. J., 'The development of the electrical system on the Bristol Brabazon I Mark I Aircraft', *Proc. IEE, Part I* **98**(113) (September 1951)

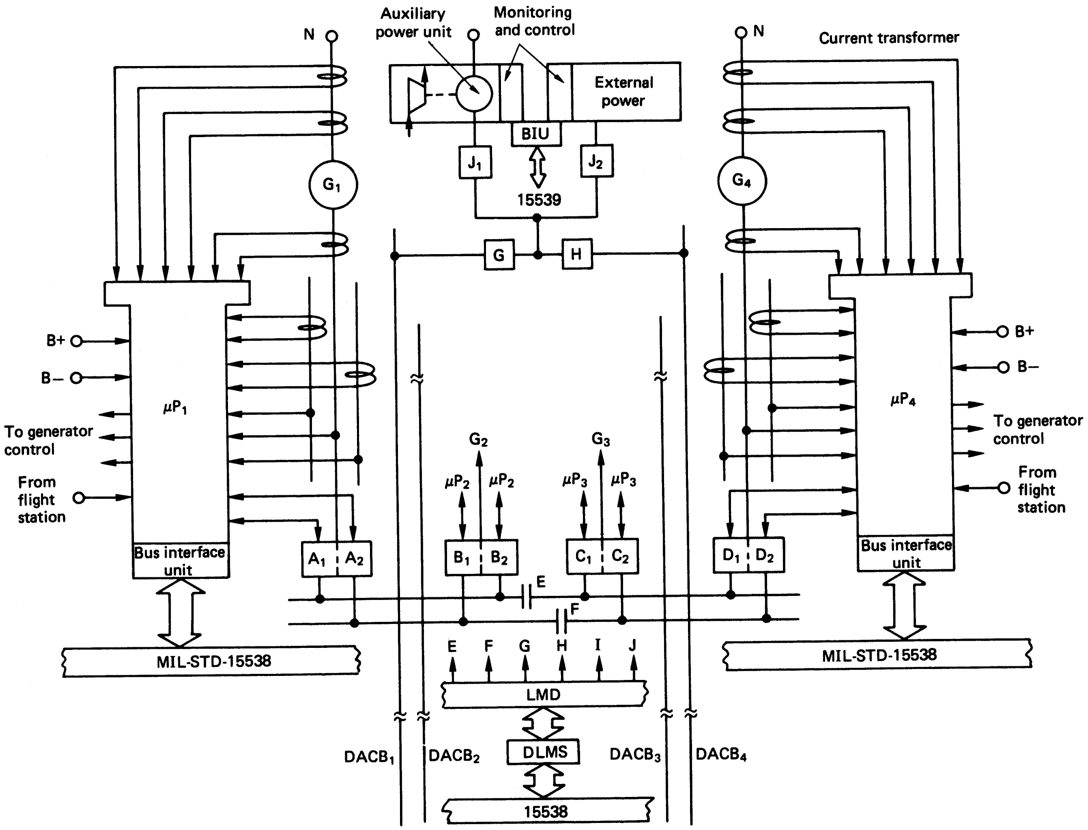

N
Auxiliary power unit
Monitoring and control
External power
N
Current transformer
BIU
J_1
J_2
15539
G_1
G_4
G
H
B+
B−
μP_1
To generator control
From flight station
Bus interface unit
A_1
A_2
G_2
μP_2
μP_2
B_1
B_2
G_3
μP_3
μP_3
C_1
C_2
D_1
D_2
μP_4
B+
B−
To generator control
From flight station
Bus interface unit
E
F
MIL-STD-15538
E F G H I J
LMD
DLMS
15538
MIL-STD-15538
$DACB_1$
$DACB_2$
$DACB_3$
$DACB_4$

Figure 47.22 Advanced power generation system with microprocessor control. BIU, bus interface unit; μP, (generator) microprocessor; G, generator; A, B, C, D, generator power contactors; DACB, distributed a.c. bus; DLMS, digital/power load management system; LMD, load management distribution; F/S, (flight station); E to J, contactors; N, neutral

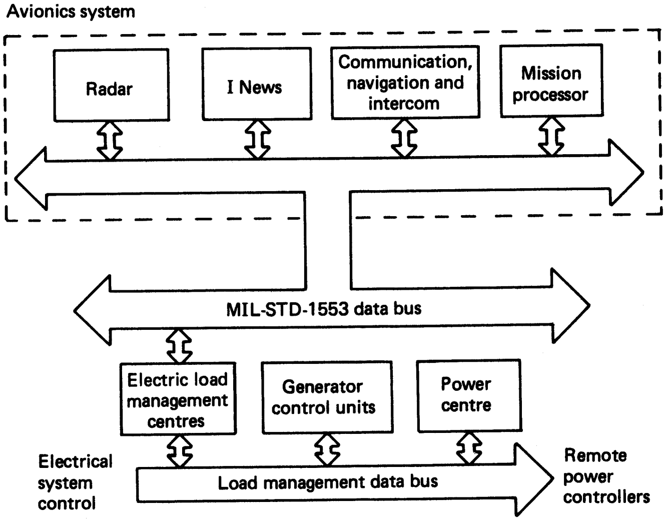
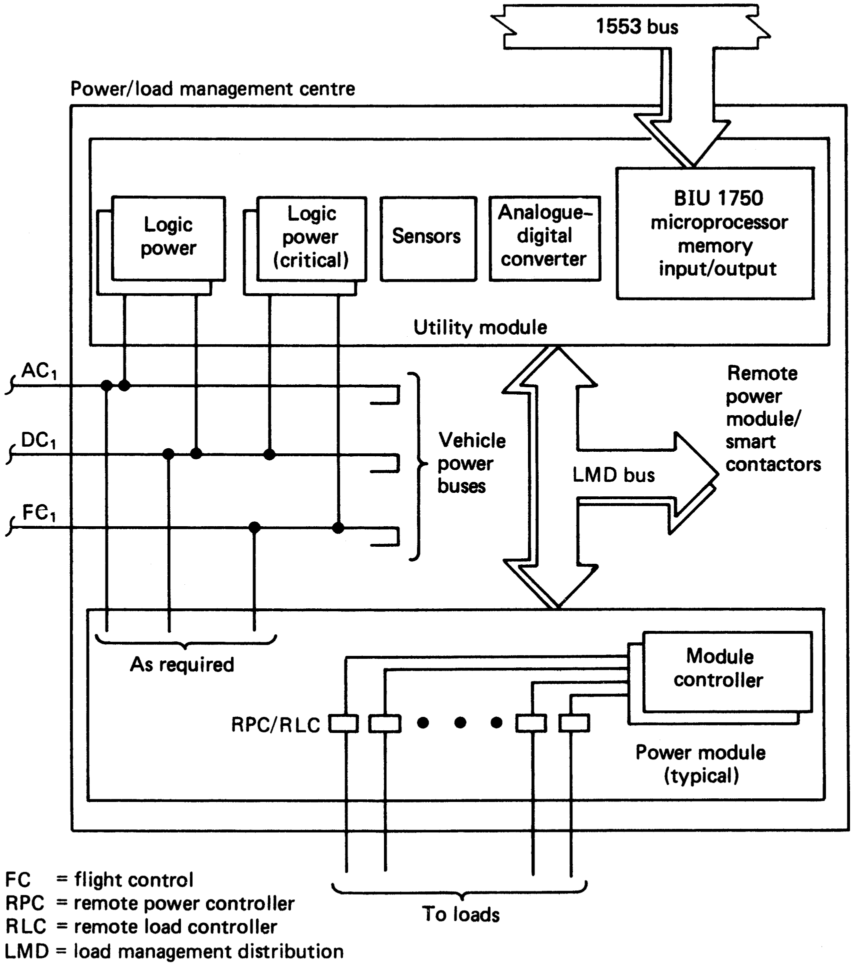

Avionics system
Radar
I News
Communication, navigation and intercom
Mission processor
MIL-STD-1553 data bus
Electric load management centres
Generator control units
Power centre
Electrical system control
Load management data bus
Remote power controllers

Figure 47.23 DP/LMS Configuration (Leach designed alternative). The electrical system is reconfigurable to ensure that the mission is completed


1553 bus
Power/load management centre
Logic power
Logic power (critical)
Sensors
Analogue-digital converter
BIU 1750 microprocessor memory input/output
Utility module
AC_1
DC_1
FC_1
Vehicle power buses
LMD bus
Remote power module/ smart contactors
As required
Module controller
RPC/RLC
Power module (typical)
To loads

FC = flight control
RPC = remote power controller
RLC = remote load controller
LMD = load management distribution

Figure 47.24 Digital-power/digital-load management

48

Mining Applications

R Hartill BSc(Hons), CEng, FIEE, Hon FIMEMME, EurIng
Trolex Ltd

Contents

48.1 General 48/3
  48.1.1 Load growth 48/3
  48.1.2 Regulations 48/3

48.2 Power supplies 48/4
  48.2.1 Distribution 48/4

48.3 Winders 48/5

48.4 Underground transport 48/6
  48.4.1 Conveyors 48/6
  48.4.2 Rope haulage 48/8
  48.4.3 Locomotives 48/8

48.5 Coal-face layout 48/9

48.6 Power loaders 48/11

48.7 Heading machines 48/12

48.8 Flameproof and intrinsically safe equipment 48/12
  48.8.1 Flameproof transformers 48/13
  48.8.2 Flameproof switchgear 48/14

48.9 Gate-end boxes 48/14
  48.9.1 Single-point sensitive earth leakage 48/16
  48.9.2 Multipoint sensitive earth leakage 48/17

48.10 Flameproof motors 48/17

48.11 Cables, couplers, plugs and sockets 48/18

48.12 Drilling machines 48/20

48.13 Underground lighting 48/20

48.14 Monitoring and control 48/21
  48.14.1 Computer system 48/22
  48.14.2 Underground sensors 48/22

48.1 General

In order to deep-mine coal, the sinking of two vertical shafts or inclined roadways must be established to access the coal seams. Underground roadways are then established to the winning area of the coal-face. There are usually two roadways to establish a ventilation system. One shaft and roadway is used to transport the coal, the other usually to transport men and materials to the coal-face. One shaft and roadway is referred to as the *intake* and the other as the *return* airway, signififying the direction of the ventilating air flow with all precautions taken to separate the air flows to maintain adequate ventilation.

Whenever coal is mined, methane gas is liberated, and the electricity regulations require that the electrical power must be removed from that part of the mine if the methane gas content exceeds 1.25% by volume. The regulations allow an exception to this rule to permit communications and certain safety monitoring equipment to be maintained even in the heavy concentrations of methane: in this case the equipment must be intrinsically safe—that is, the equipment must be tested and it must be certified that, in the event of open sparking in either normal or faulty condition, insufficient energy would be released to ignite the most easily ignitable methane concentration.

The regulations place on the manager the responsibility to state where gas may be present in areas of the mine in sufficient quantity to be a potential hazard. In these nominated areas approved equipment is certified Flameproof (FLP) or Intrinsically Safe (IS) for Group 1 (methane) gases.

It will be appreciated that the ventilation for the dilution of liberated methane is the first safety measure in the use of electricity in mining, with the added precaution of approved apparatus should concentrations of methane occur.

Ventilation of the mine is normally achieved by reducing the pressure at the surface end of the return shaft (upcast) by means of an axial- or radial-flow fan, one fan working and one fan on stand-by. A typical installation would operate at 10–15 in water gauge (2.5 kPa) and 125–250 m^3/s air flow. Since the fan is running continuously, efficiency is of prime importance: consequently, detailed attention is paid to this factor. The prime mover used is a cage or slip-ring induction motor, or a synchronous machine with speed change effected by gear or V-belt drive. More recently, variable-speed machines have been used such as a.c. commutator motor, Kramer, cascade arrangements and the pole-amplitude modulated (PAM) motor.

Where underground workings are extensive it may be necessary to provide booster fans in the underground system, usually powered by cage motors up to 400 kW.

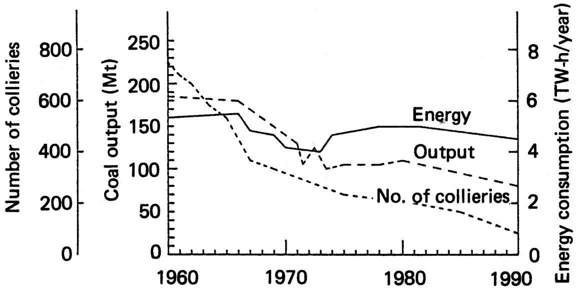


Figure 48.1 Annual coal output and electrical energy consumption (UK)

In order to ventilate single headways, i.e. roadways being driven, auxiliary ventilation fans are used providing air at the end of the tunnel by ducting. These are smaller machines of 10–35 kW and, where they are exhausting, a situation could exist where gas flows over the fan blades; in this condition it is important to ensure that sparking cannot occur at that point. For this purpose vibration monitoring is extensively used. Two frequency ranges are monitored (500 Hz and 5000 Hz) to detect bearing failure and mechanical imbalance. The equipment will first give a warning and later 'trip the power' if the force on the bearing or imbalance exceeds the preset value.

For some years the analogue computer has been used for ventilation calculations for a mine, but is now being superseded by digital computer techniques.

48.1.1 Load growth

All activities at the mine make extensive use of electricity. Steam winders are being eliminated and compressed air as a power medium is almost non-existent. Coal is now almost exclusively won by electrically powered machines, and roadways are driven by large roadheading or tunnelling machines with extensive use of electronics to provide protection, control and monitoring. The trend in electricity consumption over the recent past is shown in *Figure 48.1.*

48.1.2 Regulations

Rules were introduced for electricity in mines in 1905. These became statutory when they were replaced by Part 3 of the General Regulations dated 10 July 1913 (SR&O 1913 No. 748) made under the 1911 Coal Mines Act and relating to Electricity in Mines. In 1954 the duties of mine electrical staff were set out in the Coal Mines (Mechanics and Electricians) General Regulations 1954.

With the advent of the 1954 Mine and Quarries Act, the above were revoked, replaced and their scope increased during the years 1956 to 1965 by new legislation as follows:

The Coal and Other Mines (Electricity) Regulations 1956
The Miscellaneous Mines (Electricity) Regulations 1956
The Coal and Other Mines (Safety Lamps and Lighting) Regulations 1956
The Coal and Other Mines (Mechanics & Electricians) Regulations 1956

Other regulations, although not dealing directly with electricity, had some effect on its use. For example, Regulations relating to Qualifications (MQB Rules), Shot firing, Ventilation, Locomotives, General Duties and Conduct. All of which are printed in *The Law Relating to Safety and Health in Mines and Quarries, Parts 1 to 4.* (Note, the booklet relating to coal mines (Part 2) was reprinted in 1979 in three sections (A, B and C), but account must be taken of any amending legislation since that date, e.g. see below.)

With the arrival of the Health and Safety at Work, etc. (HSW), Act 1974 on 1 January 1975, a programme commenced to replace all existing law made under the M&Q Act 1954 by law made under the HSW Act 1974.

To date (1990), only two principal pieces of legislation have been introduced, i.e. the Mines (Safe Exit) Regulations 1988 and the Electricity at Work Regulations 1989. Both of these are accompanied by Codes of Practice approved under Section 16 of the HSW Act and having special legal status described in Section 17 of the HSW Act.

The latter legislation revoked and replaced the Coal and Other Mines (Electricity) Regulations 1956, the

Miscellaneous Mines (Electricity) Regulations 1956, and amended several sections of other Regulations to update them and eliminate the need for Special Regulations and Exemptions which have been issued over the years to take account of modern practices and techniques.

Work is at present taking place on:

(1) A Mines Administration Package which will revoke and replace the Mechanics and Electricians Regulations and others relating to mine management; and
(2) A Haulage and Transport Package and a Shafts and Winding Package which will revoke and replace the Shafts, Outlets and Roads Regulations.

Note While the nationalised coal industry has changed its name from the National Coal Board to British Coal Corporation, the industry standards continue to be printed as NCB Specifications. The author has therefore used BCC and NCB where appropriate.

48.2 Power supplies

The supply of electricity to collieries is given priority by the Electricity Boards because of the high degree of risk to human life which could arise owing to failure of supply. Winding engines, ventilating fans and pumps are the items of prime importance if men are trapped underground, and the restoration of supply to these items, in particular, must be achieved as quickly as possible.

Before 1939 it was common for collieries to generate their own electricity supply but, with the advent of mechanisation of underground operation, there was a dramatic increase in demand. It was at this period that mine generation was eliminated and supplies taken from the Electricity Boards. Small generators are, however, being introduced, powered by gas engines or turbines burning methane gas extracted from the mine.

The supply usually takes the form of an Electricity Board primary substation located adjacent to the colliery premises into which a duplicate supply is taken at 33, 66 or 132 kV. Alternatively, the colliery may be fed with at least duplicate supplies from the Electricity Board's 11 kV network.

To comply with the requirements of *NCB Mining Department Instruction PI/1957/31*, Electricity Boards are required to provide a minimum of two supplies to a colliery, each supply being taken from a separate point in the Electricity Board's network and routed so as to prevent a complete failure due to a common hazard. Each supply must be capable of handling the full output of the colliery (in the case of a large modern mine, 15–20 MV-A), the system being capable of handling the fluctuating load of electric winders, etc., which could be 3–4 MW.

48.2.1 Distribution

48.2.1.1 Surface

The supply to the colliery substation from the Electricity Board's primary substation (which is usually adjacent to the colliery) is in most cases 11 kV and consists of a minimum of two feeders. Larger collieries may have three or four feeders; a typical distribution layout for a large colliery would be as shown in *Figure 48.2*.

All incoming cables, circuit-breakers, bus-section switches and metering equipment located within the colliery substation remain the property of the Electricity Board, whereas all outgoing circuit-breakers would be the property of the British Coal Corporation (BCC).

Large drives on the surface such as winding engines and ventilating fans are generally supplied direct from the 11 kV switchboard. To ensure continuity, such supplies are usually duplicated in the form of a ring main or duplicate feeders.

The coal preparation plant, the largest energy consumer at the surface of the mine (usually 1.5–2.0 MV-A), is also supplied direct from the 11 kV switchboard by duplicate feeders.

The greatest proportion of load at a colliery is from the induction motors (mainly cage), which leads in some cases to a low power factor on the supply. In the majority of cases this is corrected by manually switched static capacitors on the 11 kV switchboard.

It was the practice to use 3.3 kV as a standard distribution voltage for underground activities. With increase of electrical powered units in operation and increased rating of coalface machines, there may be several 500 kW units being switched direct-on-line at a distance of several miles from the pit bottom; a common distance would be 8 km. The 3.3 kV distribution voltage becomes inadequate and there has been a change to the extensive use of 6.6 kV. While 11 kV is not widely used, at present for underground distribution, flameproof 11 kV switchgear and transformers have been developed and several systems are in operation.

It is envisaged that 11 kV will become the standard underground distribution voltage for systems operating 10 km or more from the source of supply, as in undersea workings.

To provide a 6.6 kV supply for underground distribution, 11/6.6 kV surface installed transformers are used with ratings ranging from 6 to 8 MV-A. Further transformation is provided from the 6.6 kV switchboard to provide lower voltages for other surface auxiliaries such as workshops, stores, stockyards, lamp rooms, offices, pithead baths, lighting, etc.

48.2.2.2 Underground

A minimum of two h.v. supplies are provided to underground workings to increase security, being installed in each of two shafts or drifts. The shaft cables are usually 185 mm^2 three-core polyvinyl chloride (PVC)-double wire armoured (DWA)-PVC and are secured to the shaft wall by large wooden cable cleats spaced at approximately 25 m intervals.

At the shaft bottom, a main h.v. substation is provided from which all supplies radiate to the various districts of the mine. The supply is taken to the coal-face, where it is transformed to the coalface utilisation voltage of 1.1 kV, three-phase 50 Hz. This voltage is proving inadequate for the latest larger coal-winning machines (500 kW), in these cases 3.3 kV is being utilised.

Substations are provided at various points along the cable route to provide a h.v. supply for the main coal conveyors/haulages, etc., or a medium voltage (m.v.) 1.1 kV supply for smaller drives such as secondary conveyors, haulages, pumps, auxiliary ventilation fans, etc. It was the practice to use the utilisation voltage of 550 V for underground activities, and this still remains in some parts of the mine; this, however, proved to be inadequate for the large modern machines, e.g. coal-winning machines (shearers), roadheading machines and armoured face conveyors.

The three-phase system at the mine is earthed to its own earth plates at the surface, and is normally maintained at 2 Ω. Earthing resistors are normally included. The practice was generally to limit the earth faults to the full-load

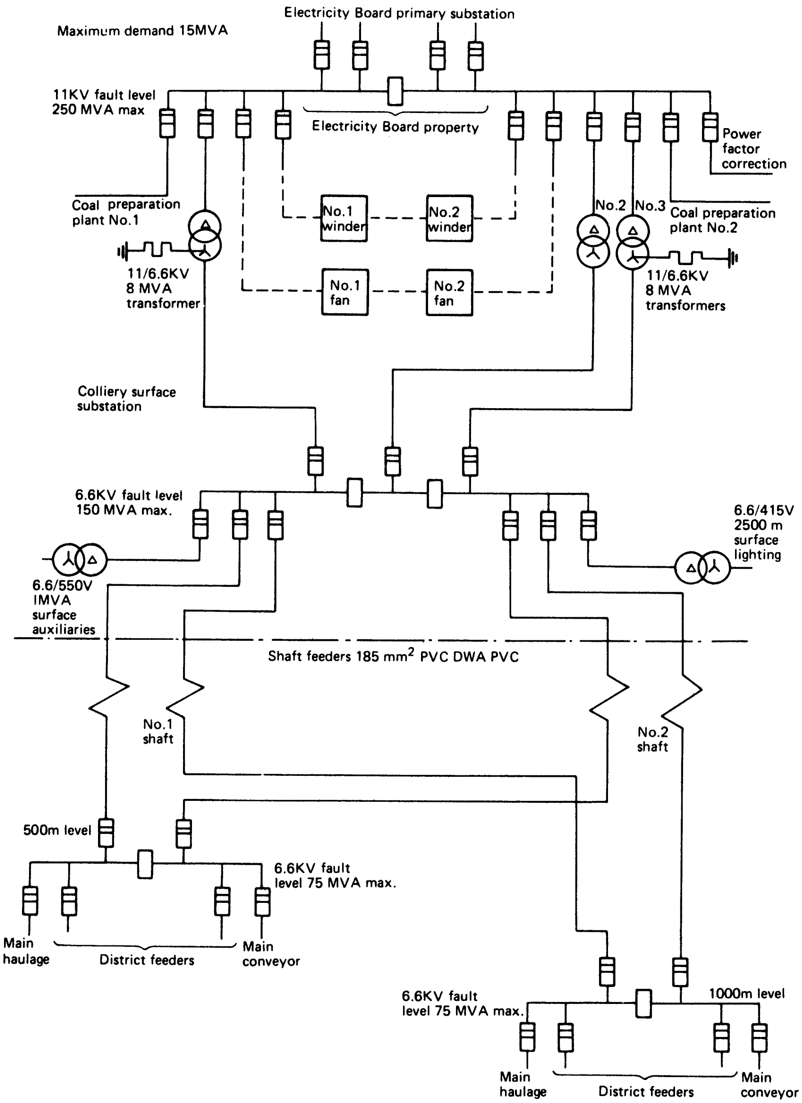

Electricity Board primary substation
Maximum demand 15MVA
11KV fault level
250 MVA max
Electricity Board property
Power
factor
correction
Coal preparation
plant No.1
No.1
winder
No.2
winder
No.2
No.3
Coal preparation
plant No.2
11/6.6KV
8 MVA
transformer
No.1
fan
No.2
fan
11/6.6KV
8 MVA
transformers
Colliery surface
substation
6.6KV fault level
150 MVA max.
6.6/415V
2500 m
surface
lighting
6.6/550V
IMVA
surface
auxiliaries
Shaft feeders 185 mm² PVC DWA PVC
No.1
shaft
No.2
shaft
500m level
6.6KV fault
level 75 MVA max.
Main
haulage
District feeders
Main
conveyor
6.6KV fault
level 75 MVA max.
1000m level
Main
haulage
District feeders
Main
conveyor

Figure 48.2 Typical colliery h.v. distribution

current of the transformer, but modern practice is to limit the 6.6 kV systems to 100 A and the older 3.3 kV systems to 150 A. Standard protection of overload, short-circuit and earth leakage is provided on the outgoing feeders.

Improved short-circuit protection is provided near the coalface, referred to as phase-sensitive short-circuit protection. This phase-sensitive protection was developed in order to permit the through current necessary to start the high-rated (500 kW) machines—which may be five or six times full-load current at a power factor of 0.2—and yet trip the supply on short-circuit of around twice full-load current which is mainly at a power factor of 0.8.

The earth-leakage protection near the face is restricted to a prospective earth fault current on the 1100 V and 550 V systems to 750 mA. Two forms are used, one a restricted core balance, the second being multiple earthing. Extensive use of electronics is made to provide the necessary detection.

48.3 Winders

Early colliery winding engines were powered by steam. Owing to the superior efficiency of electric motors and the greater ease in the provision of automatic control, nearly all winding engines in British mines are now driven by electric motors. These can vary from the very small a.c. winders in the range 100–200 kW, to the large d.c. thyristor automatically controlled winders of 4000 kW. Winding engines exist in several different forms (*Figure 48.3*), and, whereas the majority of the older designs were of the ground-mounted type (with unsightly headgear), new mines generally adopt the tower winder, with its cleaner lines.

Figure 48.4 shows the comparative power-time diagrams for the four different types of winding engines with the same output, net load, depth and decking time (net load, 12 t; depth of shaft, 1000 m; output, 450 t/h). There is little

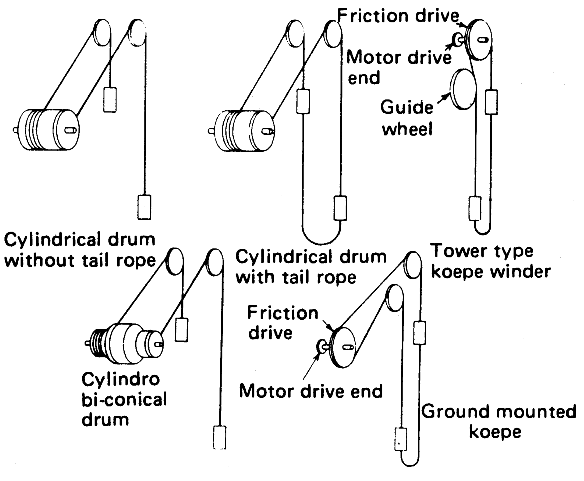


Figure 48.3 Types of winder

difference between the power requirements for a particular duty for ground and tower mounted Koepe winders, but it is obvious that a comparable Koepe winder does not need such a large motor and the energy consumption is less.

Koepe-type winders are specially suitable where extremely heavy loads are to be handled, owing to the fact that a multirope arrangement can be adopted instead of one large single rope. Two or four ropes are generally used, with special devices added to ensure that the ropes share the load equally. With Koepe winders the drive is transmitted from the winder motor by a 'friction' drive; the winding ropes may therefore have a tendency to 'creep' or 'slip'. A 'rope-creep compensating' device is provided, which automatically corrects this situation at the end of a winding cycle if it occurs, bringing the depth indicator and other safety devices into line.

Automatic winding techniques have been developed for modern winding engines, and many winders are arranged to wind coal in the automatic mode. Usually such winders employ skips attached to the winding rope/ropes, instead of cages containing tubs or mine cars. Such skips may hold some 10/12 t and would be loaded automatically from a weigh pocket located at the side of the shaft in the shaft bottom, coal being transferred from the workings to the weigh pocket by conveyor or mine car.

As the skip arrives at the shaft bottom, it is first automatically sensed for being in the correct position and stationary. The weigh pocket door is then operated, and 10/12 t of coal is deposited in the skip in a few seconds. With the skip full and all loading/unloading doors closed, a signal is automatically given to the winding engine to start the wind. The winder is automatically started and accelerated at a predetermined rate to maximum speed. Deceleration commences at a set point in the wind, causing the winder to retard at a predetermined rate to standstill.

With the winder proved stationary and the skip in line, the surface skip door is opened, discharging the coal onto the run-of-mine conveyor at the same time as the shaft-bottom skip is being filled, the whole winding cycle being repeated automatically.

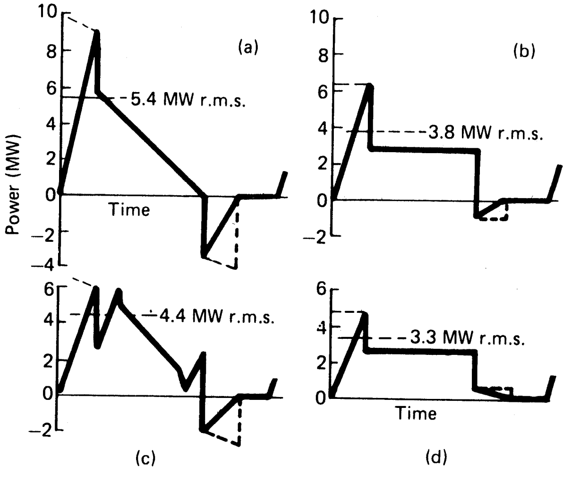


Figure 48.4 Winder power–time diagrams: (a) parallel drum without balance rope; (b) parallel drum with balance rope; (c) bicylindroconical drum

The run-of-mine conveyor transfers the newly won coal to the coal preparation plant, where it is washed, cleaned and loaded into wagons/lorries, etc., for transfer to the customer.

The most common electric winder in the UK mining industry is the a.c. winder employing the slip-ring induction motor with either liquid controller or contactor-operated metallic resistors, with a measure of speed and torque control to limit the acceleration or deceleration. Dynamic braking is used on all but the smallest winders; this is compensated to avoid saturation of the machine and ensure control. A typical layout is shown in *Figure 48.5*.

D.c. Ward–Leonard winders have been used since the turn of the century. The basic layout is shown in *Figure 48.6*. Closed-loop control was introduced to make the machine start from a signal and automatically come to rest at the surface, i.e. acceleration, deceleration, torque, current control, etc.

In the late 1950s, the mercury arc converter replaced the Ward–Leonard generator, but was replaced in the early 1970s by thyristor control. The modern machine is now fully automatic, all control and protection being solid state with thyristors in an antiparallel connection to give complete automatic winding.

48.4 Underground transport

Coal mining has two main transport problems. One is to convey the mineral from the coal-face to the pit bottom for winding to the surface. Conveyors are the main means for transporting the mineral. The second problem is the transport of men and materials to and from coal-face to pit bottom, this is in the main either by rope haulage or by locomotives.

48.4.1 Conveyors

To meet the higher levels of coal-face performance, recent advances in the technology of conventional belt conveyor design and belting have resulted in average conveyor capacities in excess of 2000 t/h. Current underground coal transport systems utilising high-capacity belt conveyor, multi-motor drives and booster drives, together with manless transfer points, remote conveyor control techniques and sufficient automatically controlled bunkerage facilities, provide the most efficient system for the tonnage rates now being produced.

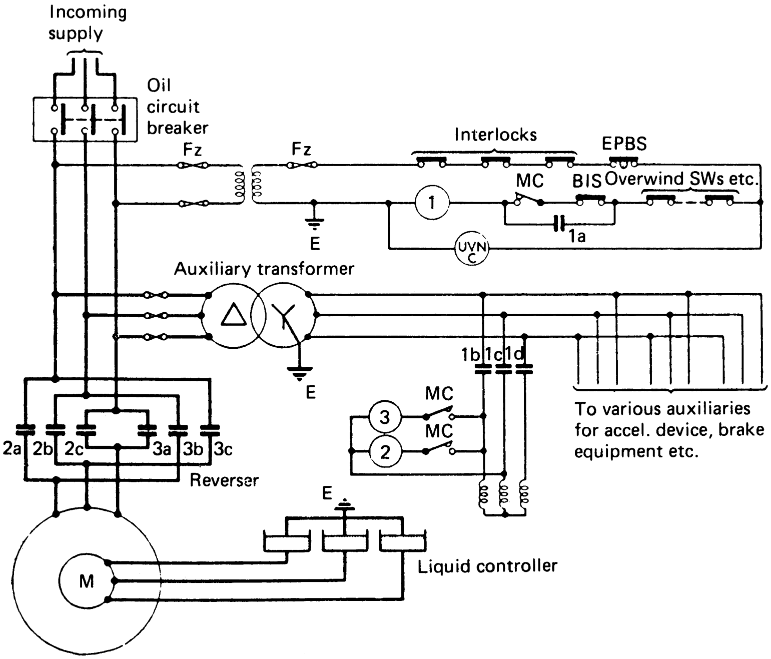


Figure 48.5 A.c. winder control

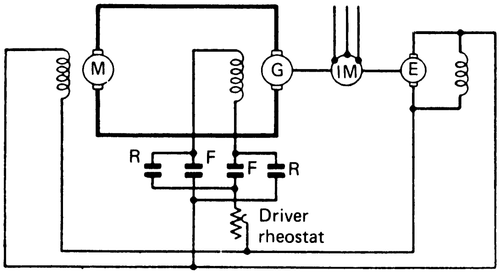


Figure 48.6 Ward–Leonard control

A Code of Practice (NCB Mining Department Instruction PI 1979/5) has been introduced, requiring full compliance to standardise on the protective devices for the safe operation of all underground roadway belt conveyor drive units. For the additional requirements appropriate to manless operation a memorandum of Guidance Minimum Requirements for Man-less Transfer Points on Conveyors was issued. The provisions of the codes are in amplification of the Health and Safety at Work, etc., Act 1974, the Mines and Quarries Act 1954, statutory regulations and mandatory BCC instructions.

The basic type of conveyor drive used underground is the solid mechanically coupled arrangement employing a single flameproof National Coal Board (NCB) specification 542 or 291 motor supplied at either 550 V or 1100 V from a standard flameproof gate-end box starter fitted with a vacuum contactor. The conveyors are started direct-on-line in sequence, and interlocked with the conveyor pre-start alarm system, signals and drive head protection sensors. Local control of starting may be used.

Main trunk roadway conveyors require higher belt speeds, coal-carrying capacities and power. To provide a soft and smooth acceleration to the belt, single- or multi-motor drives with traction-type couplings are used. A limited acceleration period is provided by the fluid coupling. Specially designed fluid couplings with coolers are fitted to keep the starting torque under the fully loaded conveyor conditions to a low level. NCB Specification 625 flameproof motors up to 112 kW rating and flameproof gate-end box starters at 1100 V may be used to supply this type of drive, but a more common arrangement is either 3.3 kV or 6.6 kV flameproof direct-switching vacuum starters supplying higher rated flameproof motors.

The conveyors are started remotely via a telemetry data transmission link from the surface control room. The high-voltage flameproof starters can be fitted with additional auxiliary equipment to provide, for example:

(1) manually controlled electrical loop-winch take-up facilities, used for the higher belt tensions and for accommodation of greater amounts of belt associated with longer conveyors;
(2) integral contactors included in the control gear at 550 V for disc or drum brakes fully interlocked with main drive motors;
(3) forward reverse contactors included in the control gear at 110 V or 550 V for electrically inserted scoop controlled fluid couplings, the arrangement allowing the main motors to be started first, and for separate insertion of scoops to provide drive to the belt; and
(4) electrical controls for acceleration-torque-limit control (ATLC) fluid couplings, which incorporate a separate hydraulic power pack unit for insertion of scoops at a pre-set rate.

Steel-cord belting is used for some drift belt installations and where drive arrangements are similar to those described for main trunk roadway conveyors, utilising ATLC fluid couplings. Microprocessor techniques have been successfully used to control and limit the torque to 150% full-load torque during start-up and acceleration.

At the Prince of Wales colliery in North Yorkshire, the second longest cable belt conveyor in the UK is installed. Powered by a single 2240 kW drive unit, the 1706 m conveyor is the sole underground-to-surface conveyor at the mine. Coal is transported to the surface at a rate of 1000 t/h. The vertical lift is 334 m. The conveyor is designed to operate 24 h per day, with an annual operating time of 4500 h. It is capable of an annual tonnage of 2 million t and the system has the capacity to be extended to a 3000 m length.

The Selby Project (the development of an integrated mine with a capacity of 10 million t/year from five mines) utilises two drift conveyors each capable of dealing with the whole mine's output. The conveyor in the South Tunnel is 14 800 m in length with a total lift of 1000 m. The belting is steel-cord type SC7100, 1.3 m wide, with a rated breaking strength of 9054 kN (923 t). The conveyor drive is a single 2.67 m diameter drum powered by two direct-coupled thyristor controlled BCC type 'E' rationalised winder motors providing an available input to the belt of 10 MW. The speed will be variable up to a maximum of 8.4 m/s (1650 ft/min) and the conveyor is capable of delivering a maximum 3276 t/h at the surface. The conveyor in the North Tunnel is a cable belt conveyor and has duties similar to those of the South drift conveyor. The conveyor is 14 923 m long, with a maximum lift of 1000 m. The drive consists of twin 6.7 m diameter friction wheels driven by two thyristor controlled type 'D' rationalised winder motors via differential and single reduction gears. Available power input to the drive is 8.3 MW.

With the introduction of multiplex data transmission equipment and computer controlled coal clearance systems, a much wider range of FLP and IS sensors is now used to protect the conveyor. The range of sensors includes protection against belt slip, motor overheat, belt misalignment, bearing overheat, blocked chute, torn belt, brake overheat and limit switches for brake and scoop 'on' and 'off' proving.

48.4.1.1 Bunkers

Underground horizontal storage bunkers of the moving-bed and moving-car type with variable speed outfeed metering conveyors are used extensively for capacities up to 500 t. The BCC bunker automation system is now fitted to bunkers of 100 t and over to provide local/remote bunker infeed/outfeed control, using a capacitance probe system, bunker contents and bunker position facilities.

Staple shaft bunkers mounted vertically in the seam incorporating variable speed thyristor controlled outfeed vibrofeeders are used for large coal storage facilities up to 1000 t.

48.4.2 Rope haulage

Underground steel wire rope haulage systems employ vehicles running on (conventional) two rails or, alternatively, on single or double captive rails and a limited use of overhead monorail, for the transport of men and/or materials, with operating speeds of 1.61–32 km/h.

The prime mover of the haulage engine is usually a cage or a slip-ring induction motor with a range of 7.5–375 kW at voltages of 550, 1100 and 3300 V three-phase 50 Hz. The electrical equipment would normally be certified flameproof to Group 1 requirements.

Motors up to around 75 kW are generally started direct-on-line with a 'soft start' feature provided by a fluid coupling of traction or scoop type, although some 10–50 kW designs use a manually operated friction clutch. Also used are electrical devices which control acceleration by automatically increasing frequency and voltage from zero.

The larger machines use slip-ring motors having FLP rotor resistors and drum-type controllers to provide a variable speed drive. Another speed control for motors of 120 kW and above is the cycloconverter, which uses a thyristor converter controlled by a signal to give a varying frequency/voltage output to supply a purpose-designed cage motor.

Generally haulage systems are operated manually from a position local to the haulage engine, with a guard travelling with the vehicle(s) (either riding or walking) to stop and start the system via hardwire transmitted signals presented in audible and visual form to the operator. For higher speed haulages (above 8 km/h) the BCC type 986 Radio System, which operates on the 'leaky feeder' principle for transmission, is used for signalling from the travelling guard to the haulage engine operator.

For small ratings only, transporting haulages are operated by a man walking with the vehicles, from frequently spaced key operated hardwired connected switches, controlling forward or reverse and brake operating contactors. Because of the unattended (i.e. remote operated) haulage, the system must comply with requirements additional to those of conventionally operated haulages to obtain exemption from mining legislation, which normally requires an operator to be in attendance at the haulage engine.

The BCC type 986 Radio Communication System has been extended to provide control, by a travelling guard, of the speed and direction of the vehicles, in conjunction with the cycloconverter drive. The control transmission uses a coded address binary-function digital signal to switch specific function relays of the haulage drive control system.

48.4.3 Locomotives

In most modern mines the coal is transported from the coal-face to the shaft bottom by belt conveyor. There are, however, some mines where the coal is conveyed from the coal-face to an inbye loading point, where it is loaded into 4–5 t mine cars.

From the inbye loading point a train of mine cars will be hauled to the shaft bottom by either diesel, battery or trolley locomotives. These are currently 75 kW diesel locomotives and battery locomotives ranging in size from 6 t, 22.5 kW up to 40 t, 70 kW.

The larger battery locomotives are powered by a 100-cell 550-A-h lead–acid battery with a nominal voltage of 200 V. The battery is contained in a large robust ventilated steel container located in the centre of the double-ended locomotive (*Figure 48.7*), the battery weight being about 4 t.

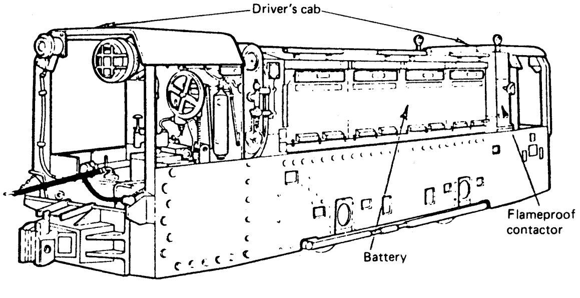


Figure 48.7 Battery locomotive

When a battery change is required, the locomotive enters the battery-charging station and positions itself between a pair of charging racks. The batteries are mechanically connected by the use of specially designed links and a racking device is set in motion. The discharged battery is racked onto the empty rack and the charged battery onto the locomotive, the whole process being completed in a few minutes.

Because the large lead–acid batteries on underground locomotives give off large quantities of hydrogen during the charging process, special requirements are laid down (Locomotive Regulations) governing the design of battery-charging stations. One of the principal requirements is that charging apparatus must be located on the intake side of the battery-charging racks and that ventilation air, having passed over the batteries, is directed into a return airway and does not subsequently ventilate a working face.

Storage battery locomotives for general use underground in coal mines are required by the terms of the Locomotive Regulations to be of a type 'approved by the Authority'. 'Health and Safety Executive—Testing Memorandum TM12' details the test and approval requirements. All electrical equipment used on storage battery locomotives, with the exception of the battery, is required to be certified flameproof.

Most locomotives have two driving motors, one on each of two sets of driving wheels. Series motors are employed with the armatures directly coupled to the driving wheels. Each motor is equipped with a bank of grid resistors controlled through a flameproof contactor and a speed controller. Each driver's cab is equipped with a speed controller, the two electrically interlocked to ensure that only one controller is in operation at any one time.

New locomotives incorporate thyristor chopper control, and methods are being devised to convert some of the existing locomotives to this form of control.

Most battery locomotives in mines are fitted with a battery leakage monitoring device, which consists of an electronic detection unit connected to the battery and an audiovisual alarm unit mounted in the driver's cab. The detector unit has a selector switch which allows the sensitivity at which the unit operates to be varied, a feature necessary to take account of the varying conditions under which the locomotives operate.

Four settings are available, giving battery leakage resistance values of 0.8, 1.3 and 2.6 kΩ. Operation of the alarm indicates that a fault has developed on one pole of the battery and that remedial action should be taken before a second fault develops on the other pole and sets up dangerous circulating currents.

Battery fires underground in coal mines are particularly dangerous and extremely difficult to extinguish once initiated.

Within the UK, trolley-wire locomotive installations have been tried in the past 30 years, but the system has not found universal acclaim, owing principally to the fact that the operating area of the locomotive is restricted to that covered by the trolley wire.

No statutory regulations exist covering the use of trolley locomotives underground in coal mines; consequently, when such installations are considered, special regulations are drawn up to suit each installation.

As coal-faces recede further from the shaft bottom and the need for increased efficiency demands quicker transport of men, materials and minerals to and from the coal-face, the advantages of trolley locomotives in certain circumstances have caused the BCC to look at further trolley installations. One such installation is currently operating on a single overhead 500 V d.c. conductor rail-return system, at a distance of 7 km from the shaft bottom, subsequently extending to 10 km.

Work is currently being undertaken on the design and development of a trolley/battery locomotive which will have complete shaft-bottom to coal-face capabilities, a distinct advantage over trolley systems. It is envisaged that the new trolley/battery locomotives will be rated at 20 t/120 kW, consisting of four 30 kW motors, and incorporating a 250 V, 500 A-h lead–acid battery capable of giving a 75 kW output over a 5-h period. The locomotives will operate on a twin 250 V overhead conductor system, and battery charging will take place while the locomotive is drawing power from the trolley wire. On reaching the end of the trolley wire the locomotive will change to battery power and will carry out excursions away from the trolley area.

The speed at which trolley locomotives operate is 15–20 km/h, compared with the 10–12 km/h of conventional diesel or battery locomotives. With improvements in roadway conditions and better standards of track, speeds of 30 km/h can be expected.

48.5 Coal-face layout

When a new coal-face is started, two roadways are driven from the main intake and return roadways, to form the intake (main gate) and return (return gate) roadways for the coal-face. In each roadway a 6.6 kV DWA PVC three-core aluminium 185 mm^2 roadway cable is installed. This supply is obtained from a local substation which, in turn, obtains its supply from the pit-bottom substation via the parallel district feeders. The main and return gate roadway cables are supported on special hangers attached to the roadway arches at approximately 2 m spacing. These are adequate to support the cable but, in the event of a roof fall, the additional weight causes the cable supports to give way and allows the cable to fall to the floor.

Each roadway cable terminates into a 6.6 kV flameproof 400 A, 150 MV-A circuit-breaker incorporating overcurrent, earth leakage and short-circuit protection. This breaker is a semi-permanent unit, being moved up periodically by the insertion of 100 m of roadway cable as the coal-face advances (*Figure 48.8*).

Advancement of the h.v. circuit-breaker and armoured roadway cable is always carried out with all power isolated, as opposed to all other equipment (flexible wire armoured cables, transformers, contactors, etc.), which is advanced automatically by hydraulic power as the coal-face advances.

From the h.v. circuit-breaker a 6.6 kV, 50 mm^2 flexible pliable wire armoured (PWA) cable takes the h.v. supply to a flameproof transformer. The PWA cable is supported in loops from a monorail attached to the crown of the roadway arches. Special cable supports with rollers permit automatic advancement.

Flameproof air-cooled transformers are used underground. Common ratings in use are 500, 750 and 1000 kV-A; they weigh about 5 t.

To permit automatic advancement, the transformer, hydrostatic power pack, flameproof contactors (gate-end boxes) and face signal/communication unit, along with spares container, oil drums, stretchers, first aid and fire fighting equipment, etc., are mounted on a robust rail-mounted pantechnicon which straddles the main roadway conveyor. The pantechnicon is securely attached to the stage loader conveyor, which, in turn, is attached to the coal-face armoured flexible conveyor (AFC).

As the coal is cut by the power loaders, the AFC is pushed forward by hydraulic rams attached to the hydraulic roof supports. This action causes the stage loader (AFC) to advance forward, which, in turn, automatically advances

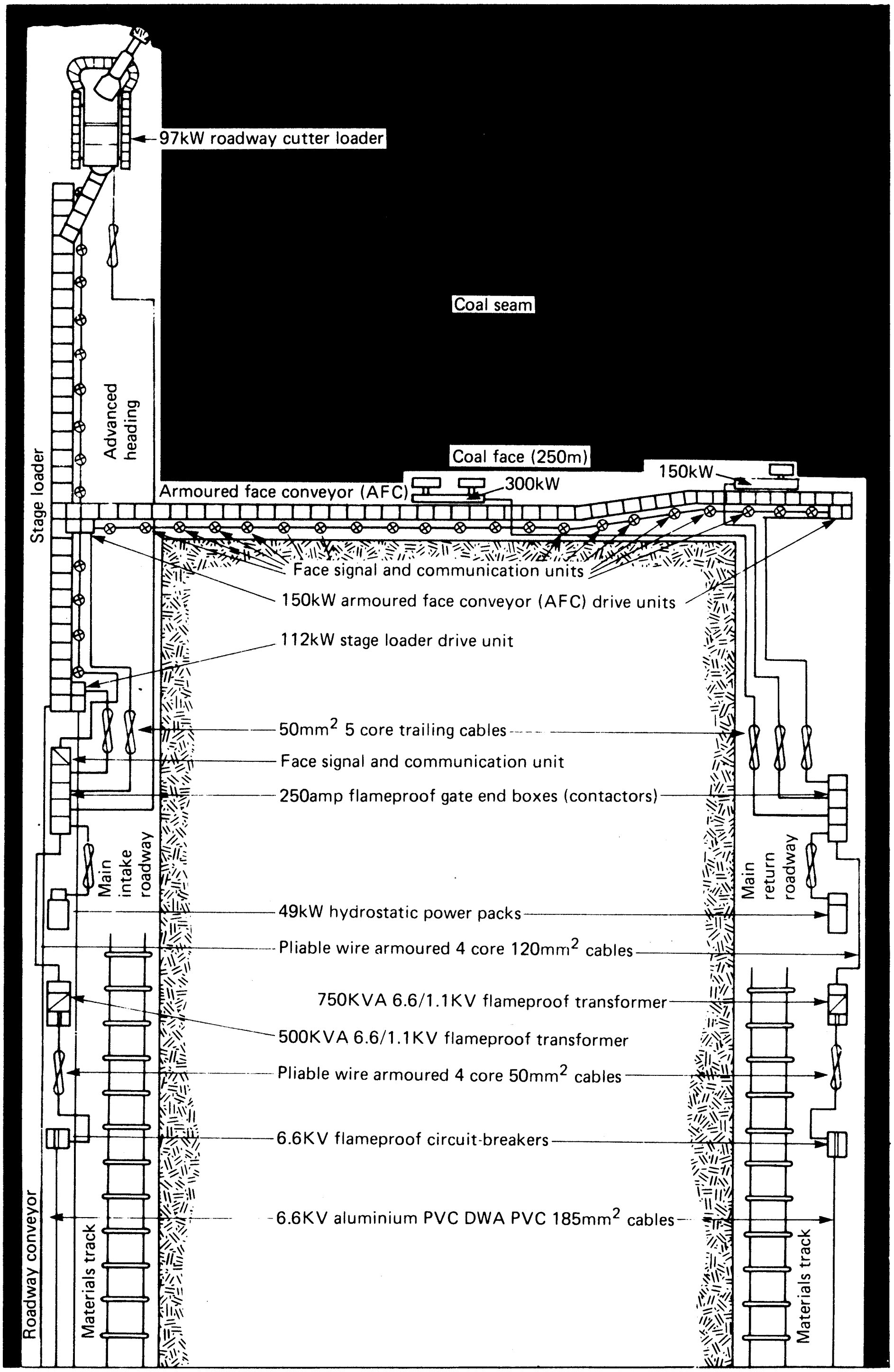

97kW roadway cutter loader
Coal seam
Advanced heading
Stage loader
Coal face (250m)
150kW
300kW
Armoured face conveyor (AFC)
Face signal and communication units
150kW armoured face conveyor (AFC) drive units
112kW stage loader drive unit
$50mm^2$ 5 core trailing cables
Face signal and communication unit
250amp flameproof gate end boxes (contactors)
Main intake roadway
Main return roadway
49kW hydrostatic power packs
Pliable wire armoured 4 core $120mm^2$ cables
750KVA 6.6/1.1KV flameproof transformer
500KVA 6.6/1.1KV flameproof transformer
Pliable wire armoured 4 core $50mm^2$ cables
6.6KV flameproof circuit-breakers
6.6KV aluminium PVC DWA PVC $185mm^2$ cables
Roadway conveyor
Materials track
Materials track

Figure 48.8 Layout of a typical 250 m coal-face

the pantechnicon. When the h.v. PWA cable has been fully extended, power is isolated, allowing the h.v. circuit-breaker and PWA cable to be moved forward, and a new length of 185 mm^2 aluminium PVC–DWA–PVC roadway cable to be installed.

Armoured roadway cables are installed in 100 or 200 m lengths and are sent into the mine already fitted with 300 or 400 A flameproof cable couplers, which are normally connected on site with copper connecting pins, rubber gasket, bolts and nuts.

PWA 120 mm^2 four-core cable is used to take the m.v. 1100 V supply from the transformer to the bank of flameproof contacters (known in the industry as gate-end boxes). Each gate-end box (Section 48.9) is equipped with a 200 A flameproof restrained plug and socket which permits supply to the machine via a 50 mm^2 five-core trailing cable. These cables pass along the side of the stage loader to the machines on the coal-face. The static parts of power-loader trailing cables are located in specially designed cable troughs attached to the AFC, whereas the part of the trailing cable which flexes backwards and forwards as the power loader moves along the coal-face is contained in a robust flexible steel or plastic cable handler. Such handlers also contain a water hose which supplies the machine with water for dust suppression and motor cooling. A typical coal-face could be established as in *Figure 48.8*.

The right-hand single-ended ranging drum shearer (SERDS) cuts the right-hand end of the face to a distance of about 25–30 m, while the main machine, the double-ended ranging drum shearer (DERDS) cuts the rest of the face.

Attached to the bank of gate-end boxes in the main gate is a flameproof and intrinsically safe face signal and communication unit. Connected to this unit and spaced approximately every 7 m along the stage loader and coal-face conveyors are face signal and communication units. Each unit is equipped with a signal push-button and lock-out stop key, and every third or fourth unit incorporates a loudspeaker and microphone.

A control point attendant at the face signal and communication unit position controls the stage loader and face conveyor in response to signals transmitted by any of the signal pushbuttons. Upon operation of the start button for the stage loader (which, in turn, automatically starts the face conveyor) a seven-second pre-start warning two-tone ‘bleep’ is transmitted along the whole length of the stage loader and face conveyor, warning faceworkers that the conveyors are about to start.

Operation of a lock-out push-button causes the respective conveyor to stop. Should this conveyor be the stage loader, the face conveyor will also stop, as they are connected in sequence.

Should the lock-out push-button be latched in the lock-out position, the conveyor cannot be started until that push-button has been reset. Each lock-out push-button has a specific number which is automatically displayed by a digital readout on the face signal and communication unit whenever a particular lock-out is operated. By the use of such communication facilities, the cause, necessary remedial action and subsequent duration of a stoppage in production can be quickly ascertained.

The face signal and communication unit is connected by cable to the main colliery control room on the surface, which permits instant and direct communication between the surface control room and any point along the working face, or vice versa.

Each power loader is controlled by an individual operator, sometimes by using radio control. Before the machine can be started, water must be turned on to the pre-start warning water jets positioned on either side of the cutting drum. This condition must persist for approximately 7 s before power can be switched on to the machine, thereby warning by wetting anyone inadvertently in a dangerous position that the machine is about to start.

48.6 Power loaders

The modern coal-getting machine is termed a ‘power loader’ because it not only cuts coal, but also loads it onto the armoured flexible conveyor (AFC), which is, in effect, a steel scraper conveyor running the full length of the working face. *Figure 48.9* illustrates a typical modern power loader which has a rotating cutting disc at each end of the machine, mounted on a ranging arm to cater for thicker seams of coal which could be 2–3 m or more. This type of machine is known as a double-ended ranging drum shearer (DERDS). Certain methods of mining call for the use of a machine with only one cutting disc. Such a machine, similar to that illustrated, is termed a single-ended ranging drum shearer (SERDS).

The majority of power loaders are driven by a single 150 kW motor, operating at 1100 V. Some larger machines, however, have been developed using a single 300 kW motor or a two 300 kW motor arrangement. The supply is obtained from a gate-end box in the roadway via a flexible trailing cable, which on the coal-face is enclosed in a robust flexible cable handling device for protection.

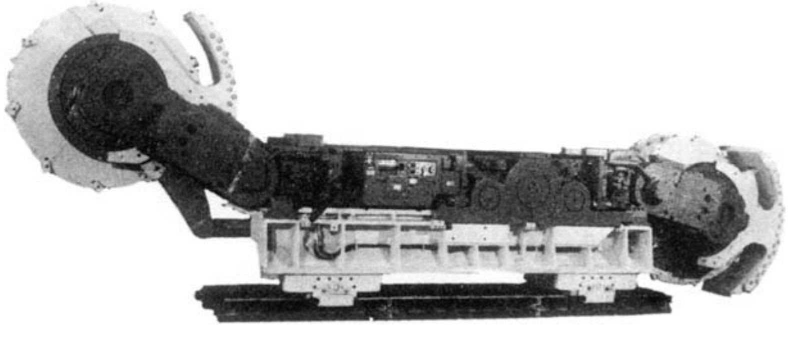


Figure 48.9 Power loader

Each complete machine is built up of a number of sections, consisting basically of electric motor, haulage unit and gearhead.

Incorporated in the motor are a reversing isolator, control facilities and fault diagnostic equipment, which are connected to the flexible cable by a flameproof plug and socket. A drive shaft protrudes from each end of the motor to transmit power to the adjacent units.

The motor at one end is attached to the haulage unit, which provides the hydraulic power, to haul the machine along the coal-face via a driven pinion on the machine, and a static rack attached to the AFC. Speed control is automatic, following the limits determined by the operator who travels with the machine.

Attached to the opposite end of the motor (single-ended machine) is a gearbox, which drives the rotating coal cutting disc. In a double-ended machine a similar gearbox is attached to the haulage unit at the opposite end of the machine.

Radio control techniques have been developed for power loaders, which allow the operator to control the machine from a comparatively safe area, some 15–20 m from the machine.

48.7 Heading machines

One of the principal requirements of modern mining is the ability to drive roadways quickly and safely. This is achieved in the main by the use of heavy-duty roadway cutter loaders (*Figure 48.10*). These machines are equipped with one or two rotary cutting heads, which cut out the stone and shape the profile of the roadway to 4 m × 3.5 m.

Debris from the cutting heads falls onto a rotary scraper conveyor at the front of the machine, which transfers the cut material to a bridge conveyor at the rear of the machine, and then on to the roadway conveyors (*Figure 48.8*).

Roadway machines are fed by flexible trailing cables at 1100 V and typically have a 50 kW motor driving a hydraulic power pack and a 50 kW motor driving each cutting head. The machine traverses on hydraulically powered caterpillar tracks, the whole machine being controlled by an operator sitting in the middle of the machine.

48.8 Flameproof and intrinsically safe equipment

During the process of extracting coal from the seam, methane gas is given off. It combines with the normal mine air flow and is eventually disposed of at the surface of the mine through the mine ventilation system. By this means the methane content in the mine air is kept to low and safe proportions.

Owing to possible malfunctioning of ventilation apparatus, power supply failure or heavy emissions of methane, the methane/air ratio can increase. Between approximately 5 and 15% methane to air, the mixture becomes explosive, the most explosive mixture being 8.3% methane to air. Electrical equipment in which sparking during normal operating may occur is capable of igniting an explosive methane/air mixture and must, therefore, be given special consideration. For equipment operating at a low voltage and current levels, the circuits can be designed such that the energy released at the spark is insufficient to cause ignition. This can be achieved by the use of non-inductive resistors, non-linear resistors, capacitors, shunt diodes, Zener diodes, full-wave rectifiers, etc., to give ‘intrinsically safe’ (IS) apparatus.

Apparatus classed as ‘intrinsically safe’ includes telephones, signals, communications, testing instruments, methanometers, sensors and remote control and monitoring. Since any open sparking produced within such equipment is incapable of igniting an explosive methane/air mixture, no other form of protection is required other than to house the components in a robust enclosure.

In the UK, it has been the practice to design IS apparatus to conform to the requirements of BS 1259 (IS Electrical Apparatus and Circuits). Equipment now, however, is designed to the CENELEC Standard, now BS 5501: Part 7.

IS equipment must be certified by an approved certifying authority for use in mines and the IS certificate number along with other specified information must be clearly marked on each item.

One of the principal requirements of the intrinsically safe certificate is that the equipment must be supplied from an approved source of supply, which can be either a.c. or d.c. The current British Standard covering such supplies is BS

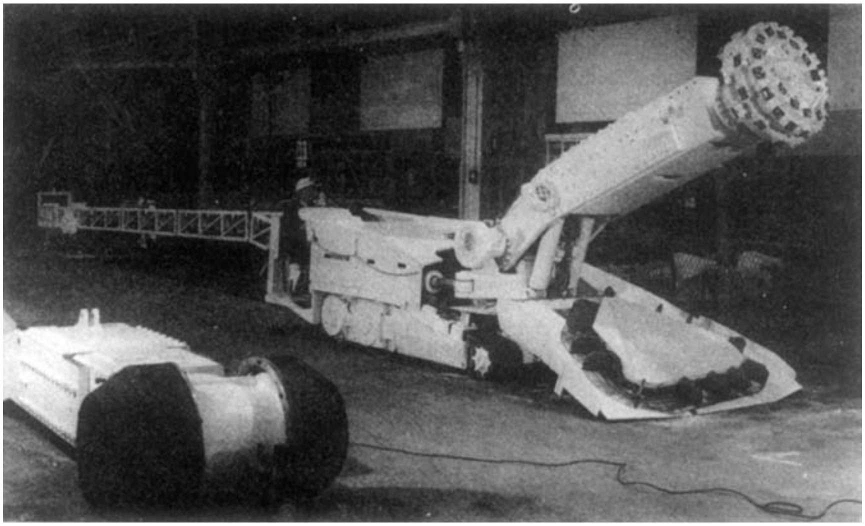


Figure 48.10 Boom miner with bridge conveyer

6182 (Intrinsically Safe Power Supplies) and caters for a.c. and d.c. supplies and rechargeable battery units.

There are three categories of d.c. supplies, i.e. 7.5, 12 and 18 V, and two of a.c. supply, i.e. 12 and 15 V. Rechargeable batteries of 8 and 14 V are specified, each capable of being charged from the respective source of supply, i.e. the 8 V battery from the 12 V d.c. supply and the 14 V battery from the 18 V d.c. or 15 V a.c. supply.

Power equipment operating on higher current levels which produce sparks during normal operation, and which cannot be designed to be intrinsically safe, must be enclosed in a robust enclosure. If an explosive methane/air mixture exists in a mine roadway and enters the apparatus containing spark producing components, the flame resulting from the ignition will not be transmitted to the ambient atmosphere and so ignite the general body of the mine air. The enclosure in which spark producing components are housed is termed a 'flameproof enclosure'.

Equipment coming within this category covers such items as motors, contactors, switchgear, transformers, light fittings, plugs and sockets, cable couplers, etc., and requires to be certified flameproof by the approved certifying authority as safe for Group 1 gases (methane). In the past, the relevant British Standards were BS 229 and BS 4683. Equipment now, however, is designed to comply with the CENELEC Standard, now BS 5501: Part 5.

A 'Flameproof Certificate' is issued, the number of which must be permanently displayed on each item of equipment along with other relevant details required by the Standard, i.e. manufacturer's name, type reference, number of British Standard, etc.

The harmonisation of European Standards resulted in the issue in 1977 by CENELEC (European Committee for Electrotechnical Standardisation) of the EN 50 series of standard, i.e. EN 50-014 to EN 50-020 Electrical Apparatus for Potentially Explosive Atmospheres. Equipment designed and certified to this standard is accepted by the European Community without further testing, etc.

The European Standards of 1977 were issued as British Standards: a list is given in *Table 48.1*. Since then, these Standards have been periodically updated and two further Standards have been issued, i.e. CENELEC Standard EN 50028 Encapsulation type 'm', i.e. BS 5501: Part 8, and Cenelec Standard EN 50039 (IS) systems, i.e. BS 5501: Part 9.

Gaps associated with any removable covers, doors, motor shafts, etc., must conform to minimum dimensions laid down by the standard. 'Flameproof' apparatus must be designed, installed and maintained at all times within those limits. As a typical example of the toleranes permitted, gaps with a length of 12.5 mm must not exceed 0.4 mm and for 25 mm flanges the maximum is 0.5 mm. During normal maintenance procedures in the mine these gaps are periodically checked by colliery craftsmen using feeler gauges.

Table 48.1 Standards for electrical apparatus for use in potentially explosive atmospheres

CENELEC Standard, EN 50-	*British Standard BS 5501*	*Subject*
014	Part 1	General requirements
015	Part 2	Oil immersion 'o'
016	Part 3	Pressurised apparatus 'p'
017	Part 4	Power filling 'q'
018	Part 5	Flameproof enclosure 'd'
019	Part 6	Increased safety 'e'
020	Part 7	Intrinsic safety 'i'

48.8.1 Flameproof transformers

Providing sufficient electrical energy at the modern coal-face with a demand of some 1–1.5 MV-A at a distance of 6–8 km from the shaft bottom and at a depth of 900–1000 m requires an efficient h.v. electrical distribution system.

From the voltage regulation point of view it is essential that the h.v. supply be taken up to the working face, therefore transformation facilities are needed to provide the coal-face utilisation voltage of 1100 V.

Before the advent of flameproof dry-type transformers, standard industrial oil-filled transformers were utilised, often with flameproof oil circuit-breakers attached to each end. Mining legislation at that time, however, decreed that such transformers could not be used nearer than 300 yards from the coal-face and in no circumstances could they be used in a return roadway. As coal-face loading increased in the 1950s and early 1960s as a result of the mechanisation of coal-getting, it became of paramount importance to move the transformer right up to the working face and strengthen the h.v. distribution system. This led to the development of the flameproof air-cooled transformer, which has a flameproof h.v. circuit-breaker on the h.v. side and a flameproof m.v. chamber mounted on the opposite end to house the overcurrent, sensitive earth leakage and short-circuit protection equipment, which on operation causes the h.v. circuit-breaker to trip. Typical transformer ratings would be 500, 750 and 1000 kV-A. Comparing these with the old oil-filled types (which were usually of the order of 250 kV-A) indicates the change which has taken place in the mining industry since nationalisation in 1947.

Figure 48.11 shows a typical 750 kV-A flameproof transformer viewed from the h.v. end, the circuit-breaker being a 6.6 kV, 400 A, 150 MV-A sulphur hexafluoride (SF_6) unit complete with incoming cable adapter suitable to accept the 6.6 kV, 300 A, six-bolt cable coupler attached to the end of the incoming 6.6 kV h.v. cable. The m.v. chamber at the opposite end is approximately three-quarters the size of the h.v. circuit-breaker and is similarly equipped with flameproof adapters to accommodate the outgoing 1100 V cables.

The transformer is equipped with lifting lugs for loading and unloading and adjustable wheels for transportation underground. The total weight would be of the order of 5–$5\frac{1}{2}$ t.

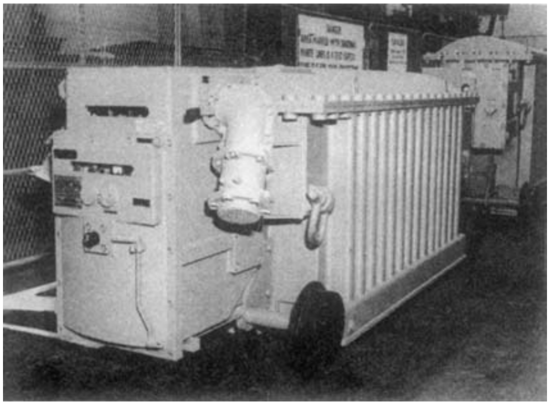


Figure 48.11 Flameproof transformer

48.8.2 Flameproof switchgear

Prior to the mid-1960s, nearly all the h.v. and m.v. circuit-breakers used in mines were of the oil-break type, usually of the order of 150/200 A, 3.3 kV, 25 MV-A and, although certified flameproof, they constituted a hazard owing to the oil-fire risk. In addition, maintenance of the oil was a problem due to the oil transport and cleanliness, and the disposing of waste oil. With the introduction of no-oil switchgear, maintenance was reduced and the oil-fire risk eliminated.

Modern flameproof mining-type circuit-breakers operate on either the air-break, vacuum interrupter or sulphur hexafluoride gas principle and can be arranged such that they can be utilised as single free-standing units, on a complete switchboard, or mounted on the h.v. end of a flameproof transformer.

Owing to the increase in demand for electrical power underground and the uprating of underground distribution systems, switchgear ratings have also increased. A typical modern flameproof circuit-breaker as shown in *Figure 48.12* would be a 6.6 kV, 400 A, 150 MV-A unit. This circuit-breaker is of the vacuum interrupter type and the illustration shown is a classical example of the construction of flameproof switchgear. The flameproof enclosure is divided into separate flameproof compartments electrically linked by the use of flameproof bushed terminals. In the bottom compartment the circuit-breaker is housed on a withdrawable chassis complete with overcurrent, earth leakage and short-circuit protection.

Two separate compartments are provided in the centre of the circuit-breaker which accommodate the isolator(s) and incoming or outgoing cable terminations. Also incorporated in the section are the isolator and circuit-breaker operating handles, which are mechanically interlocked to ensure that the isolator can be operated only with the circuit-breaker in the 'off' position. Mechanical interlock is also provided between the isolator and circuit-breaker handles and the front cover of the circuit-breaker compartment, to prevent access until the interior has been made dead.

With slight modification to the basic design, flameproof mining-type circuit-breakers can be used in either of four different modes, i.e. incoming unit, feeder unit, bus-section unit or transformer switch unit.

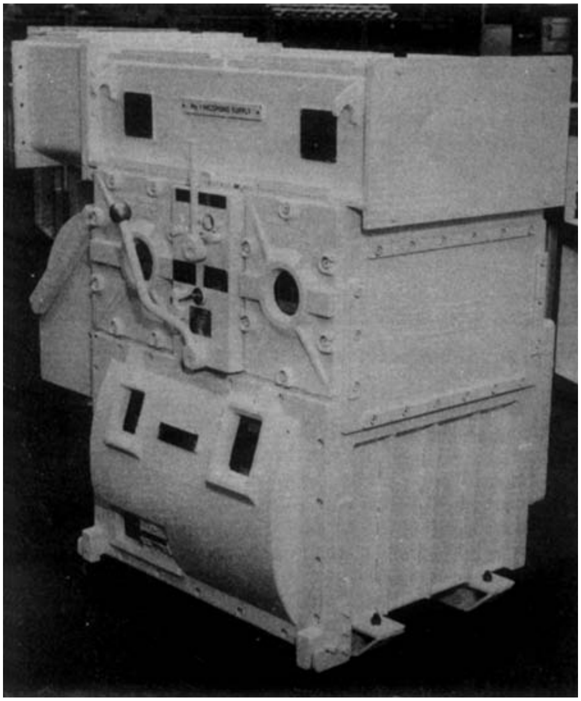
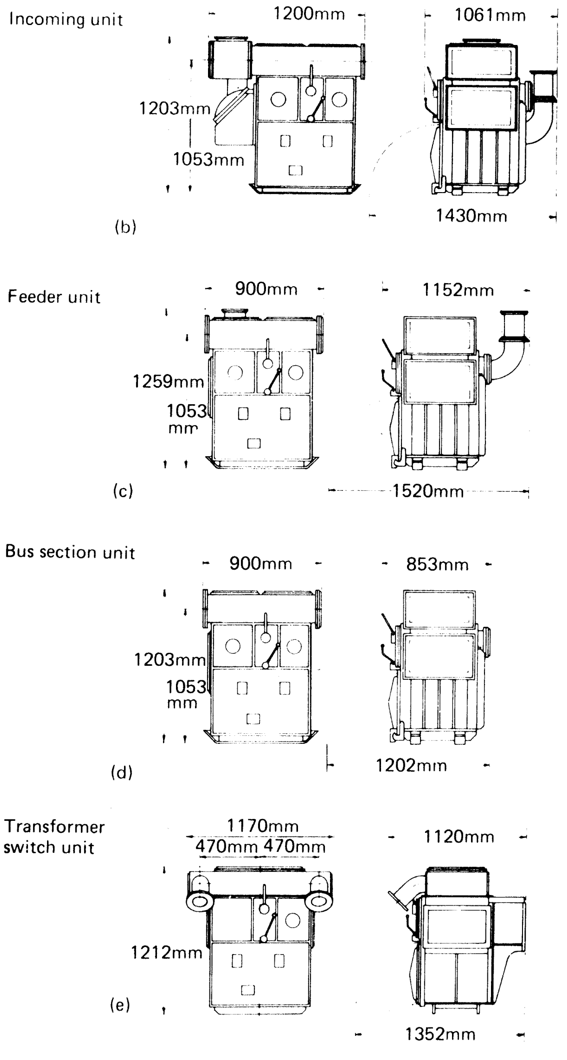


Figure 48.12 Flameproof circuit-breaker

48.9 Gate-end boxes

The control of individual drives in a coal mine, such as conveyors, power loaders, pumps, haulages, etc., is achieved as in any other industry, i.e. by use of contactors.

For mining purposes contactors must be enclosed in a robust flameproof enclosure or box. In the early days of electricity in mines, such a box would be installed at the end of the roadway leading to the coal-face. To use mining parlance, a roadway is a 'gate'; therefore, the 'box' installed at the 'gate end' became known as the 'gate-end box'.



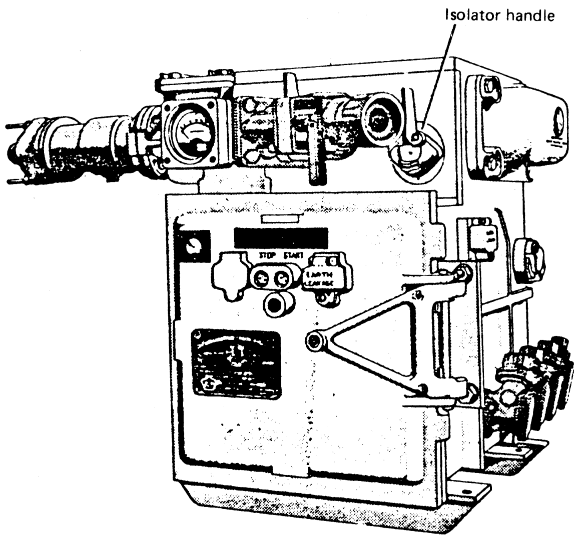


Figure 48.13 Gate-end box

The term is still used today to refer to a flameproof contactor unit.

Figure 48.13 illustrates a typical flameproof gate-end box suitable for use on an 1100 V, three-phase, 50 Hz system, and rated at 300 A. The box can be adapted for use as a single unit or assembled to form a 'bank of panels'. *Figure 48.14* shows a typical bank of panels mounted on a pantechnicon over a roadway conveyor adjacent to a coal-face.

A gate-end box consists of upper and lower chambers interconnected by flameproof bushed terminals. The upper compartment contains 400 A throughgoing bus-bars and a three-phase isolator. The ends of the bus-bar chamber are designed to accept flameproof bus-bar trunking units and links to enable individual panels to be built up into a bank or to accept a bus-bar blank at one end and an incoming cable adapter at the other to form a single unit. In the lower compartment is housed the contactor of rating 150, 250 or 300 A, and a control unit, all mounted on a removable chassis for ease of maintenance, repair or removal.

A mechanical interlock is provided between the isolator operating handle and the contactor compartment cover to ensure that access to the contactor compartment is prevented until the isolator has been placed in the 'off' position.

Power is transferred to the drive from the gate-end box via a flexible trailing cable which, owing to the hazardous environment of the coal-face, is always susceptible to damage. Facilities must therefore be provided to enable the trailing cable to be changed quickly. This facility is provided on the gate-end box and the motor in the form of a 200 A restrained flameproof plug and socket (*Figure 48.15*).

Taking into consideration the control facilities, sensitive earth leakage protection, static overcurrent protection equipment, etc., the modern gate-end box is a complex piece of electrical equipment which has to work with a high degree of reliability in an atmosphere that can be hot, cold, dry, wet and, on occasion, subject to considerable vibration. Some of the electrical/electronic components can therefore be subject to abnormal abuse while in service. It is for this reason that the evolution of the modern gate-end box has resulted in the majority of the small electrical/electronic components being contained in a 'control unit', which is a plug-in unit on the contactor chassis. This control unit can be quickly changed in the event of trouble, and transported to the surface workshop for overhaul and repair.

Control of the gate-end box can be effected locally or remotely, the selection being by way of a changeover switch on the control unit. In the majority of cases the remote control facility is adopted: it utilises a pilot control core in the five-core flexible trailing cable and a flameproof starting device at the motor end. In the past, the pilot circuit or the remote control IS circuit was designed to the NCB specification P130, based on the following principles:

1. the circuit is energised from an intrinsically safe constant voltage transformer within the gate-end box, designed to give a constant 12 or 7.5 V secondary output over a wide variation of primary input—one side of the 12 or 7.5 V winding is earthed;
2. a pilot relay is provided in the gate-end box which will operate on half-wave but not on full-wave a.c.; and
3. at the far end of the trailing cable a diode is provided along with a start switch across which is connected a 30 Ω resistor.

With the upgrading of the specification for intrinsically safe (IS) equipment it has become difficult to design this

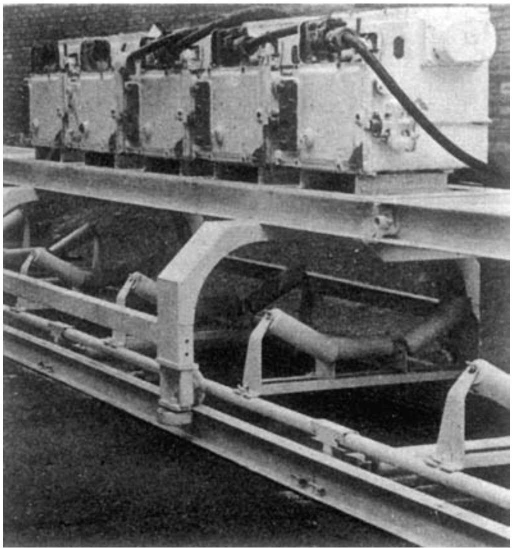


Figure 48.14 Bank of gate-end boxes

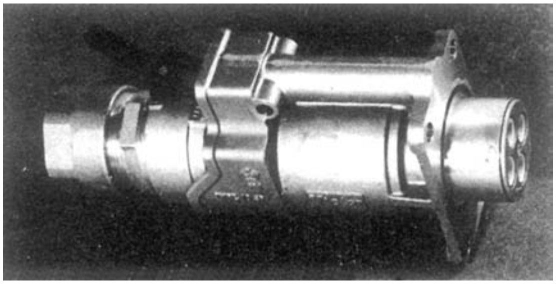


Figure 48.15 Flameproof plug and socket

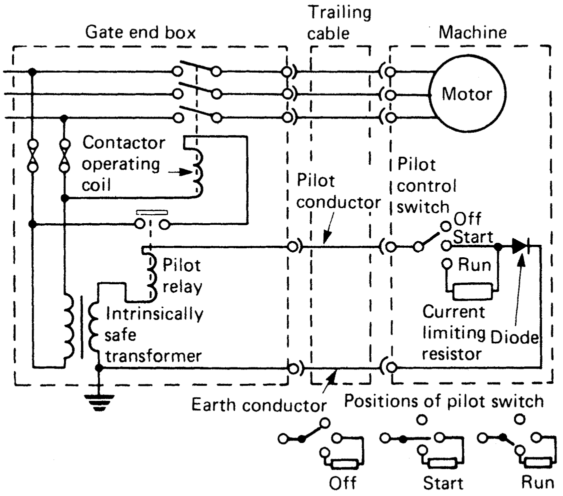

Trailing cable
Gate end box
Machine
Motor
Contactor operating coil
Pilot conductor
Pilot control switch
Off
Start
Run
Pilot relay
Intrinsically safe transformer
Current limiting resistor
Diode
Earth conductor
Positions of pilot switch
Off
Start
Run

Figure 48.16 Contactor-coil/pilot circuit

pilot circuit to the latest IS requirements. A new British Standard (BS 7202: 1989 Non Incendive Low Voltage Control/Interlock and Low Voltage Earth Fault Monitoring Circuits for Use in Mines) has been issued to which the equipment is now being designed.

Figure 48.16 shows a basic arrangement of the contactor coil/pilot circuit within a gate-end box, the associated trailing cable and a face machine with in-built start switch. To start the machine, the switch is moved to the 'start' position. With only the diode in circuit the pilot relay (PR) energises and the machine starts. The start switch now reverts automatically to the 'run' position inserting the 30 Ω resistor into circuit.

Should a power failure occur with the control switch in the 'run' position, following which the power is restored, the relay PR will not energise at a voltage below 120% of the declared voltage of the incoming supply. Operation of the start button causes PR to energise: it must do so down to 75% of the declared incoming voltage supply. Once energised and with the 30 Ω resistor in circuit, PR must remain energised down to 60% of the declared voltage but under no circumstances continue to operate at 20% or below.

Should a damaged trailing cable result in a pilot core-to-earth fault, full-wave a.c. would be applied to PR which (on account of the increased impedance) would de-energise if in operation, or fail to energise upon operation of the start button. This condition, known as pilot core protection (PCP), would be indicated by a lamp at the front of the gate-end box.

Prior to 1959 most transformers used underground to supply coal-face equipment had the neutral point of the secondary windings solidly earthed, and earth leakage protection in gate-end boxes operated on the core-balance principle. Trailing cable damage on such systems resulted in very severe incendive arcing which, on the coal-face especially, was a very serious hazard. Following an explosion at Walton Colliery near Wakefield in 1959, H.M. Principal Inspector of Mines recommended that a further attempt should be made to devise an electrical protective system capable of eliminating, or at least substantially reducing, the dangers of incendive sparking resulting from damaged trailing cables. This led to the development of the sensitive earth leakage (SEL) circuit, which exists in two forms: single point and multi-point.

48.9.1 Single-point sensitive earth leakage

The basic principles of single-point earthing systems are similar to those of solidly earthed systems in that a core-balance transformer more sensitive than used on solidly earthed systems is employed. This system is sometimes referred to as sensitive core balance, and the main difference between the two systems is in the method of earthing the neutral point of the transformer secondary winding.

In the single-point system an impedance is inserted between the neutral point and earth of such value as to limit the earth fault current to a maximum of 750 mA (*Figure 48.17*). Although this is the maximum earth fault current permitted, individual earth fault trip circuits are set to trip at between 80 and 100 mA, giving a safety factor of approximately 7 to 1.

The core-balance transformer output under fault conditions is very small and an electronic amplifier is used to control the earth fault relay, which is energised under healthy conditions and de-energises on the occurrence of a fault. This arrangement results in a fail-safe system. The earth leakage relay contacts are inserted in the contactor coil circuit, which opens on the occurrence of a fault.

To ensure that a contactor cannot close onto a system on which an earth fault condition exists, an additional arrange-

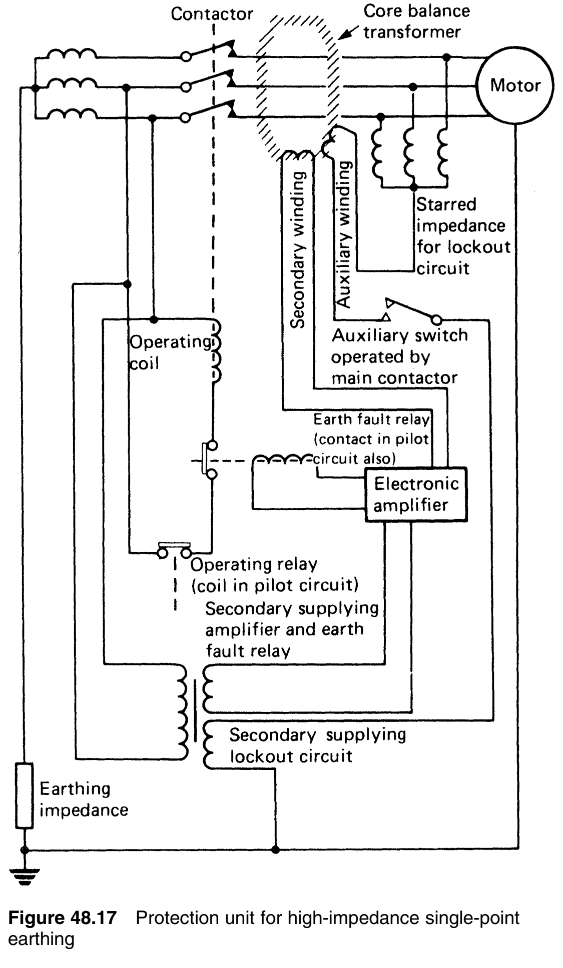

Contactor
Core balance transformer
Motor
Secondary winding
Auxiliary winding
Starred impedance for lockout circuit
Operating coil
Auxiliary switch operated by main contactor
Earth fault relay (contact in pilot circuit also)
Electronic amplifier
Operating relay (coil in pilot circuit)
Secondary supplying amplifier and earth fault relay
Secondary supplying lockout circuit
Earthing impedance

Figure 48.17 Protection unit for high-impedance single-point earthing

ment is provided which is termed the electric lockout circuit. It consists of three impedances, one end of which is connected to the three outgoing phases; the other ends are star-connected and, in turn, are connected to an auxiliary core-balance transformer winding, a pair of contactor auxiliary contacts (closed when the contactor is open) and an intrinsically safe source of supply which has one end of its winding earthed. Should an earth fault develop while the contactor is de-energised, the earth leakage lock-out circuit would operate to prevent the contactor coil energising, a condition that would persist as long as the earth fault was in existence.

48.9.2 Multipoint sensitive earth leakage

In the multipoint system the transformer secondary is completely insulated from the earth, i.e. it is a free neutral (*Figure 48.18*). Each contactor is provided with a false neutral which, similar to the single-point system, consists of three impedances connected to the three outgoing phases. The star point is connected via a pair of contactor auxiliary change-over contacts to either the failsafe earth leakage detection circuit or the earth leakage lock-out circuit, depending on whether the contactor is energised or not.

The earth fault current on the multipoint system is limited to 40 mA on a 1100 V system. Since an earth fault on a system supplied from one transformer could cause all gate-end boxes on the system to trip, and in order to keep the maximum earth fault current to 750 mA, the number of gate-end boxes on a system must be limited to $750/40 = 18$.

When a contactor trips on earth leakage on either system, that contactor locks out and displays earth leakage trip conditions, which can only be re-set by an authorised craftsman with the appropriate specialised equipment.

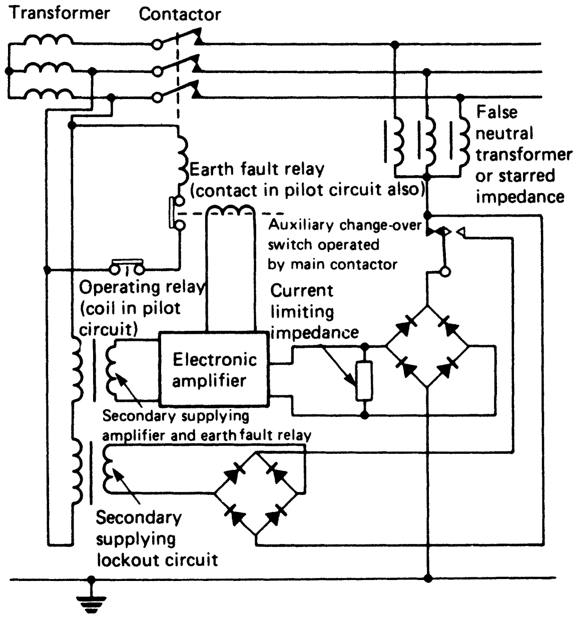

Transformer
Contactor
False neutral transformer or starred impedance
Earth fault relay (contact in pilot circuit also)
Auxiliary change-over switch operated by main contactor
Operating relay (coil in pilot circuit)
Current limiting impedance
Electronic amplifier
Secondary supplying amplifier and earth fault relay
Secondary supplying lockout circuit

Figure 48.18 Protection unit for multipoint earthing

Table 48.2 National Coal Board (NCB) Specifications for flameproof motors

NCB Specification No.	*Motor rating* (kW)	*Mounting*	*Voltage* (kV)
291	37.5–50	Flange	≤1.1
420	67.5–90	Flange	≤1.1
542	2.25–30	Foot/flange	0.11–1.1
625	112	Flange	0.55–1.1
634	112	Flange	2.2–3.3*
635	112	Flange	2.2–3.3†
636	150	Flange	1.1
637	225	Flange	1.1
651	75	Foot	0.55, 1.1, 3.3‡⇐

*Single-stack conductors.
†Other than single-stack conductors.
‡For booster fans.

48.10 Flameproof motors

Mechanisation of the coal-mining industry in the 1950s, followed by further mechanisation and automation in the 1960s and 1970s completely changed the face of the industry. Elimination of 'self-generation' at collieries and the introduction of duplicate and more substantial power supplies from the Electricity Boards plus the strengthening of colliery distribution systems made the direct-on-line starting of large cage-induction motors possible. BCC Specifications have been produced covering the majority of FLP motors used in British mines (see *Table 48.2*).

Motors associated with coal-face equipment need changing more often than those operating in roadways and engine houses, owing to the hazardous conditions in which the equipment has to operate on the coal-face, physical damage and ingress of water moisture being the prime causes of failure. To facilitate speed and accuracy in changing and lining up, such motors are designed for flange mounting. *Figure 48.19* shows a typical flange-mounted motor complete with flameproof terminal box and flameproof 200 A plug and socket.

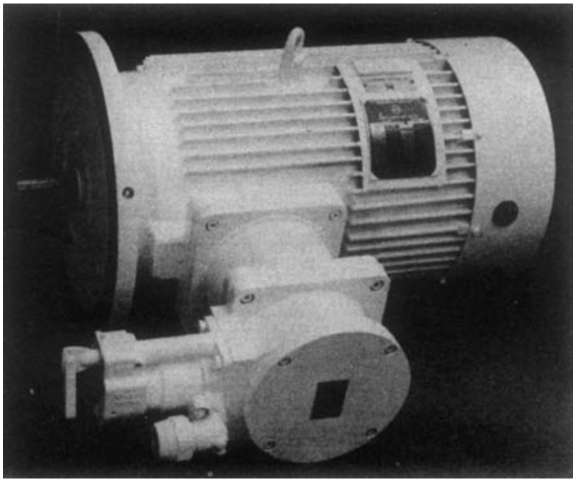


Figure 48.19 Flange-mounted flameproof motor

Table 48.3 National Coal Board (NCB) Specifications for cables

NCB Specification No.	*Subject*
P115	Short-firing cables (other than in shafts)
P188	Flexible trailing cables (for coal cutters and similar use)
P295	PVC-insulated wire armoured and sheathed cables
P492	PVC-insulated wire armoured telephone cables
P493	PVC-insulated wire armoured signalling cables
P504	Flexible trailing cables with galvanised steel pliable armouring
P505	Flexible trailing cables (for drills)
P610	Flame-retardant properties of flexible trailing cables
P648	Multicore PVC-insulated wire armoured and PVC sheathed 0.6–1.0 kV cables (with special screening for mine winder safety and control circuits)
P653	Flexible multicore screened auxiliary cables with galvanised steel pliable armouring
P656	EPR-insulated wire armoured and PVC sheathed cables

EPR, ethylene propylene rubber; PVC, polyvinyl chloride.

48.11 Cables, couplers, plugs and sockets

All cables used in mines for any purpose must conform to the requirements of the relevant BCC Specifications, which are listed in *Table 48.3*. Colliery surface and underground h.v. and m.v. distribution systems utilise, in general, PVC insulated and sheathed mains cables. DWA cables have up to a few years ago been exclusively used for underground systems but recently single wire armoured (SWA) cables have received favourable consideration, owing to reduced cost and flexibility in handling. Such cables conform to NCB Specification 295 or 656.

Wire armoured roadway cables are usually received at the colliery on 100 m drums with 300 A flameproof couplers already fitted at each end. On completion of assembly the cable coupler is filled with either bituminous compound or a cold-pouring compound, consisting of a bituminous oil and a hardener. A typical 6.6 kV, 300 A flameproof cable coupler is shown in *Figure 48.20*, the halves being connected by the use of three connector pins, a rubber sealing gasket and six connecting bolts.

Mobile machines such as power loaders and roadheading machines, which require to move while energised, must be powered by the use of a flexible trailing cable to satisfy the requirements of NCB Specification 188. Several types are available, ranging in size from 16 to 95 mm^2. *Figure 48.21* illustrates a typical trailing cable (type 7). It consists of three power cores, a pilot control core and an earth core. The earth core is uninsulated, and is located in the centre of the cable, with the three power cores and the pilot core equally spaced around it. All four cores are insulated with ethylene propylene rubber (EPR); the power cores have an additional copper/nylon screen over the insulation. Overall protection is provided by a tough heavy-duty polychloroprene (PCP) oversheath. A similar cable but of slightly different construction is the type 10 cable shown in *Figure 48.22*. The three power cores and pilot core are again EPR insulated, but each has a conductive rubber screen, and all four cores are laid up around a conductive rubber cradle separator. The overall protective sheath is PCP.

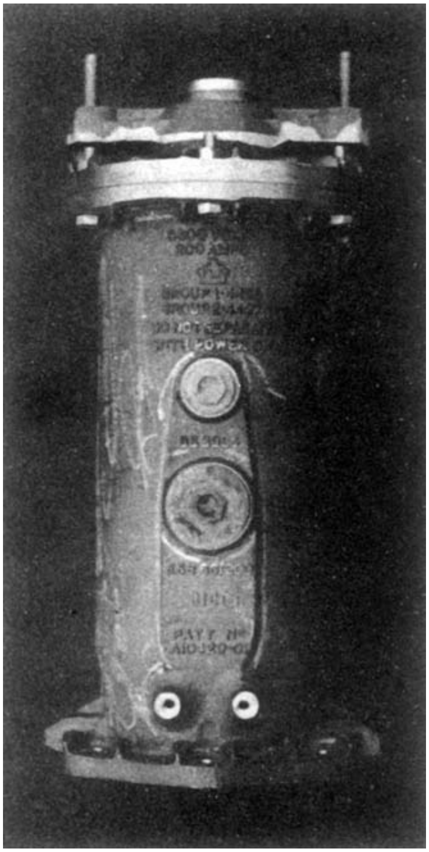


Figure 48.20 Flameproof cable coupler

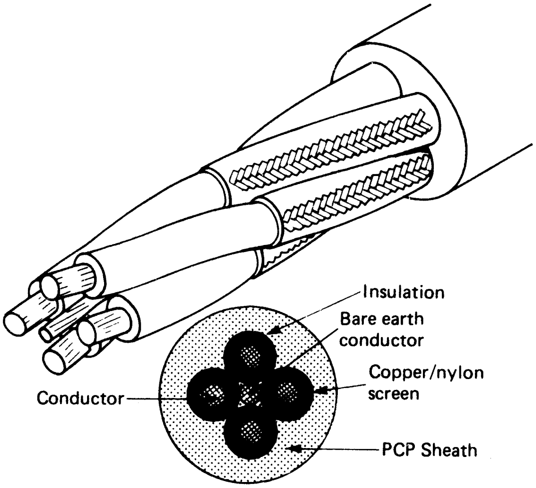


Figure 48.21 Trailing cable (type 7)

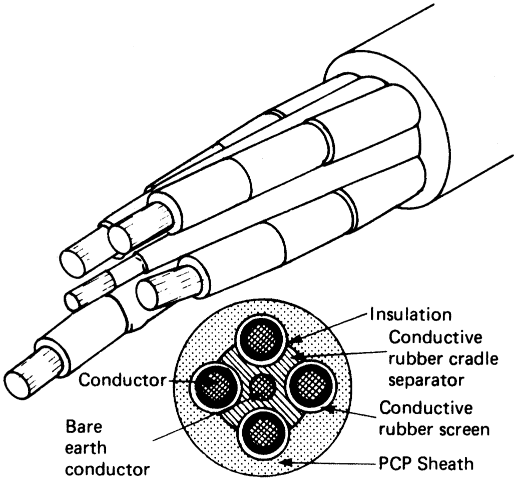


Figure 48.22 Trailing cable (type 10)

Since trailing cables associated with mobile machines are susceptible to damage, means must be provided to quickly connect or disconnect the cable from the motor/gate-end box. This facility is provided in the form of a 200 A, 1100 V flameproof restrained plug and socket, a typical example of which is shown in *Figure 48.23*. The socket is attached to the motor/contactor and the plug is attached to the cable. A scraper earth facility is provided on the socket which mates with the nose of the plug as it is inserted to maintain the necessary earth connection. Four pins are provided on the plug-and-socket assembly, three power cores and one pilot control core. The pilot pin is shorter than the power pins, so that on insertion it makes contact after, and on withdrawal it breaks before, the power pins, preventing their making or breaking on load.

Pliable wire armoured cables are used to power permanent and semi-permanent apparatus such as transformers and gate-end box assemblies which move up automatically as the coal-face advances. Such cables must conform to the NCB Specification 504, and range in size from 10 to 150 mm^2 at voltages of between 660 V and 6.6 kV. *Figure 48.24* illustrates a typical 6.6 kV pliable wire armoured cable (type 631) which has four cores insulated with EPR. A copper/nylon screen is provided over the insulation on the power cores, and all four cores are laid up around a PCP centre, over which is provided a PCP sheath. Over the inner PCP sheath is a galvanised steel strand armouring and an overall sheath of PCP.

Two, three-, four- and five-core pliable wire armoured cables of similar construction but with much smaller conductor size, e.g. 4 mm^2, are used for control circuits and coal-face lighting installation. Such cables are termed types 62, 64, 70 and 71, respectively.

Five-core 6 mm^2 flexible cables are used to power handheld drilling machines which operate at 125 V, three-phase, 50 Hz. Type 43 has three power cores, one pilot core EPR insulated and one earth core conducting rubber covered laid around a conducting rubber cradle separator, screened with conducting rubber and a heavy-duty overall sheath of PCP. Type 44 has five EPR-insulated cores with the three power cores copper/nylon screened, laid up round a PCP centre with a heavy-duty PCP sheath overall.

Cables for telephone communication are designed to NCB Specification 492, and can be either PVC-SWA-PVC

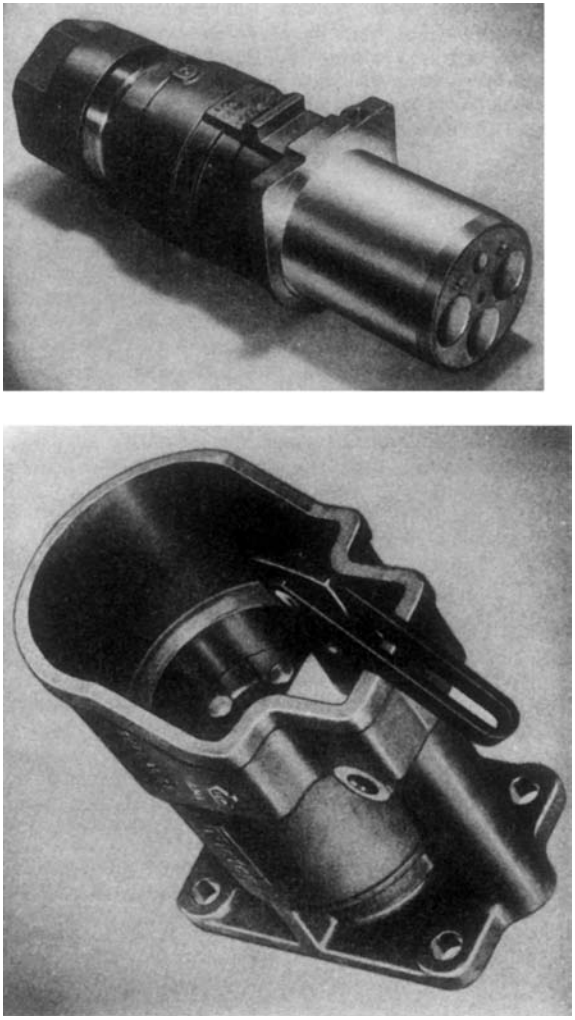


Figure 48.23 Flameproof restrained plug and socket

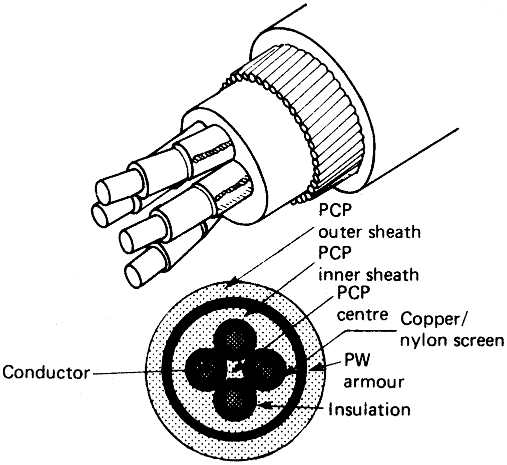


Figure 48.24 Flexible wire armoured cables

or PVC–DWA–PVC. The cores are laid up in one to 91 pairs of 1.5 mm^2 conductors and follow a set colour code.

Signal cables conform to NCB Specification 493, and can also be PVC–SWA–PVC or PVC–DWA–PVC. These cables have conductor sizes of 1.5 mm^2, ranging from 2 to 91 cores, and are laid up as single cores to a colour code.

48.12 Drilling machines

Mechanisation and automation have reduced the application of hand-held drilling machines. A typical machine is shown in *Figure 48.25*. Such machines are rated at 1.1 kW, 125 V, three-phase, 50 Hz, and operate from a purpose-designed flameproof drill panel. Drill panels are approximately the same size and shape as gate-end boxes and are designed so that they can be connected together mechanically. The drill panel contains its own step-down transformer to feed the drilling machine, contactors, protection, etc. The supply to the drill is taken from the drill panel via a 30 A flameproof plug and socket, and a five-core drill cable, consisting of three power cores, a pilot core and an earth core.

48.13 Underground lighting

Illumination underground is provided in accordance with the Coal Mines Safety and Health Regulations and to improve environmental conditions. The areas illuminated are pit bottoms, haulage stations, locomotive stations, main trunk conveyors, assembly areas and main roadways where men pass to and from their place of work.

Lighting underground is provided in four ways:

(1) by a portable lamp carried by each person underground;
(2) by permanent lighting installations supplied by a power transformer at 120 or 240 V, 50 Hz;
(3) by mains lighting which forms part of a mobile machine; or
(4) by portable compressed-air turbines.

The cap lamp (*Figure 48.26*) comprises a headpiece provided with a main 4 V, 0.9 or 1.0 A lamp and a 4 V, 0.46 A pilot lamp. The headpiece, which can be carried or worn on a special helmet, is connected by flexible cable two-core 1.0 mm^2 vulcanised-rubber-insulated PCP sheath. The battery consists of two lead–acid cells in a polycarbonate moulded case giving 4 V output. A fuse is provided in the battery top to afford protection. The capacity of the batteries is 13 or 16 A-h.

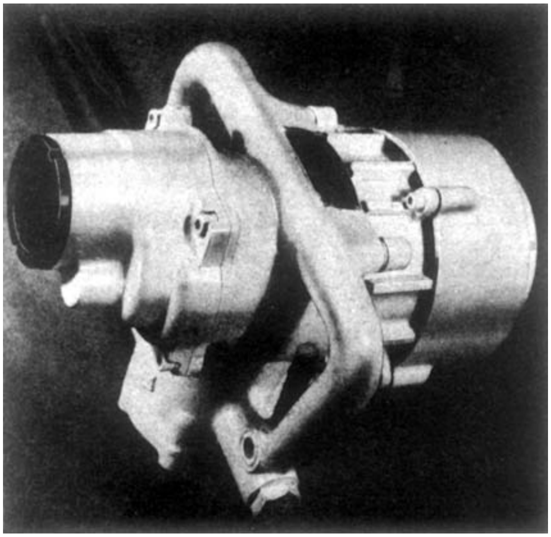


Figure 48.25 Hand-held drilling machine

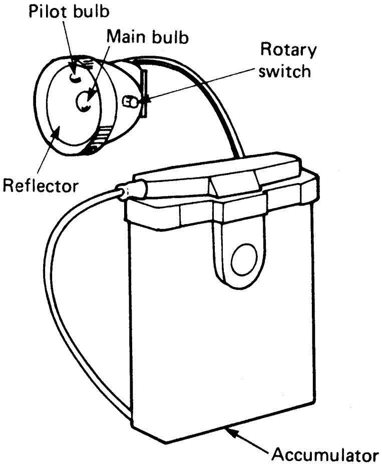


Figure 48.26 Cap lamp

The assembly was covered by BS 4945, but this has been superseded by the European Norm CENELEC 50 033: 1986, i.e. BS 6881:1987.

To prevent interference underground, all lamps are locked and sealed before issue at the surface lamp room. *Figure 48.27* shows a typical lamp-room layout. The miner enters the lamp room and collects his own personal cap lamp before entering the mine. On his return he places his lamp on the charging rack in the lamp room, when recharging commences automatically. During the period between shifts, the lamps are examined, cleaned and topped up ready for the next period of duty.

Permanent lighting installations below ground are similar to those provided on the colliery surface, *viz*. filament, fluorescent, discharge, sodium, and mercury lamps, in substantial

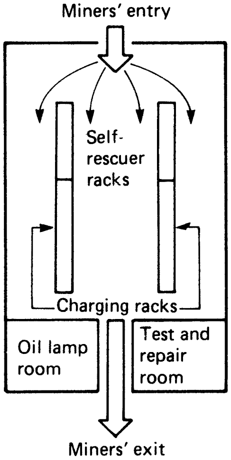


Figure 48.27 Layout of a lamp room

dust-proof fittings outside hazardous areas, and flameproof to group 1 (methane) in the hazardous areas. Special attention is paid to the design of electric circuits, proper loading, and fault protection, i.e. earth fault, overload and short circuit.

Illumination on mobile machines is provided by equipment designed and tested for the automotive industry but fitted into flameproof lights with protection to prevent damage. In mining development work where the heading is progressively moving, the provision of lighting on machines is advantageous.

Coal-face lighting has had a long history. Numerous approaches and types have been used but have failed for various reasons. The conditions imposed by the Coal Mines Act 1911 limited such installations to 'naked-light mines' only; it was not until 1934 that mains lighting on the face in safety lamp mines was permitted. In 1973 mains fed intrinsically safe lighting was developed using 12 in., 8 W fluorescent tubes on a high-frequency supply. This has resulted in a smaller, lighter, more easily maintained system.

48.14 Monitoring and control

Different types of monitoring and control systems are in use at collieries: either 'surface only' installations, designed as for other industries, or 'surface/underground' systems, which are specifically designed for use underground, requiring IS.

A typical system for use underground consists of a surface control or central station, a number of two-way transmission links (called 'rings') to convey data over distances up to 8 km, to and from numbers of underground FLP/IS out-stations to which are connected sensors for monitoring and control facilities. These systems may perform individual functions (e.g. coal clearance by remotely controlling conveyors, environmental monitoring, pumping, or fan monitoring), or may be multifunctional.

A surface control or central station would consist of a mini-/micro-computer, controlling a number of transmission rings. Transmission rings are usually from two- to six-wire time division multiplex data transmission systems, capable of approximately 500 bit/s, coupled to a maximum of 32 outstations. Some frequency-multiplex systems are use: in this case the signals are passed along armoured low-loss coaxial cable. Audio frequencies, modulated low radio frequencies or modulated v.h.f. frequencies are employed.

The out-stations, sited underground near to the plant concerned, are constructed part FLP to give a number of IS supplies from 110/240/550 V, and part non-FLP to house the printed circuits and terminations. The out-stations are required to collect information from various sensors and other circuits, pass this information to the transmission rings and receive from them surface information and commands. Action can then be taken by the out-station on these commands, or independently appropriate shutdown and alarm action can be taken through fault-tripping logic, according to the state of the protective sensors and other monitors associated with the plant.

Where possible, all circuits are designed for 'failure to safety'; therefore a.c. signals are preferred, with the final control operator being a relay, supplied through an isolating transformer or 'diode pump' circuit, fed through two high-voltage series capacitors, the relay being energised in the healthy state.

On/off sensor signals are obtained using a half-wave rectifier at the sensor end and an a.c. IS supply to provide open- and short-circuit cable protection.

The surface control (*Figure 48.28*) could be in an air-conditioned room and may house many separate control and monitoring systems, together with a data collection computer system, using larger disc storage. The electrical supply is taken from a source as near to the colliery feed as possible, through constant voltage transformers (CVTs) to give as clean and permanent a supply as possible.

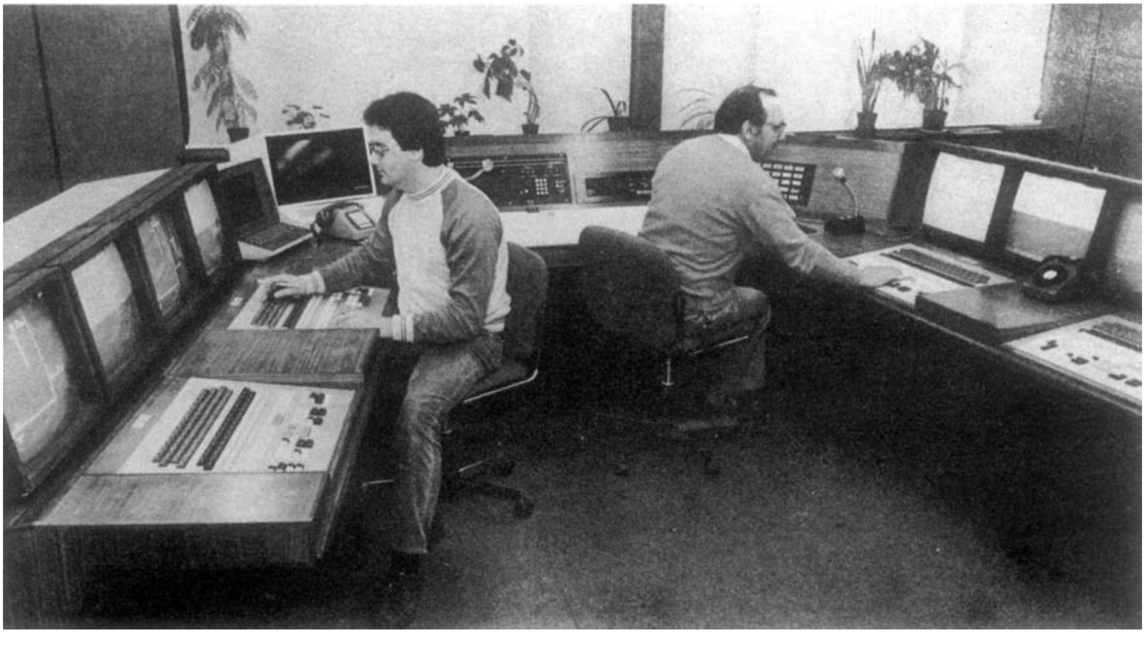


Figure 48.28 A surface control room

48.14.1 Computer system

48.14.1.1 Hardware

A list of the hardware used is given below.

(1) Mini- and/or micro-computers;
(2) floppy or hard disc loading and data storage;
(3) all parity checks utilised; 'watchdog'-type timer system; system scan time with 1 s;
(4) two-colour graphic visual display units (VDUs) displaying mimic representation of system (faults, alarms etc.) in different areas of the screen;
(5) keyboard used to effect system changes and issue commands;
(6) print—300 baud rate of print-out shift analysis of data, breakdown, etc.;
(7) suitable intrinsically safe interface to the underground transmission system;
(8) all changeable-type ROMs, marked to a standard by the suppliers; and
(9) computer interconnections via 20 mA serial loops, terminals, VDUs, etc., connected via either 20 mA loop or VT24.

48.14.1.2 Software

Software is normally written in assembler of Coral-66 languages. Most systems have executive type control software to give safer systems and provide rapid response to real-time requirements. Software is written to be as inherently safe as possible and is resident in the memory at all times. Some systems perform periodic check sums on ROM and RAM memory, as well as transmitted data. Entry to make system changes or give commands is restricted by level of password. Some diagnostic software is available to check much of the hardware, including on-line checks of transmission data. A software system is supplied to a colliery in package form, the package containing an operating system and a set of modules or application programs from which a particular system may be configured. Configuring, carried out in password, is on a questionnaire basis and allows the operator or engineer to: add or delete items of plant control facilities (sensor, etc.); change set points, levels and limits of analogue signals; set out-station types; build, using graphics, a representative mimic of the colliery system; select data to be periodically printed; etc.

48.14.2 Underground sensors

Listed below are some of the many different types of underground sensor, together with a few of the varied monitoring applications of each.

(1) *Temperature*: bearings, brakes, and temperature rise.
(2) *Pressure*: barometric, fluid, and differential (for flow).
(3) *Flow*: liquids, ventilating air, and gas.
(4) *Proximity*: position and movement (for velocity and acceleration).
(5) *Weight, volume, force*: belt load, static weighers, chain force, torque, and tension.
(6) *Vibration*: bearings.
(7) *Gas and atmosphere analysis*: methane, oxygen, carbon monoxide, carbon dioxide, humidity, dust, and smoke.
(8) *Electrical analysis*: power, voltage, and current.
(9) *Level*: coal/stone bunkers, and fluid level (e.g. of lubricating oil).

Most sensors, including any electronic circuitry, are designed to be IS 'fail-safe'. With on/off switching, a.c. IS voltages are used, with a half-wave rectifier at the sensor. Analogue signals are 0.4–2.0 V, or 4–20 mA. Underground sensors must be manufactured from approved materials, such as brass, to eliminate incendive sparking risks.

Analogue temperature In these measurements, by virtue of the low energy capacity of the IS supply low impedance devices such as, for example, 100 Ω platinum resistors or thermistors are used in order to enable a number of points to be monitored.

Measuring electronics are designed to be 'fail-safe' throughout and provide a variable shutdown/warning level or in some cases (such as monitoring of compressors) fixed maximum shutdown levels, for example the maximum air temperature would be 160°C and maximum oil temperature 80°C.

Pressure Measurement ranges vary from 0–0.25 kPa to measure the differential pressures across mine stoppings to 1500 bar for hydraulic pressure.

In general the lower pressures are measured using a diaphragm-type transducer, higher pressures are usually monitored by strain-gauge bridges.

49

Standards and Certification

H W Turner BSc, CPhys, FInstP

Contents

49.1 Introduction 49/3
  49.1.1 The need for standards and the aims of standardisation 49/3

49.2 Organisations preparing electrical standards 49/5
  49.2.1 British Standards Institution (BSI) 49/5
  49.2.2 International Electrotechnical Commission (IEC) 49/6
  49.2.3 European Committee for Electrotechnical Standardisation (CENELEC) 49/7
  49.2.4 Underwriters Laboratories Inc. (UL) 49/8
  49.2.5 Other bodies (ISO other national and regional committees etc.) 49/8
  49.2.6 Communication with organisations concerned with standards 49/9

49.3 The structure and application of standards 49/10
  49.3.1 Different types of British Standard documents 49/11
  49.3.2 International equivalents of British Standard documents 49/13
  49.3.3 Structure of a typical standard 49/13
  49.3.4 Implementation of standards 49/13

49.4 Testing, certification and approval to standard recommendations 49/14

49.5 Sources of standards information 49/14
  49.5.1 Addresses of organisations concerned with standards 49/15

49.1 Introduction

Standards are vital components of UK marketing strategy. Ordering and procurement is increasingly done over the Internet and in order to ensure that the products are suitable for the purpose required, compliance with a nationally and internationally approved standard is of increasing importance to customers. Reflecting the growth of electronic communication, standards authorities such as the British Standards Institution have been making considerable investment in Information Technology in order to exploit the rapid communication possibilities of IT in providing industry with the standards it needs. From the year 2000 this will become the major means of communication in the standards field. Details of information sources are given in subsequent parts of this chapter.

A standard is defined as a technical specification or other document available to the public, drawn up with the cooperation and consensus or general approval of all interests affected by it, based on the consolidated results of science, technology and experience, aimed at the promotion of optimum community benefits and approved by a body recognised on the national, regional, or international level.

The National Standards Body recognised in the United Kingdom is the British Standards Institution (BSI) whose principal function is the preparation and/or publication of national standards and/or the approval of standards produced by other bodies. BSI serves as the national member of the corresponding international and regional standards organisations. The international standards organisation principally concerned with electrical standards is the international Electrotechnical Commission (IEC) and the corresponding regional standards organisation for Europe is the European Committee for Electrotechnical Standardisation (CENELEC). CENELEC is very important to UK manufacturers, because it has the power to issue *Directives* which are binding on all EEC member states and effectively outlaw any product which does not fully comply with the standard requirements contained in that directive. More detail of these and other standardising bodies is given in Section 49.2.

Standards are revised, when necessary, by the issue of amendments, or, in the event of the need for major alterations or additions, by the issue of new and revised editions. Users of standards must ensure that they use the latest amendments or editions. Standardising bodies supply regular information on all such changes (in the United Kingdom, this information is summarised annually by BSI who also issue their catalogue in electronic form. The information is up-dated monthly in *Update-Standards* and general information about standards is published bi-monthly in *Business Standards*. Both publications are available from BSI.)

More immediate access is provided over the Internet on the BSI Website : www.bsi-global.com

49.1.1 The need for standards and the aims of standardisation

The primary need for standardisation is to provide a set of criteria by which a product can be judged to be suitable for the purpose for which it is intended in comparison with like products from a variety of sources. This is of value to both manufacturer and user (see Section 49.1.1.1). Achievement of all the targets of performance set in the standard by samples taken from production at regular intervals is also a measure of the sustained quality of manufacture, the standard again providing the yardstick by which consistency of manufacture can be assessed (see Section 49.1.1.5).

Those preparing standards have common agreed aims (see Section 49.1.1.2) designed to make the standard of optimum applicability and usefulness to meet the needs of the application.

Standardisation is needed also to avoid the pitfalls of proliferation of sizes and performance specifications which in the absence of standardisation would make difficult the replacement of a damaged component by one of different manufacture or even from the same manufacturer if modifications have been subsequently made to the design (see Section 49.1.1.3).

Last but by no means least, safety requirements can be carefully thought through by a standardising committee which has representation of health and safety bodies, to ensure that there are clauses in the standard which spell out clearly the necessary requirements and design criteria which the product must meet in order to ensure that it is safe in normal use (see Section 49.1.1.4).

49.1.1.1 The value of standardisation

Users are increasingly aware of the value of purchasing a product certified to national and international standards. They can rely on its performance to a known set of criteria, and if its quality of manufacture is regularly assessed, they know that the standard of performance is maintained and that the product does not have design faults which are likely to make it fail prematurely in normal service. This can ensure a financial saving not only of the cost of early replacement but also of the cost of the 'down-time'. Furthermore, with the more strict personal liability for health and safety at work applying today, the user is even keener to ensure to the best of his ability that the product he is using is not unsafe in use. If he buys a product which has not been tested to the appropriate safety requirements of the standard the chance of an accident is much higher. He also may consequently be legally liable, whereas if the product had been certified to the appropriate standards and correctly installed and applied, he would have done all that he could do to avoid an accident.

New regulations introduced in 1966 (detailed in 'RIDDOR 95' published by the HSE) put extra responsibilities on employers when accidents occur in the workplace, and legislation is becoming increasingly intolerant of those employers who neglect to provide adequate protection. Compliance with all safety aspects of standards is therefore paramount in producing a satisfactory risk assessment.

The commercial advantage to the manufacturer in having a range of products fully certified to the appropriate standards is therefore very apparent. His customers need to buy his products of demonstrated performance and safety, rather than making false economies by purchasing cheap untested competitive products which are likely to fail or malfunction in service, possibly dangerously. The certification to safety standards gives also a degree of protection against legal action in the event of accident involving the product.

The value to the general public at large of standardisation is the improved reliability of the equipment around them, giving longer uninterrupted service and greater personal safety.

Under the Directives of the EEC, larger users will be compelled to change their installations, replacing components not conforming with harmonised European standards. This is considered further in Section 49.2.3.

Standards are also needed to simplify contractual agreements, because the supplier can ensure that his products comply with all the relevant clauses of the appropriate standard and such compliance can be written into the contract for any of his customers. Otherwise it would be necessary to prepare a detailed technical performance specification for each product for each customer. Having stated the appropriate standard it remains only to add any special requirements, delivery dates, price agreement etc. and both sides are presented with an unambiguous contract. Savings on stock levels are also facilitated, because it is generally only necessary to stock standardised items for which there is more general demand.

49.1.1.2 Specific aims of standardisation

From considerations of the need for standards similar to those summarised above, five aims of standardisation have been identified in BS 0:1991:

(1) provision of means of communication amongst all interested parties;
(2) promotion of economy in human effort, materials and energy in the production and exchange of goods;
(3) protection of consumer interests through adequate and consistent quality of goods and services;
(4) promotion of the quality of life: safety, health and the protection of the environment; and
(5) promotion of trade by removal of barriers caused by differences in national practices.

In the following sections we can see how well these aims have been achieved by national, regional, and international standardisation. BS 0 has only recently been completely revised, but with the advance of the electronic age it will doubtless soon require further revision.

49.1.1.3 Reduction of proliferation

The issue of a standard promotes the reduction in proliferation of product sizes which do not fit in standard equipment, product ratings which are too restrictive for general application and products which are non-compatible with all the systems of which they are intended to be a component.

Standardisation therefore represents a major economy in the stocks of spares which have to be kept to replace spent or worn-out components, and considerably eases the task of designers of equipment incorporating a range of products, with standard ratings based where possible on the R10 series. This harmonisation means that the standard product can be specified, knowing that it will fit and will have the required technical performance with the specified level of safety.

This principle is increasingly being extended internationally through the work of the IEC (see Section 49.2.2) and in Europe by CENELEC (see Section 49.2.3). After 1992 many CENELEC harmonised documents will have the force of law, and the corresponding products will only be saleable in the UK and Europe if they comply in all respects.

49.1.1.4 Safety

Standards also cover aspects of safety. In this respect they seek to define the acceptably reasonable level of risks of personal harm in the foreseeable use or misuse of a product, process or service.

An example of a standard specifically classifying different levels of acceptable risk is BS 60529 (identical with IEC 60529) 'Classification of Degrees of Protection Provided by Enclosures'. This standard specifies tests for a series of levels of protection and defines an IP (International Protection) code which identifies the degree of protection against direct contact, against ingress of solid objects and/or water, etc. For example an enclosure classified as IP2X has been tested to ensure that an average finger cannot be poked into the enclosure in such a way that it can come into contact with live parts or dangerous moving parts. However, wires can be poked in. If protection is provided against touching by ingress of wires down to 1 mm. diameter, the category is IP4X. Space does not permit further detail here, and readers are referred to the standard, which is applicable to enclosures containing any electrical equipment, providing its rated voltage does not exceed 72.5 kV.

Standardising bodies also publish documents summarising up-to-date knowledge on safety related matters. An example of this is IEC 61479-1 'Effects of Current passing through the Human Body', which provides basic guidance on the effects of shock currents on the human body. This provides a most useful source of reference when establishing electrical safety requirements for products, or when devising regulations for electrical safety with respect to avoidance of death or injury from electric shock. Hazards due to overheating of electrical equipment can lead to oxidation of contact surfaces, deterioration of insulation, burns to personnel, or in an extreme case to ignition and fire. Consideration of these matters becomes complicated when a number of heat producing units are contained in a single enclosure because the temperature rises permitted in switchgear standards result from conventional situations which may differ appreciably from the situation within an enclosure and connected to other heat producing equipment. In this case, guidance can be obtained from IEC 60943 'Guide for the specification of permissible temperature rise for parts of electrical equipment, in particular for terminals'. IEC also publishes a bound volume containing the text of all its electrical safety standards up to the time of publication. This is a very useful compendium, sold at a price significantly less than the total cost of the standards bought separately, and provides a most handy source of reference although it is necessary to check that a particular standard has not been up-dated since the date of publication.

Safety aspects contained within product specifications establish matters such as marking of the product to inform the user of the product ratings and limitations, and where necessary to warn of dangers of misapplication etc.

Safety tests are also included in standards, e.g. by limiting the emission of arc products, restricting temperature rise to prevent injury, decomposition and degradation of insulation or other properties of materials, or ignition leading to fire hazard, etc. Sometimes the level of safety is indicated in the title of the standard e.g. if it states that it is intended for use by authorised persons, which implies that there can be an additional level of hazard in untrained hands.

It is impossible to make any product completely safe. One cannot do without water, but one can drown in it. Safety regulations reduce the risk. In the same way the standard will generally set a defined level of acceptable risk, at a point where most hazards are eliminated as far as is possible without extravagant expenditure on remote chances of danger through misuse. A warning will normally be found in the standard where there could be any hazard that may arise in the use or foreseeable misuse of the products covered by the standard.

49.1.1.5 Quality assurance

The quality of goods and services are provided for in standards by defining features and characteristics which establish their ability to satisfy the stated needs. A quality assurance procedure provides a standard framework for regular checking to ensure maintenance of quality in production, and continued adherence of the product to the standard to which it is claimed to conform. It would be inappropriate to go into any great detail here, but the reader is referred to the following British Standards: BS 4891 contains a guide to quality assurance, terminology is in BS 4778, and specification of quality systems can be found in BS 5750.

Although not an electrical standard, firms of quality often also seek approval to ISO 9000 to demonstrate quality of management. Year 2000 revisions to the ISO 9000 series of quality management standards are expected to bring a great many new business benefits to companies making the transition to the new standard (details available from BSI head office).

49.2 Organisations preparing electrical standards

One of the problems facing a manufacturer developing a product for export is that the product is required to conform with the appropriate standard which applies to the particular country to which he is exporting. He may already have test certificates to certify complete compliance with the British Standard, as is necessary for sales within the UK, but the other country may have national standards which differ in significant details from the British Standard and local regulations and trade practices which enforce an expensive series of re-testing to slightly different prescriptions in a test station in the other country. This is particularly troublesome in the USA where manufacturers of tried and tested products find that they must spend considerable sums of money and time having their products tested once again at the Underwriters Laboratories (UL) or some similar test company before they are able to market that product in the USA. Now that Canada has joined with USA in a common market activity, it is likely that the same will apply to Canada in future. The organisation of standards in the USA is somewhat chaotic, compared with other developed industrialised countries. There are approximately 35 different bodies producing electrotechnical standards for the USA. These are loosely coordinated by the national body, The American National Standards Institution (ANSI) which does not develop standards, but may approve standards produced by the other bodies as American National Standards. Apart from UL standards and similar product standards, there are two basic standards generally recognised in the USA:

(1) The National Electrical Code, produced by the National Fire Protection Association (NFPA); and
(2) The National Electrical Safety Code, produced under the auspices of the Institute of Electrical and Electronic Engineers (IEEE).

The National Electrical Manufacturers' Association (NEMA), the US equivalent of BEAMA is also active in producing standards. The system is illustrated below for the case of the Underwriters' Laboratories Inc.

A parallel situation existed for many years in Europe, where approvals were needed in each country, especially for electrical equipment sold to the general public, and equipment sold in several countries would be studded with approvals marks for KEMA, VDE, SEMKO, NEMKO or DEMKO, etc. as proof to the purchaser that the local approvals authority had passed it. The European Committee for Electrotechnical Standardisation (CENELEC) since its formation in 1972 has reduced much of the unnecessary duplication of testing activity and has introduced a new approvals mark which should eventually replace all the existing marks within the CENELEC regional group of countries. However, to ensure that all electrical equipment in the EEC is up to a uniform standard of quality and safety, it has become mandatory for all electrical goods to carry the 'CE' mark.

As detailed below, CENELEC is far advanced in actively converting all National Standards within the EEC into European Standards and in the process eliminating national differences in preparation for the removal of trade barriers in 1992. Although this gives many technical and commercial problems, it will undoubtedly simplify European trade when the process has worked through the system. Canada and the USA are going through a similar transition, having formed a similar regional group.

Other international and regional standardising bodies exist as detailed below, but most of the electrotechnical groups tend to follow the standards produced by the International Electrotechnical Commission (IEC).

The IEC was formed in 1906 to harmonise all electrotechnical standards world-wide and to rationalise the test and certification requirements in such a way that all nations can work to one set of electrotechnical standards. Considerable progress has been made this century in that direction, and it is only to be hoped that the strengthening of the regional groups will not delay the progress to World harmonisation by leading to the erection of more formidable regional barriers to trade.

49.2.1 British Standards Institution (BSI)

In the UK, the British Standards Institution is the recognised body for the preparation and promulgation of national standards in all fields. In 1901, it was set up by the professional engineering associations with the title 'Engineering Standards Committee' changed at the end of World War I to the 'British Engineering Standards Association'. A Royal Charter was granted in 1929 supplemented in 1931 when the body was retitled the 'British Standards Institution', an independent body whose Royal Charter was consolidated in 1981 as an independent body with the following objectives:

(a) to co-ordinate the efforts of producers and users for the improvement, standardisation and simplification of engineering and industrial materials so as to simplify production and distribution, and to eliminate the national waste of time and material involved in the production of an unnecessary variety of patterns and sizes of articles for one and the same purpose;
(b) to set up standards of quality and dimensions, and to prepare and promote the general adoption of British Standard Specifications and schedules in connection therewith, and from time to time to revise, alter and amend such specifications and schedules as experience and circumstances may require;
(c) to register, in the name of the Institution, marks of all descriptions, and to approve and affix or license the affixing of such marks or other proof, letter, name, description or device; and
(d) to take such action as may appear desirable or necessary to protect the objects and interests of the Institution.

BSI acts as the national member of the corresponding international and regional standards organisations and as such is responsible for paying the subscriptions to these bodies (more than a million pounds!). As a completely independent body, BSI has to raise income to meet its expenditure. A large proportion of its income is derived from quality assurance activities, sales of publications, testing and technical help. The remainder of its income was in the main divided between subscriptions from the tens of thousands of members and Government grant-in-aid. Electrotechnical Standardisation represents a significant fraction of BSI activity, due to the high technological level and national importance of the industry, and its rapid rate of advance into new and exciting fields of discovery and development, necessitating the preparation of standards for products which did not exist in any form in earlier generations.

Preparing standards for such innovatory products necessitates the formation of new committees to add to the present total of over 3000 technical and subcommittees, each committee being provided with a secretary by the BSI, with over 25 000 committee members nominated by organisations representing the views of users, manufacturers, health and safety authorities, testing authorities etc. BSI maintains more than 10 000 current British Standards and a similar number of standards projects. Approximately one third of these are electrotechnical in nature.

Membership of BSI is divided into several categories, for example: individuals, professional firms and partnerships, industrial and commercial firms in both the private and public sectors, professional institutions, research associations and similar bodies, local authorities, and other organisations. Members pay subscriptions related to the size of the organisation.

Committee members are mainly nominees of the organisations which they represent and are responsible for standardisation project work, and obtaining the consensus of opinion of interested parties on any matter to be raised at committee meetings, reporting back to their organisations.

Each of the tens of thousands of committee members is therefore reporting back to a large number of other persons and organisations which indicates the vast amount of effort and time devoted to standardisation matters, and the importance attached to it by all concerned.

British Standards may be used to promote standardisation in any of the following stages:

(1) terminology, symbols;
(2) classification;
(3) methods of measuring, testing, analysing, sampling, etc.;
(4) methods of declaring, specifying, etc.;
(5) specification of materials or products;
(6) dimensions, performance, safety, etc.;
(7) specifications for processes, practices, etc.;
(8) recommendations on product or process applications; and
(9) codes of practice.

The most familiar type of British Standard Specification is one that lays down a set of requirements to be satisfied by the material, product, or process standardised and which embraces, often by reference, relevant methods by which compliance may be determined.

More information on the different types of British Standard Specifications and other British Standard publications are given in Section 49.3.1.

British Standards were originally all publicly available documents voluntarily agreed. However, the publication by the BSI does not at present ensure their use. This position is progressively changing, notably where the standard is also a harmonised European standard where compliance becomes mandatory. Otherwise a British Standard only becomes binding if a claim of compliance is made, if it is invoked in a contract, or if it is called up in some other legislation. Regulations made under a number of British Acts of Parliament call up approximately 300 British Standards.

Support for the application of British Standards as agreements produced in the public interest is given by the Restrictive Trade Practices Act 1976. The care exercised in the production of British Standards is relied upon by users who themselves owe a similar duty to the public. The compliance with the requirements of a British Standard does not in itself necessarily imply that a product is safe and suitable for all possible applications. It remains the responsibility of users to ensure that a particular British Standard is appropriate to their needs. Within their scope, British Standards provide evidence of an agreed ‘state of the art’ and may be taken into account by the Courts in determining whether or not someone was negligent.

In July 1982, the Department of Trade on behalf of HM Government issued a document entitled ‘Standards, Quality and International Competition’ (Cond 8621) in which the BSI was recognised as the main producer of standards in the UK, and HM Government agreed to maintain the annual grant-in-aid based on the level of contributions from other subscribers and to support BSI’s efforts to achieve international harmonisation of standards through ISO, IEC, CEN, CENELEC, etc. in the interests of UK industry and trade, and to encourage fuller participation of public purchasing authorities in the preparation of and compliance with British Standards in their purchasing decisions, quality assurance requirements, and in their operational procedures. The BSI undertook the huge task of reviewing, and where necessary revising existing British Standards to ensure that these, and any new British Standards, will be suitable for reference in Government Regulations as unambiguous statements of technical requirements. With the rapid movement to international harmonisation, in particular within the European Regional harmonisation within CENELEC, BSI is somewhat hindered in fully carrying out this task, since agreement must be obtained with all the international partners before any change whatever can be made in a harmonised standard. However, the Public Procurement Directive enforces mandatory compliance with the corresponding European standard, and thus some apparent ambiguities may be ineradicable.

However, both BSI and CENELEC are agreed that IEC standards are the most suitable for harmonisation, since they have been formulated with the benefit of world-wide expert opinion and agreed by a substantial majority vote of all interested nations. They thus offer the best chance of achieving a common set of standards world-wide. There are two major benefits:

(1) a major international trade barrier is removed if the trading countries have common standards; and
(2) users specifying a common standard are ensured that they have a common and valid base for examining and comparing competing products.

British Standards and new European standards for electrotechnical products etc. are therefore almost exclusively generated by the IEC. In order to have any influence on the content of future British Standards it is therefore vital to have adequate representation on IEC committees and even more important to maintain active participation in all the Working Groups of IEC, since it is in the Working Groups that basic standards are formulated and contentious points are resolved.

49.2.2 International Electrotechnical Commission (IEC)

The International Electrotechnical Commission (IEC) was established in 1906, and now comprises the national electrotechnical committees of over 50 countries in all continents throughout the world. During the twentieth century the IEC had produced over 3000 standards compiled by over 80 technical committees, and over 120 sub-committees that collectively represented over 80% of the world's population that produced and consumed 95% of the total electrical energy. National committees include representatives from manufacturers, users, testing authorities, trade associations, governments, and professionals and academics from research organisations and colleges.

IEC Standards are widely adopted as the basis of national electrotechnical standards so far as local customs and conditions permit. They are also quoted in manufacturers' specifications and by users when calling for tenders This widespread adoption facilitates international trade in the electrical and electronic engineering sectors.

The International Conference on Large High-Voltage Electric Systems (CIGRE), meeting in Paris when necessary, has a number of working groups which produce papers discussed at the meetings and subsequently form the basis of work to produce standards within the IEC. The IEC works in close cooperation with the International Organisation for Standardisation (ISO) which is mainly concerned with standards in non-electrical fields. However there are overlapping areas, for example the ISO deals with automotive electrics (see Sections 49.2.5), and IEC is responsible for standards for steam and hydraulic turbines, which are almost exclusively employed for generation of electrical power.

In 1982, the IEC began its Quality Assessment System for Electronic Components (IECQ). This was initially introduced for quality assessment of mass-produced components, such as resistors and capacitors, but it has been much extended to embrace components made for special purposes. The major exporting countries of electronic components, more than 20 countries, are members of IECQ.

The British Standards Institute is the UK member body for the International Organisation for Standardisation (ISO) and the European Committee for Standardisation (CEN).

The Electrotechnical Council forms the British Electrotechnical Committee, the UK national committee of the parallel International Electrotechnical Organisation, the International Electrotechnical Commission (IEC) and the parallel regional electrotechnical organisation the European Committee for Electrotechnical Standardisation (CENELEC). BSI also participates as the UK member in producing European Standards (EURONORMS) for the European Coal and Steel Community (ECSC).

Towards the end of the twentieth century BSI was working on over 10 000 standards projects, the major proportion of which were involved in international standardisation. At the same time they held the UK Secretariats of about 200 technical- and sub-committees of international standardising bodies, of which 17% were within IEC and CENELEC, and many more secretariats of working groups. The major importance to the UK of its electrical industry would suggest that Britain should be deeper involved in this activity, but BSI is compelled to keep its activities within the provision of services within the UK, due to the constraints of its limited budget. The sale of standards was anticipated to be enhanced by the pace of international standardisation in which case it could have offered some growth of income for promoting such work. Even this source could be threatened if, for example, a decision were made to sell all European standards from Brussels or some other profit centre outside the UK. With the revolution in Information Technology it has become possible to deliver Standards over the Internet, and BSI has invested large sums in the provision of advanced IT facilities to meet this challenge. Standards are available to subscribers on-line (www.bsi-global.com), and many standards are also available on CD ROM as well as the traditional hard-copy format all available from the BSI bookshop (open to all every weekday from 0900 to 1730). 'Print on demand' has been introduced from images stored on electronic files and enquiries can be transmitted to BSI by e-mail: info@bsi.org.uk

Many new strategies are being developed to support the continuation of BSI's standardisation work, and a substantial range of new services is being made available to customers in addition to the loan facilities available at modest cost from BSI library.

BSI has also established many centres overseas, which are listed in Section 49.5.1.

Every international committee in which the UK participates has an equivalent committee, usually a BSI committee which appoints and briefs the UK delegation. Each delegation has a leader who is the principal UK spokesman at the international meeting, and a Rapporteur to assist the leader in technical interpretation of documents at the meeting and to keep a record of the proceedings from which he produces a brief report of the main decisions of the meeting for BSI. Members chosen as delegates in international standards work are chosen for their special knowledge and powers of advocacy, and they are responsible for presenting the viewpoint of the UK as agreed in the corresponding BSI committee. Five responsible and properly briefed delegates would normally be the maximum number on any committee, ideally comprising the leader, a manufacturer, a user, a testing specialist, and a technical expert with an academic or industrial research background (although the size and composition of the delegation will vary depending on the availability of funding for the attendees, and the nature and importance of the matters to be discussed). If regulatory matters are to be discussed, the delegation should also, if possible, include a representative of the appropriate Government department serving also on the corresponding BSI committee. As stated above, this representative should also actively participate in the work of the working group preparing the voting document if there are any special UK requirements which need to be included. Special European standards which need to be developed where there is no equivalent standard available or under consideration in the IEC.

The IEC is governed by IEC council and its committee of action. It operates also certain special committees referred to by the following acronyms:

ACET	Advisory Committee on Electronics and Telecommunications
ACOS	Advisory Committee on Safety
CISPR	International Special Committee on Radio Interference
ITCG	Information Technology Coordination Group

49.2.3 European Committee for Electrotechnical Standardisation (CENELEC)

In 1959 the standardisation institutions of Western Europe formed the European Committee for Standardisation (CEN) which co-ordinated the drafting of standards within the two regional trading groups: The European Economic Community (EEC) and the European Free Trade Area (EFTA). Electrotechnical standards were then made the responsibility of CENEL (the electrotechnical counterpart of CEN). The CEN certification body was referred to under the acronym CENCER.

However, as the EEC began to be enlarged in 1972, the European Committee for Electrotechnical Standardisation (CENELEC) was set up which creates harmonisation documents (HDs) and European standards (ENs). CENELEC members are the national electrotechnical committees of Austria, Belgium, Denmark, Finland, France, Germany, Greece, Ireland, Italy, Luxembourg, Netherlands, Norway, Portugal, Spain, Sweden, Switzerland and the UK. Once an EN is approved, by the agreed voting procedure, it must be printed or endorsed as a national standard in all the countries within the EEC. Electrotechnical ENs and HDs are preferably derived from, or exact copies of, corresponding IEC standards. However, after 1992, when all trade barriers were removed in the EEC (under Article 100 of the treaty of Rome), it became apparent that there were many national variants in the standards of the various EEC countries which would need to be resolved before a truly free-trade situation could exist. The Vilamoura procedure was rather quickly introduced to provide for such national standards which might be converted into regional standards. In the UK this has given many problems of possibly losing standards that are urgently needed to be maintained and gaining others which are not wanted. This is liable to produce problems unlikely to be resolved until well into the 21st century.

With the advent of CENELEC there has been a considerable change in the method of drafting British Standards in order to avoid conflict with HDs. When CENELEC begins work on a given subject, a 'standstill' or '*status quo*' arrangement comes into effect. Changes to national standards relating to that subject cannot then be made until harmonisation has been agreed, or permission obtained from CENELEC. It is still possible to have a national standard for a particular type of equipment which is of no interest to any other member countries of CENELEC, provided that the type of equipment is not just a local variant of an HD or EN. At the time of writing, individual customers may still obtain equipment to their own specification, sometimes as a 'special' carrying perhaps a delivery and/or a price penalty. However, freedom to obtain such 'specials' has been restricted from 1992 within the EEC especially for companies like the electrical supply industry, which are subject to the Public Procurement Directive whether or not they are privatised.

As the CENELEC standards are based on IEC standards wherever possible, active participation in IEC committee meetings and working groups has become more important than work in the BSI or CENELEC committees alone. Once a European Standard or a EURONORM is called up in an EEC Directive it is binding on Governments of all Member States, and the terms of the Directive will determine its status.

49.2.4 Underwriters Laboratories Inc. (UL)

The Underwriters Laboratories Inc. (UL) is an independent not-for-profit organisation in the USA which was originally set up by the insurance underwriters to carry out testing for public safety, but is now completely independent. It was originally founded in 1894 and is chartered as a not-for-profit organisation without capital stock, under the laws of the State of Delaware, to establish, maintain, and operate laboratories for the examination and testing of devices, systems, and materials. A complete description of the organisation, purposes, and methods of UL can be found in a pamphlet '*Testing for Public Safety*' obtainable from UL. The UL is given here as an example of a major test organisation in the USA, which is also very active in producing national standards in that country, in cooperation with US manufacturers of the products standardised.

UL will grant a 'listing' which 'observes' the requirements of the UL standard. In this respect, a product employing materials or having forms of construction differing from those detailed in the requirements of the standard may be examined according to 'the intent' of the requirements, and if found to be substantially equivalent, may be listed. Listings may be granted in certain circumstances, partly on tests at UL and tests carried out by the manufacturer. Under the UL arrangements a US manufacturer can obtain a listing and get a product on the market much more quickly than with the normal standardisation approval delay elsewhere. However, this is by no means a 'soft option' because UL usually requires 'follow-up' testing to be carried out at regular intervals, and any failure at that stage will result in withdrawal of listing, stopping production until the matter is put right. In this respect compliance with UL standards could be regarded as intermediate between 'type testing' (tested once only) and a service such as the BSI 'kite mark' scheme (of which there is more detail in Section 49.4).

49.2.5 Other bodies (ISO other national and regional committees etc.)

The International Organisation for Standardisation (ISO) was founded in 1947 in the aftermath of World War II, and the national standards bodies of approximately 90 different countries now participate in its work. ISO produces standards which are published on approval by 75% of the member bodies. ISO is governed by the ISO Council, and within the organisation there are in excess of 2000 technical committees, sub committees, and working groups involved in the preparation of International Standards. ISO had published approximately 8000 standards in the 20th century. In principle, ISO does not produce electrical standards, this being the province of its electrical counterpart the IEC. However, in practice, where the electrical standard relates to a product assembled within a larger product, ISO will take over the task of preparing a standard. An example of this is in automotive electrics. Electrical components in a motor vehicle constitute a significant proportion of the product, but such electrical components are not standardised by IEC but by ISO. A specific example is the standardisation of the fuses which would normally be the province of IEC/TC32. However, these are covered by an ISO committee concerned with the vehicles, although IEC/SC32C (Miniature Fuses) is kept informed of their decisions through a liaison system. This is typical of the type of liaison which exists between different standardisation bodies world-wide. Further detail of the ISO/IEC Code of principles on 'reference to standards' can be found in the following: Annex B to BS 0: Part 1: 1991.

The ISO Council organise their activities through a number of committees referred to under the following acronyms:

CERTICO	Committee on certification
COPOLCO	Committee on consumer policy
DEVCO	Development Committee (aimed at the needs of developing countries)
EXCO	Executive Committee (which is also responsible for finance)
INFCO	Information Committee
PLACO	Planning Committee
REMCO	Reference Materials Committee
STACO	Standardisation Principles Committee

The International Commission for Conformity Certification of Electrical Equipment (CEE) was set up immediately after World War II to harmonise electrical equipment standards. It comprised the national electrotechnical committees of 23 European countries together with observers from Australia, Canada, Hong Kong, Iceland, India, Japan, South Africa and the USA. A Certification Body (CB) was set up which could validate certificates of testing valid in all European member countries for electrical equipment tested in two of them at recognised testing authorities. This was known as the 'CB Scheme'. CEE also established the 'E' mark for marking a product to signify approval. This should be superseded by the 'CE' mark which should become mandatory throughout the EEC. CEE became later IECEE when it was taken over by IEC in 1985.

Most of the testing to this system is therefore now incorporated in the *IEC System for Conformity Testing to Standards for Safety of Electrical Equipment* (IECEE). IECEE's objective is the reciprocal recognition of test results among all participating countries, and offers the only world-wide recognition scheme for the safety of electrical equipment tested against IEC Standards. About half of the member organisations of the IEC are also members of IECEE. Addresses of most of these can be found in Section 49.5.1. The scheme is intended for electrical equipment used in homes, offices, workshops and similar locations and covers safety-related standards for the following categories of equipment:

(1) electronic entertainment equipment;
(2) measuring instruments;
(3) cables and cords;
(4) lighting equipment;
(5) household and similar electrical appliances;
(6) portable power tools;
(7) accessories, including certain components;
(8) switches for appliances and automatic controls;
(9) protective equipment for installation, including fuses and fuse-holders;
(10) low-voltage switchgear and controlgear;
(11) transformers;
(12) office machines and IT equipment; and
(13) electromedical equipment.

Under this scheme a manufacturer can contact the national certification body in his own country (BSI in the UK) and have his equipment tested in conformity to standards for safety at the designated testing laboratory. The equipment is then included in the '*List of CB Test Certificates*', where new certificates are published each year. Importers can procure conforming equipment by having it tested fully to the same scheme, but it only needs to be tested once to cover all countries for testing standards included in the scheme.

Importers can confirm which recently successfully type-tested equipment is and where, by consulting the regularly published *CB Bulletin* available from the IECEE Secretariat at the IEC Central Office.

There are many other international and regional groups concerned with aspects of electrical standardisation, but in the main they tend to apply IEC standards or North American variations. An international body is IFAN, the International Federation for the Application of Standards, which comprises the official standards-user bodies recognised by their national standards organisations. Regional bodies include ARSO, the African Region Organisation for Standardisation, which was founded in 1977 by the United Nations Economic Commission for Africa (ECA) and comprises the national standards bodies of the African countries who are members of ECA and the Organisation of African Unity (OAU), ASMO, the Arab Organisation for Standardisation and Metrology which was set up for the League of Arab States in 1965 as a specialised technical body in the field of standardisation, metrology and quality control, the Council for Mutual Economic Assistance (CMEA) which since 1949 acted to harmonise the national standards of USSR and Eastern Europe, and also later Cuba, Mongolia, and Vietnam, the Pan American Standards Commission (COPANT) comprising the standards bodies of the United States of America and 11 Latin American countries who since 1961 have been working to co-ordinate their policy of implementing IEC standards and recommendations within the Pan American region, the Pacific Area Standards Congress (PASC) which was founded in 1973 to assist countries in the Pacific area in international standards activities and promote closer cooperation between its members (active members Australia, Canada, China, Hong Kong, Indonesia, Japan, Republic of Korea, Malasia, New Zealand, Philippines, South Africa, Thailand and the USA), and the Nordic Electrotechnical Standards Cooperation Committee (NOREK) comprising Denmark, Finland, Norway and Sweden, which reviews the work of national organisations and formulates regional policy for presentation to IEC and CENELEC and ensures that standards prepared by IEC can be implemented in the Nordic countries with as few modifications or deviations as possible.

Although all major developed countries and regions belong to the IEC and are committed to eventual harmonisation of electrotechnical standards, there are still significant differences in details and special requirements and necessity for verification in a local test station which can so add to costs that a product from outside the region can be priced out of the market. It appears also that those countries which are most vociferous in wanting their own products admitted to other world markets with as low tariffs as possible and without testing other than to their own local requirements in their own country tend to be those with the most standards barriers to reciprocal trade. However, progress in the last quarter of the 20th century tends to indicate, with occasional minor setbacks, a trend towards a world-wide set of harmonised electrotechnical standards with common wording and interpretation in the early 21st century. It is also remotely possible that in that century, test certificates and approvals to an IEC standard from an internationally recognised test authority in any region might be recognised as good and sufficient proof of compliance with the appropriate electrotechnical standards in any other region in the world without further local testing.

49.2.6 Communication with organisations concerned with standards

It is clear from the above that anyone wishing to export products has first to penetrate the minefield of local standards regulations imposed either by government or by local traders in the target region. Also any contractor wishing to undertake a construction project, large or small, must apprise himself of all the standards mandatory in the region where his project is to be completed, since such compliance will be implicit in the contract, even if not specifically elaborated.

Appropriate addresses for enqiries in the various lands or international organisations are to be found in Section 49.5.1 (addresses of organisations concerned with standards).

However, the best source of information in the UK on standards matters is the British Standards Institution which runs a service of Technical Help to Exporters which was a popular free enquiry service, originally supported by DTI for the BOTB. However, in response to requests from users of the service and others to make it more comprehensive, the service has been considerably extended in its scope. This greatly expanded service, however can no longer be provided free of charge because of the possible incalculable very large cost of an open-ended service, and the considerably improved service is now self financing. With the advance of IT systems at the beginning of the 21st century, the internet is developing as the best and quickest route to information on standards. How this source is developing is detailed in Section 49.2.6.1.

49.2.6.1 The impact of the Internet on the distribution of standards

Any reader interested in standards who is not connected with the Internet should install the facility as soon as possible, with the fastest means of access available. Year 2000 saw the commencement of a big change in the distribution of standards and related material. At the beginning of the millennium more than 10 000 users had registered with the new service WEBSITE: bsonline.techindex.co.uk, email: bsonline@techindex.co.uk

This service, known as 'British Standards On-line' provides subscribers anywhere in the world with new and existing standards from its database. This method ensures that all documents transmitted are updated with the latest up-to-the-minute revisions, which previously might appear much later due to the delay in typesetting and printing. In addition, the user of the standard gets immediate access to the standards information he requires without any postal distribution delays. It is still necessary to pay for standards, because the work of standards publication is largely supported by sale of the standards. In previous years, free distribution of international standards might have led to vastly escalating publishing costs. In future years, however, if printing ceased to be used altogether, and some reliable system of financing standardisation work could be agreed, it might be possible to make all standards free on-line. At the time of writing this chapter there is no remote possibility of such a move, and the sale of standards is essential to support the continuation of the service.

In addition, most standardising bodies have installed websites where the latest freely distributed information is displayed, and from some of which standards can be purchased over the Internet. The two most important sites for UK users are the British Standards Website http://www.bsi-global.com and the IEC Website http://www.iec.ch which are good sources of general information on the standardisation of electrical products. Throughout the text of this chapter readers may find many other websites and e-mail addresses which may be used to gather further information concerning individual aspects of standardisation, and within the major websites more detailed addresses can be specially useful. For example, the IEC created within their website a possibility of viewing a slide presentation of seminars using Microsoft Powerpoint.

The leader of the project team on IEC 80416 promoted the use of 'virtual meetings' over the world-wide web. An interesting idea, though troublesome for those experts sitting in on a meeting when it is the middle of the night in their location!

The use of one system for the circulation of documents has great advantages because there are incompatibilities between different word-processing products. IEC have operated for several years with document circulation using Microsoft 'Word' for this purpose which can also readily be circulated by e-mail among working group and committee members. This was a step in the right direction but which gave many problems when some committee members invested in a later version of the product, distributed early in the USA, which could not be read on earlier versions and produced hundreds of pages of obscure symbols instead of the document. This matter needs to be standardised. Even at the Houston meeting the IEC provided a battery of PCs all programmed to read Word 6 which were all found to be unable to read documents on disk for the meeting brought by a convenor who had just bought a later version.

BSI committee members, commencing in year 2000 are to have all their committee documents distributed via the Internet. Eventually no paper copies of documents will be distributed to committee members (except of course to those who agree personally to meet the additional cost of providing such a service). The file format adopted by BSI at the time of writing is that of the Acrobat PDF file system which has been found very satisfactory by most users except for users with certain makes of printer which show difficulty in printing graphics in the latest version of Acrobat. PDF files are quite secure, however, and cannot be edited. This is a disadvantage for a committee member wishing to prepare a new draft, but the difficulty may be overcome by copying text into a Word document and then making the desired changes to that.

Committee members can read these files, and only need to print out those in which they have a particular interest or need for a forthcoming meeting.

IEC also distribute documents in PDF format. However they also use FTP sites, for which you have to be an FTP client. Some files are compressed and need to be un-zipped, for example with Winzip.

A further problem at the moment is the long period of time on the Internet sometimes required for downloading. However, contact with the Internet by faster means than the enhanced modem should undoubtedly overcome this difficulty in the future.

49.3 The structure and application of standards

The drafting and presentation of British Standards is itself the subject of a British Standard i.e. BS 0: Part 3: 1991 'A Standard for Standards: Part 3. Guide to Drafting and Presentation of British Standards'. In this part of BS 0, strong emphasis is placed on the need for precision in drafting, particularly in relation to titles, scopes, requirements, and test methods appropriate to specifications, and the advice in this standard is still of great value in preparing draft clauses for standards. However, this guidance can now only apply at the drafting stage, because most British (and CENELEC) standards are now to be IEC standards often with identical text. Any wording has therefore to be agreed by all International partners in the agreement, and the final version made into an identically worded British Standard even when it is not in 'good English' as spoken in England. The 1991 revision required alterations to the earlier editions of BS 0: Part 3, in order to accommodate the changes required by the IEC/ISO directives—Part 3; 1989 *Drafting and presentation of International standards* and the CEN/CENELEC Internal Regulations—Part 3: 1990 *Rules for the drafting and presentation of European standards*.

The example given below of structure of British Standards therefore reflects typical IEC structure, which sometimes conflicts with the aspirations of BS 0.

49.3.1 Different types of British Standard documents

British Standards are used to promote standardisation of a wide range of technical matters, terminology and symbols, marking and classification, methods of measuring, testing, analysing , sampling, etc.; methods of declaring, specifying, etc.; specifications for materials or products, specified dimensions, performance, safety requirements, etc.; specification of processes and practices etc.; recommendations on applications of products and processes, and desirable safe procedures.

The type of British Standard document for each of these purposes must obviously vary since the purpose of the document is different, varying from a set of instructions to a code of advice on good practice. There are thus British Standard Specifications, Drafts for Development (where a new technology is rapidly being introduced), Published Documents of information and guidance, British Standard Codes of Practice, Guides, Handbooks, Glossaries, etc. Standards may also be referred to in Regulations to avoid inclusion of detailed technical provisions and criteria in the body of the law.

In the following parts of this section some more detail is given of these different forms of documents produced by the BSI.

49.3.1.1 British Standard Specifications

A British Standard Specification is the most well known and understood type of BSI standards publication. It lays down a set of requirements to be satisfied by the product, material, or process in question, and embraces, often by reference, the relevant methods by which compliance may be determined.

The general series of British Standards uses the prefix BS and is numbered sequentially. This is followed by a colon and the part or section number (if any) and then another colon and year of issue (or year of revision when the standard is revised). At one time this nomenclature was mainly restricted to British Standard Specifications, but now is also used for codes of practice (see Section 49.3.1.4). The number has no special significance, although many British Standard committees keep one number for a product (e.g. BS88: Part 4: 1988 has been retained for low-voltage fuses for the protection of semiconductor devices) and only the date is changed when the standard undergoes revision. At the time of writing it is possible that the parts of IEC 20269 corresponding to BS 88 may become Harmonised Documents (prefix HD), but it is unlikely that they will become ENs. This is because the harmonisation document permits a country to choose only those parts of the document which are to be used in that country. If the document becomes a full EN, all parts become the British Standard. Although the performance standards are practically the same for all fuses in IEC 60269, the dimensions are very different. If all the parts of the IEC standard were to be included in the EN, fuses would be supplied which would not fit into any existing UK installation, and users would not want to rip out their installations just in order to replace them with fuses of exactly the same performance but of new dimensions.

The automobile series are indicated by the letters BS AU plus a sequential number with a suffix letter which advances with every amendment (e.g. BS AU 242a: 1998).

The prefix BS ISO indicates the British version of an ISO standard, and BS CISPR similarly a CISPR document.

The prefix PAS implies a product assessment specification issued as an interim measure, and QC in the prefix relates to quality control.

The marine series are indicated by the prefix BS MA and is numbered sequentially.

The aerospace series uses a still different nomenclature, being numbered sequentially within classes, each class being indicated by a letter, e.g. ‘F’ for fabrics, ‘S’ for steels, etc. Editions other than the first bear a prefix number to show the edition. e.g. BS 3M 60:1998 is the third edition whereas in the case of BS Z 15:1998 no current standard is superseded.

Amendments became numbered sequentially under the prefix AMD, although earlier amendments were numbered as published documents (e.g. in BS 3036 Amendment No. 3 was classified as PD 5141, whereas Amendment No. 4 is AMD 2463).

Every British Standard used to be also classified according to the Universal Decimal Classification (UDC) (details in BS 1000). The UDC number was assigned at the manuscript stage and for older British Standards it can be found in small print beneath the BS number on the front cover. Every BSI publication for sale (except amendments) carries also an International Standard Book Number (ISBN) on the outside back cover. This appears usually near the top left-hand side of the back cover (e.g. on the rear cover of BS 88: Section 2.2:1988 you will find ‘ISBN 0 580 16846 8).

Other general information, outside the main body of the standard, tends also to be printed on the inside and outside of the front and back cover. A group number may also be included which indicates the purchase price of the standard.

In the case of a dual numbered standard where there is identical text, the international or regional document number is printed beneath the British Standard number on the top right-hand corner of the front cover: The following example is a European Standard (EN) which is now a part of BS88:

BS EN 60269-3:1995
BS 88 Section 3.1:1995
IEC 269-3 :1987

In this example, in the numbering of the EN it can also be detected that the standard is derived from IEC 269-3. The general series of British Standards have white covers, which identifies them from other types of British Standard publication. This will have decreasing relevance as distribution goes electronic.

49.3.1.2 Drafts for development

Drafts for Development (DDs) can be regarded as equivalent to the ‘provisional’ or ‘tentative’ standards that are issued in certain countries outside the UK. They are published when guidance is urgently needed, for example in a new and rapidly developing technology, but where such guidance, though theoretically sound, has not yet been subjected to enough practical application to justify the publication of a British Standard. DDs are particularly advantageous for setting a framework for new test methods where extensive use of the tests is needed to establish satisfactory repeatability and reproducibility. The IEC equivalent to a DD is the Technical Trend Document (TTD). An example is the dual numbered document:

DD 183:1989
IEC 127-4 TTD

Drafts for Development may be converted into British Standards of any type or withdrawn when sufficient experience has been obtained and considered by the appropriate committee.

Drafts for Development are bound in vivid orange covers to distinguish them from other British standard publications. After 1992 IEC ceased publication of TTDs, and only a few DDs were current at the start of the millennium.

49.3.1.3 Published documents

Published Documents (PD) are miscellaneous documents containing supplementary information relating to standardisation. A good example is PD 5688 covering the use of 'SI Units'. This is a useful little booklet which was published by BSI when the metric system was being introduced into British Standards and everyday life in the UK. It gives a brief account of the way the SI system of units has evolved, tabulates the basic units and the principal supplementary and derived units and a few conversion factors relating SI units to the imperial units. (The SI system of units is used in all IEC Standards as well as in British Standards.)

Published Documents were all published in blue covers, except for the Education Section of PDs which had yellow coloured covers.

49.3.1.4 Codes of practice and guides

Codes of Practice have the main function of recommending good accepted practice as followed by competent practitioners. Codes of Practice assemble the results of practical experience and scientific investigation in a form that enables the reader to make immediate use of proven developments in particular branches of industry.

Before 1975, Codes of Practice were numbered in a separate CP series with a red cover. Nowadays, new and revised Codes of Practice are each given a new number in the BS series, retaining the title '*British Standard Code of Practice for . . .*'.

In contrast to British Standard Specifications, British Standard Guides and Codes of Practice are written in the form of guidance only, and are not intended to provide objective criteria by which compliance may be judged. Consequently the word 'must' is never used in a Code of Practice, and 'shall' is replaced by 'should' because none of the requirements are mandatory. Neglect of safety measures prescribed in Codes of Practice and Guides might now however be taken as evidence of liability in the case of an accident.

Care is taken by drafting committees to ensure that Codes of Practice do not become textbooks, the principles behind a particular practice being discussed or explained only when absolutely necessary. Specific recommendations for avoiding certain existing practices are made only where tacit approval of these practices would otherwise be assumed by the reader or if the practices in question may be hazardous. Where appropriate, a code offers a series of options and identifies the implications of accepting them.

49.3.1.5 Glossaries and methods

British Standard Glossaries are documents bringing together agreed sets of terms and definitions for reference. They normally contain a contents list, classified sections giving terms and definitions numbered by section and subsection in a special code which assists retrieval of information, and an alphabetic index referring to the same special code.

British Standard Methods comprise a variety of standards describing formalised ways of performing given tasks. The methods of doing this are clearly stated and explanations given of any calculations etc., necessary to complete each task. British Standard methods can be included in specifications that require them, or alternatively may be published as separate standards with a general BS reference. This is a useful arrangement when the methods are required to be called up in two or more other standards.

49.3.1.6 Handbooks

British Standards Handbooks comprise texts taken from a number of British Standards, complete and/or in part, together with related material previously published elsewhere relating to a particular field. Since a handbook might contain only an extract from a British Standard, problems of interpretation may arise. In any such case, the full text of the British Standard in question should be consulted.

An example of a Handbook was published around the time of 'metrication' mentioned above, the British Standard Handbook No. 18 'Metric Standards for Engineering', which brought together in one volume all the information then currently available on metric standards for general engineering. When first published it was a 580 page book with all available standardised metric information. Comparison of this with the little booklet PD 5686 mentioned above will give some concept of the difference between a Handbook and a Published Document.

49.3.1.7 Regulations

A Regulation is a binding document which contains legislative, regulatory or administrative rules and which is adopted and published by an authority legally vested with the necessary power. References made to standards in regulations may have one of two effects:

(1) *Standards made mandatory*: the standard or the part of the standard which is referred to must be followed, or a specific result in a standard test must be achieved in order to obey the statutory requirement. This means that the text in the standard ceases to be voluntary in the context of the legal requirement.

(2) *Standards deemed to satisfy*: in this case, compliance with the standard is indicated as one way of fulfilling a regulatory requirement. It is possible to choose another route to fulfil the requirement, but those doing so may be required to prove that their alternative complies with the regulation.

Until recently, Regulations might call up British Standards, but are not British Standards. The separation of the regulatory authority and the Standardisation body has needed modification to fit in with the arrangements with CENELEC in co-ordination of Electrical Regulations. Under EEC Directives, European Standards, dual numbered with British Standards cease to be voluntary, and thus effectively become Standard Regulations. This may require changes to the present UK organisation of the regulatory process.

The principal example of this is the case of the *IEE Wiring Regulations—Regulations For Electrical Installations*, popularly known as the 'wiring rules'. These are at present drawn up by the Wiring Regulations Committee of the Institution of Electrical Engineers. The parallel international work takes place in IEC, which is the international equivalent of the electrical aspects of BSI. The international work on regulations for electrical installations is undertaken by IEC TC64 and the position of the UK regulatory system would fit better into the system if parallel work was carried out by BSI committees issuing standards which take into account CENELEC harmonisation documents. More detail of the problems of co-ordination with International Standards (such as IEC 60364—'Electrical Installations of Buildings') was made in the preface to the 15th edition (amended) of the IEE wiring regulations. Such a reorganisation would not necessarily change the organisations or expertise used in doing the technical work but the Regulations have had to become harmonised standards. These difficulties were resolved in the year 1990/91 by the formation of BSI committee PEL/64 and its subcommittees matching equivalent IEC/TC64 and CENELEC TC64 committees and sub-committees.

In 1997, the 16th edition of the wiring regulations was also published as the second edition of BS 7671.

49.3.2 International equivalents of British Standard documents

As has been noted in Section 49.3.1.1 and Section 49.3.1.2 and elsewhere there are equivalent IEC and European Standards to the more important electrical British Standards. These are indicated by the dual numbering system. As time progresses, however, the distinctions between all such documents tend to decrease, and national deviations diminish. This is particularly the case within Europe, where CENELEC regulations stipulate conditions giving European Standards the status of a national standard without any alteration in the member states.

Many standards current in North America do not have direct equivalents elsewhere. If the USA with its North American bloc could move a little faster in aligning its national standards with IEC, we could see most national deviations vanishing by the year 2050.

49.3.3 Structure of a typical standard

Since most electrotechnical British Standards are directly derived from IEC standards, the structure and wording of the standards originates in IEC decisions rather than independent national decisions. The following is the structure of a typical IEC standard specification:

(a) Foreword (followed by preface and introduction where appropriate), giving the background to the specification and its international development;
(b) Scope;
(c) Object;
(d) Definitions;
(e) Conditions For Operation in Service;
(f) Classification;
(g) Characteristics;
(h) Marking;
(i) Standard Conditions for Construction; and
(j) Tests.

This structure has been criticised in the past in that it sometimes leads to a certain degree of repetition in different clauses. However this disadvantage is offset by the considerable advantage that all the requirements are clearly stated in each clause, and *all* the test requirements are gathered together in the final section rather than being distributed throughout the text. It is thus perfectly clear exactly what has actually been tested and what the product has achieved if it successfully complies with the specification.

49.3.4 Implementation of standards

The implementation of standards is controlled by factors seen above to be changing with the growth of international regulation. British Standards have historically always been *voluntary* in their implementation. However, the changes implied by the EEC Directives are shown above to make them *mandatory* when they become ENs. A similar situation is seen in Section 49.3.1.7 to exist with Regulations.

A British Standard Specification forms part of the trade description of a product when quoted by a BS number or when compliance with it is claimed. Marking with a BS number constitutes a unilateral claim that the product complies with *all* the requirements of the BS quoted. The person making such a claim is responsible for its accuracy under the Trade Descriptions Act 1968. To support their claims, manufacturers may have their products certified (see Section 49.4) as complying with the requirements of the appropriate British Standard Specification.

The existence of a British Standard Specification facilitates the preparation of contracts, and any British Standard Specification invoked in a contract becomes legally binding on the contracting parties. However, the compliance with the British Standard does not of itself confer immunity from legal obligations. The user is only guaranteed that the product supplied complies with all the requirements of the standard, and must satisfy himself that these are sufficient to fulfil the tasks for which he is purchasing the product. Many British Standard Specifications contain options and other matters which need to be clarified in any contract to ensure that the correct variant of the British Standard is being specified.

As explained above, British Standard Codes of Practice, Guides and similar recommendations are written in the form of guidance only, and are not intended to provide objective criteria by which compliance may be judged. Such publications are not appropriate for simple reference in contracts.

There are three methods of reference to British Standards used in Regulations and elsewhere:

(1) *Reference to standards by exact identification (strict reference)*: one or more specific British Standard(s) are designated in such a way that later revisions of any such British Standard will not be applied unless the reference is modified. The British Standard is usually indicated by its number and date. This is the method of reference normally used in the UK.
(2) *Reference to standards by undated specification (undated reference)*: one or more specific British Standard(s) are designated in such a way that later revisions of any such British Standard will be applied without the reference needing to be changed. In this case the British Standard is usually designated by its number only.
(3) *General reference to standards*: reference is made in a general way to present and future standards, which means that the relevant reference includes a general clause so that all the present and future standards in a specific field are regarded as meeting the aims of the document containing the reference (e.g. the document could be a regulation).

49.4 Testing, certification and approval to standard recommendations

To confirm compliance with British Standards it is necessary to test and to mark the appropriate products to identify compliance with the appropriate British Standard and/or that the manufacture is maintained at an acceptable level of quality. BSI operates four certification marking systems:

(1) The '*kitemark*', the BSI's certification mark indicates not only compliance with the appropriate standard, but also compliance with a rigorous program of surveillance, inspection and follow-up testing at regular intervals.
(2) The '*safety mark*', which was introduced in 1974 to provide manufacturers with a means of demonstrating compliance with a British Standard specifically related to safety.
(3) The '*registered firm symbol*', operated for manufacturers who produce goods which are not at present covered by a British Standard.
(4) The '*BS 9000 mark*' for electronic components certified under the BS 9000 system.

In its certification activities, BSI does not have, or seek, a monopoly position, but responds to UK needs and aims to provide a service which can be used by industry. Through the Quality Assurance Council the BSI cooperates with other organisations concerned with certification of compliance with standards. The more important of these are:

(1) British Electrotechnical Approvals Board (BEAB);
(2) British Approvals Service for Electric Cables (BASEC);
(3) Association of Short-Circuit Testing Authorities (ASTA);
(4) British Approvals Service for Electrical Equipment for Flammable Atmospheres (BASEEFA).

The BSI has compiled a register of test houses of assessed capability indicating their fields of testing.

The BSI Certification and Assessment Department is responsible not only for the certification of products but also for the assessment of the capabilities of the firms in manufacturing and service industries. The growing realisation of the importance of quality and reliability in goods and services has caused rapid growth in the percentage of firms seeking registration in schemes such as BS 5750. Registration and maintenance to such a quality standard has a dual advantage. Compliance is not only an attraction to customers, it is also a benefit to the efficient business of the registered firm improving its competitiveness and ensuring the maintenance of that quality by independent assessment to the British Standard.

The BSI Test House centred at Hemel Hempstead provides a wide range of testing facilities which are generally available to clients world-wide and embraces tests which are electrical, electronic, mechanical, chemical, photometric, automotive, etc. (www.inspectorate.com)

BSI also carry out import/export inspections, arranging for inspection before shipment at the place of manufacture. The BSI inspectorate is accredited by the following organisations in the UK:

(1) BSI,
(2) BASEC,
(3) Department of the Environment (Property services agency)
(4) Department of Transport,

and it also has accreditation from the following overseas organisations:

(1) Canadian Standards Association (CSA),
(2) Centre Technique du Bois (*France*),
(3) Institute for Industrial Research and Standards (IIRS) (*Eire*),
(4) Instituto del Marchio di Qualita (IMQ) (*Italy*),
(5) Staatliches Materialpruefungsamt (MPA) (*Germany*),
(6) Standards Association of Australia (SAA),
(7) Standards Association of New Zealand (SANZ),
(8) Statens Planverk (*Sweden*),
(9) Technishe Uberwachungs Verein (TUV),
(10) Underwriters Laboratories Inc. (UL),
(11) Underwriters Laboratories of Canada (ULC)
(12) Verband Deutscher Electrotechniker (VDE) (*Germany*).

BEAB certifies household appliances and all home laundry equipment, heating and cooking appliances, refrigerators and freezers, home sound and vision equipment, etc. for compliance with relevant British Standards, for example the many parts of BS 3456. These are mostly already in agreement with CENELEC HDs and the ENs which are superseding parts of BS 3456.

Many retailers in the UK large and small will only stock BEAB-approved appliances.

BASEC provides a certification scheme for manufacturers of electric cables and flexible cords, and ensures that ongoing quality control procedures are adequate to ensure consistently high standards. BSI provides the assessment and inspection facilities (but not the operations management service which it provided in earlier years).

ASTA operates a certification and product marking scheme for circuit-breakers, fuses, fuse-links, fuse-boards, switches, isolators, starters, transformers, reactors and electrical wiring accessories. It offers a number of classes of test certificate, including a certificate of complete compliance with the requirements of a relevant standard.

BASEEFA certifies electrical equipment using any of the recognised forms of explosion protection, and certifies equipment intended for zone II areas. Certification is against recognised standards, where these are available, and otherwise to standards prepared by BASEEFA.

International Standards are prepared with great care, and every attempt is made to avoid ambiguity in specifying tests. However, where any possible ambiguity becomes apparent in the practical performance of the tests, the above bodies assess the most appropriate interpretation of the test. Such interpretations are today often agreed internationally by groups representing test authorities so that tests are of equal severity wherever carried out.

49.5 Sources of standards information

Many libraries throughout the UK have stocks of British Standards and some also stock foreign standards. The most comprehensive source of information, however, is the BSI which has established an extensive standards information service.

BSI library and enquiry service The BSI Library contains a full set of all current and superseded British Standards, a collection of International Standards and complete sets of standards from 80 foreign standardisation bodies. This is one of the world's largest collections of standards, comprising over half a million documents. Documents are available on loan to members of BSI at a small charge.Tel: 0208 996 7004 fax: 0208 996 7005

BSI PLUS (private list updating service) Available to BSI subscribing members. PLUS monitors and updates subscribers' standards library automatically.

BSI enquiry service This deals with direct enquiries concerning British Standards, international regional and foreign standards. e-mail: info@bsi.org.uk or look in at the website http://www.bsi-global.com

BSI catalogue This publication, up-dated annually, lists all published British Standards and other special series, together with short abstracts.

BSI update STANDARDS These are issued monthly cumulatively up-dating the latest catalogue and gives details of all new and revised British Standards, amendments, withdrawals, renewals, drafts issued and new work started, as well as details of IEC and CENELEC documents.

BSI business STANDARDS This publication is issued bi-monthly, and in addition to articles and general information on current standards topics, it gives BSI information on topics such as the Reader Enquiry Service (Tel: 020 8996 9001), a comprehensive updated directory of BSI senior staff and services, details of membership services and of the British Standards Society, and details of events such as the communication days 2000 and the training services of BSI Business Solutions Ltd.

BSI annual report This is published each year in October, and it gives a review of the year's work and a statement of accounts.

49.5.1 Addresses of organisations concerned with standards

British Standards Institution
Head Office; BSI Standards
389 Chiswick High Road
London W4 4AL
United Kingdom
Tel: +44 (0) 181-996 9000
Fax: +44 (0) 181-996 7400
e-mail: info@bsi.org.uk

Hemel Hempstead
Maylands Avenue
Hemel Hempstead HP2 4SQ
United Kingdom
Tel: +44 (0) 1442 230442
Fax: +44 (0) 1442 231442
e-mail: info@bsi.org.uk

Units 1, 5 and 6
Finway Road
Hemel Hempstead
Herts
HP2 7PT
United Kingdom
Tel: +44 (0) 1442 278504
Fax: +44 (0) 1442 232442

Brazil
BSI Brazil
Av. Ana Costa 151-Cj 32-3 andar
Vila Mathias
Santos
Brazil
CEP 11060-000
Tel: 55 13 232 1144
Fax: 55 13 235 4750
e-mail: ellie_borges@bsi-inc.org

China-Shenzhen
BSI Pacific Ltd.
Room F, 22/F , Shangbu Building
Nan Yuen Road
Shenzhen 518031
China
Tel: +86 755 323 5472
Fax: +86 755 321 2434
e-mail: bsisz@public.szonline.net
Website: www.bsi-pacific.org

China-Shanghai
BSI Pacific Ltd.
Unit3, 28/F Nan Zheng Building
580 West Nanjing Road
Shanghai 200041
China
P.R. China
Tel: +86 21 5234 1101
Fax: +86 21 5234 1102

France
BSI France
Quai Southampton
F-76600 Le Havre
France
Tel: +33 2 35 21 90 00
Fax: +33 2 35 41 21 41
e-mail: BSI.FRANCE@wanadoo.fr

Hong Kong and Macao
BSI Pacific Ltd.
Unit C, 5/F Garment Centre
576 Castle Peak Road
Kowloon
Hong Kong
Tel: +852 2742 5638
Fax: +852 2743 8727
e-mail: ilam@bsi.com.hk
web: www.bsi-pacfic.org

India
BSI India Pvt. Ltd.
201, Ansal Bhawan
K. G. Marg
New Delhi 110 001
India
Tel: +91 11 371 9002/3, +91 11 373 9003/4
Fax: +91 11 294 2920

Japan
BSI Japan K.K.
Nanpeidai Aie Aie BLDG. 3F.
15-12, Nanpeidai-Cho
Shibuya-Ku
Tokyo 150-0036

Japan
Tel: +81 (0)3 5459 9331
Fax: +81 (0)3 5459-9332
e-mail: james azim@bsi-japan.com

Korea
BSI Quality Services Korea Ltd.
Suite #321, KCCI Building
45, Namdaemoonro-4ka
Chung-gu, Seoul 100-743
Korea
Tel: +82 2 777 4123
Fax: +82 2 777 4446

Mexico
BSI Mexico
Av. N. Bravo No 1203
96400 Coatzacoalcos
Veracruz
Mexico
Tel: +52 921 29646

Poland
BSI (Poland Branch)
Al Jerozolimskie 49, m5,
PL-00-697 Warsaw
Tel/Fax: +48 22 628 1917
e-mail: stephens@pol.pl

Scotland
Quality House
2000 Academy Park
Gower Street
Glasgow G51 1PP
United Kingdom
Tel: +44 (0)141-427 2825
Fax: +44 (0)141-427 5989
e-mail: Info@bsi.org.uk

Singapore
1 Maritime Square #09-21,
World Trade Centre,
099253 Singapore
Tel: +65 270 0777
Fax: +65 270 2777
e-mail: isospore@mbox3.singnet.com.sg

South Africa
BSI Quality Services South Africa
PO Box 2079
Southdale 2135
South Africa
Tel: +27 11 835 2830
Fax: +27 11 496 3704
e-mail: elsie@inspml.co.za

Spain
BSI Espana
Paseo de la Castellana, 111, 4f,
E-28046 Madrid
BSI Espana
Ctra. Fuencarral a Alcobendas KM14-5
C./ Sepulveda 6
E-28108 Alcobendas
Tel: +34 91 662 3857
Fax: +34 91 661 8864

Taiwan
BSI Pacific Ltd—Taiwan Branch
14/F Huei Fong Building
No. 27 Chung Shan North Road
Section 3
Taipei 10451
Taiwan
Tel: +886 2 2586 2674
Fax: +886 2 2594 4234
e-mail: bsitwn@ms21.hinet.net
web: www.bsi-pacific.org

USA
BSI Inc
12110 Sunset Hills Road
Suite 140
Reston, VA 20190
1-800-862-4977
Tel: +1 (703) 437-9000
Fax: +1 (703) 437-9001
email: BSI Inc@compuserve.com
web: www.bsi-inc.org

Wales
QED Centre
Main Avenue
Treforest Estate
Pontypridd
Mid Glamorgan CF37 5YR
United Kingdom
Tel: +44 (0)1443 841381
Fax: +44 (0)1443 841373
e-mail: Info@bsi.org.uk

BSI Quality Assurance
BSI QA is a part of Inspectorate (details from BSI Hemel Hempstead or e-mail: info@inspectorate.co.uk)

Technical Help To Exporters
Tel: 0208 996 7111
Fax: 0208 996 7048

Overseas Trade Services
website: http://www.dti.gov.uk/ots

BSI Business Solutions
0208 996 7559

The Institution of Electrical Engineers (IEE)
Savoy Place
London WC2R OBL
Tel: 0207 240 1871

The British Electrotechnical Approvals Board (BEAB)
9/11 Queens Road
Hersham
Walton-on-Thames
Surrey KTI2 SNA

The Association of Short-circuit Testing Authorities (ASTA)
23/24 Market Place
Rugby
Warwickshire CV21 3D

The British Approvals Service for Electrical Equipment for Flammable Atmospheres (BASEEFA)
Health and Safety Executive
Harpur Hill
Buxton
Derbyshire

BEAMA
Westminster Tower
3, Albert Embankment
LONDON
SE1 7SL
Tel: 0207 793 3000
Fax: 0207 793 3003

ERA Technology Ltd.
Cleeve Road
Leatherhead
Surrey KT22 7SA

International

The International Electrotechnical Commission (IEC)
3 Rue de Valembé
CH-1211 Geneva 20
Switzerland
Website: http://www.iec.ch

The International Organisation for Standardisation (ISO)
Case Postale 56
CH-1211 Geneva 20
Switzerland

CENELEC General Secretariat
2 rue Briderobe, Bte 5
1000 Brussels
Belgium

International Commission for Conformity Certification of Electrical Equipment (CEE)
Utrechtseweg 310
Arnhem
Netherlands

International Conference on Large High-Voltage Electric Systems (CIGRE)
112 Boulevard Haussmann
F-75008 Paris
France

North American

American National Standards Institute (ANSI)
1430 Broadway
New York
NY 10017
USA

American Society for Testing and Materials (ASTM)
1916 Race Street
Philadelphia
PL 19013
USA

Institute of Electrical and Electronics Engineers (IEEE)
345 East 47th Street
New York
NY 10017
USA

National Electrical Manufacturers' Association (NEMA)
2101 L Street NW
Washington
DC 20037
USA

National Standards Association (NSA)
1321 14th Street NW
Washington
DC 20005
USA

Underwriters' Laboratories Inc. (UL)
333 Pfingsten Road
Northbrook
IL 60062
USA

Association of Home Appliance Manufacturers (AHAM)
20 North Walker Drive
Chicago
IL 60606
USA

Aerospace Industries Association (AlA)
1725 de Seles Street NW
Washington
DC 20036
USA

American Welding Society (AWS)
2501 NW Seventh Street
Miami
FL 33125
USA

Computer and Business Equipment Manufacturers' Association (CBEMA)
1828 L Street NW
Washington
DC 20036
USA

Electrical Apparatus Service Association (EASA)
1331 Bause Boulevard
St Louis
MO 63132
USA

Edison Electric Institute (EEI)
1111 19th Street NW
Washington
DC 20036
USA

Electrical Testing Laboratories Inc. (ETL)
Industrial Park
Courtland
NY 13045
USA

Illuminating Engineering Society (IES)
345 East 47th Street
New York
NY 10017
USA

Insulated Power Cable Engineers' Association (IPCEA)
South Yarmouth
MA 02664
USA

Instrument Society of America (ISA)
400 Stanwix Street
Pittsburgh
PA 15222
USA

National Fire Protection Association (NFPA)
Battermarch Park
Quincy
MA 02269
USA

Pacific Region

Standards Association of Australia (SAA)
Standards House
80 Arthur Street
North Sydney
NSW 2060
Australia

Standards Association of New Zealand (SANZ)
Private Bag
Wellington
New Zealand

Bodies affiliated to the IECEE

Due to likely changes in the constitution of this body in the early years of the millennium, enquiries are best routed through the Central Office:

IECEE
c/o Central Office of the IEC,
3, rue de Varembé
CH-1211 GENEVA 20
Tel: (+41 22) 34 01 50
Fax: (+41 22) 33 38 43